

LREC-COLING 2024

**First Workshop on  
Patient-Oriented Language Processing  
@LREC-COLING-2024  
(CL4Health)**

Workshop Proceedings

Editors

Dina Demner-Fushman, Sophia Ananiadou,  
Paul Thompson and Brian Ondov

20 May, 2024  
Torino, Italia

**Proceedings of the First Workshop on Patient-Oriented Language Processing  
@LREC-COLING-2024 (CL4Health)**

Copyright ELRA Language Resources Association (ELRA), 2024  
These proceedings are licensed under a Creative Commons  
Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-25-8  
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association  
and the International Committee on Computational Linguistics

## Preface

The First Workshop on Patient-Oriented Language Processing (CL4Health) aims to establish a general venue for presenting research and applications focused on patients' needs. These include summarizing health records for patients, answering consumer-health questions using reliable resources, detecting misinformation or potentially harmful information, and providing multi-modal information, such as video, if it better satisfies patients' needs. Such a venue is needed both to invigorate patient-oriented language processing research and to build a community of researchers interested in this area. The growing interest in this topic is fueled by several current trends, which include a proliferation of online services that target patients but do not always act in their best interests; policy changes that allow patients to access their health records written in the professional vernacular, which may confuse the patients or lead to misinterpretation; replacement of customer services with chat bots; and the increasing tendency of patients to consult online resources as a second or even first opinion on their health problems.

Broadly, CL4Health is concerned with the resources, computational approaches, and behavioral and socio-economic aspects of the public interactions with digital resources in search of health-related information that satisfies their information needs and guides their actions.

## Invited Speakers

The invited speakers have devoted significant parts of their research to patient-centered language processing. We are grateful and excited to present the following talks:

### **Barbara Di Eugenio, University of Illinois Chicago, USA**

#### **Engaging the Patient in Healthcare: Summarization and Interaction**

Effective and compassionate communication with patients is becoming central to healthcare. The talk discusses the results of and lessons learned from three ongoing projects in this space. The first, MyPHA, aims to provide patients with a clear and understandable summary of their hospital stay, which is informed by doctors' and nurses' perspectives, and by the strengths and concerns of the patients themselves. The second, VIRTUAL-COACH, models health coaching interactions via text exchanges that encourage patients to adopt specific and realistic physical activity goals. The third, HFChat, envisions an always-on-call conversational assistant for heart failure patients, that they can ask for information about lifestyle issues such as food and exercise.

#### **Brief Biography**

Dr. Di Eugenio's work is characterized by: large interdisciplinary groups of investigators who bring different perspectives to the research; grounding computational models in ecologically valid data, which is small by its own nature; and the need for culturally valid interventions, since the University of Illinois Health system predominantly serves underprivileged, minority populations.

**Natalia Grabar, University of Lille, France**

### **Linguistic Foundations of the Simplification and its Current State**

The purpose of text simplification is to adapt the content of documents in order to make their reading and understanding easier for a given type of population. If the simplification usually aims specific language levels (lexical, morphological, syntactic, semantic...), the available data cannot always provide precise indications required for this process. The talk discusses some sources of such available data. Dr. Grabar also analyzes the current situation related to the exploitation of linguistic indicators during the definition of language complexity and the simplification.

### **Brief Biography**

Dr. Grabar is a CNRS Researcher at the University of Lille. She studied philology at Lviv University, Ukraine and obtained her PhD in Medical Informatics from the Université Paris 6, France. She develops linguistic and statistical methods to access information and knowledge within scientific and technical texts and terminologies. The results are used in information retrieval, information extraction and text simplification. Dr. Grabar has co-authored over 200 publications.

**Graciela Gonzalez-Hernandez, Cedars Sinai Medical Center, USA**

### **Patients are speaking - are we listening? Incorporating patient perspectives posted online into clinical trials**

Research that aims to be equitable and effective at treating chronic diseases and improving patient outcomes must incorporate a broad range of patient perspectives (health-related uncertainties, beliefs, and experiences). Setting research priorities and designing trials is complex since clinicians, researchers, and patients differ on what is considered important. Patients often prioritize outcomes that directly impact their quality of life, such as symptom relief, functional status, and treatment side effects, while clinicians prioritize outcomes related to survival, disease progression, and biomarker endpoints. Methods commonly used for gaining patient perspectives are often limited are subject to recall and other biases, are expensive and time-consuming, are limited in recruitment number and diversity, and may not comprehensively capture factors important for research design.

A vast amount of data from the patient's perspective is already publicly available: patients openly share useful perspectives on different social media platforms. Despite its potential, approaches for the systematic integration of such data to inform the prioritization and design of health research are still to be developed and validated.

In this talk, Prof. Gonzalez-Hernandez discusses her ongoing efforts to enable the extraction of relevant patient perspectives posted online using state-of-the-art natural language processing (NLP) methods, and the promise of their integration into clinical trial design.

## **Brief Biography**

Dr. Gonzalez-Hernandez has over 23 years of experience and more than 200 publications in health AI and NLP, funded by multiple NIH grants. She is currently a Professor and Vice Chair for Research and Education in the Cedars-Sinai Department of Computational Biomedicine. She launched the #SMM4H (Social Media Mining for Health) Workshop and Shared Tasks, which has run annually for the last 8 years.

## **Abeed Sarker, Emory School of Medicine, USA**

### **Learning and Educating via NLP of Social Media: the Use Case for Substance Use and Overdose in the United States**

Substance use and overdose is an ongoing crisis in the United States and growing globally. The sphere of substance-related overdose also evolves continuously as novel psychoactive substances enter the supply. Nonmedical substance use surveillance via social media has the potential to provide low-cost and more timely insights than traditional approaches. In our research, we leverage natural language processing (NLP) and machine learning to obtain insights from targeted cohorts of people who use substances about emerging patterns and problems in substance use disorder and treatment. This talk outlines our NLP pipeline for analyzing substance use-related chatter from Twitter (X) and Reddit, and how insights derived from these sources may be used to educate medical practitioners at the forefront of the opioid crisis in the United States, facilitating more patient-centered care.

## **Brief Biography**

Dr. Sarker is an Associate Professor and the Vice Chair for Research at the Department of Biomedical Informatics, School of Medicine, Emory University. He leads several large-scale projects focusing on the application of NLP for health-related tasks, particularly those involving vulnerable populations such as people with substance use disorders, victims of intimate partner violence, and people at risk of self-harm and suicide. His research is primarily funded by the National Institutes of Health (NIH) and Centers for Disease Control and Prevention (CDC). Dr. Sarker's research has been covered by various national and international media outlets such as the Wall Street Journal, Forbes, and Scripps National News.

## **Submissions**

CL4Health received 40 valid submissions, of which 8 were accepted as oral presentations and 25 as posters. The work covers a wide range of topics focusing on patients' well-being and proper care. The topics include retrieval augmented generation, communications (including plain language, sign language, and dialog), mental health issues, and patients' sentiment.

As always, we are deeply grateful to the authors of the submitted papers and to the reviewers (listed elsewhere in this volume) who produced thorough and thoughtful reviews for each paper in a fairly short review period. The Organizers are truly grateful to our amazing Program Committee, whose members helped us determine which studies are ready to be presented

and those which would benefit from additional experiments and analysis, as suggested by the reviewers. We hope that this workshop will inspire new collaborations and research into patient-centered language technologies, in order to continue the valuable contributions made by our community towards public health and well-being.

*Dina Demner-Fushman, Sophia Ananiadou, Paul Thompson and Brian Oudov (Organizers)*

## **Organizing Committee**

Dina Demner-Fushman, National Library of Medicine, USA  
Sophia Ananiadou, National Centre for Text Mining and University of Manchester, UK  
Paul Thompson, National Centre for Text Mining and University of Manchester, UK  
Brian Ondov, National Library of Medicine, USA

## **Program Committee**

Sophia Ananiadou, National Centre for Text Mining and University of Manchester, UK  
Luiz Henrique Bonifacio, University of Waterloo, Canada  
Leonardo Campillos-Llanos, Spanish National Research Council, Spain  
Dina Demner-Fushman, National Library of Medicine, USA  
Manas Gaur, University of Maryland, Baltimore County, USA  
Natalia Grabar, Université de Lille, France  
Cyril Grouin, Université de Paris-Saclay, CNRS, LISN, Orsay, France  
Tudor Groza, Curtin University, Australia  
Deepak Gupta, National Library of Medicine, USA  
Anna Koroleva, Springbok AI, UK  
Alberto Lavelli, Fondazione Bruno Kessler, Italy  
Aurélie Névéal, Université de Paris-Saclay, CNRS, LISN, Orsay, France  
Brian Ondov, National Library of Medicine, USA  
Anthony Rios, University of Texas at San Antonio, USA  
Miguel Rocha, University of Minho, Portugal  
Roland Roller, German Research Center for Artificial Intelligence, DFKI, Germany  
Abeed Sarker, Emory School of Medicine, USA  
Sarvesh Soni, National Library of Medicine, USA  
Paul Thompson, National Centre for Text Mining and University of Manchester, UK  
Grigorios Tsoumakas, Aristotle University of Thessaloniki, Greece  
Aswathy Velutharambath, University of Stuttgart, Germany  
Amelie Wüthrl, University of Stuttgart, Germany  
Pierre Zweigenbaum, Université de Paris-Saclay, CNRS, LISN, Orsay, France

## **Invited Speakers**

Barbara Di Eugenio, University of Illinois Chicago, USA  
Graciela Gonzalez-Hernandez, Cedars Sinai Medical Center, USA  
Natalia Grabar, Université de Lille, France  
Abeed Sarker, Emory School of Medicine, USA

## Table of Contents

<i>Improving Sign Language Production in the Healthcare Domain Using UMLS and Multi-task Learning</i> Jonathan David Mutal, Raphael Rubino, Pierrette Bouillon, Bastien David, Johanna Gerlach and Irene Strasly .....	1
<i>It's Difficult to Be Neutral – Human and LLM-based Sentiment Annotation of Patient Comments</i> Petter Mæhlum, David Samuel, Rebecka Maria Norman, Elma Jelin, Øyvind Andresen Bjertnæs, Lilja Øvrelid and Erik Veldal .....	8
<i>Simulating Diverse Patient Populations Using Patient Vignettes and Large Language Models</i> Daniel Reichenpfader and Kerstin Denecke .....	20
<i>Annotating Emotions in Acquired Brain Injury Patients' Narratives</i> Salomé Klein, Amalia Todirascu, Hélène Vassiliadou, Marie Kuppelin, Joffrey Becart, Thalassio Briand, Clara Coridon, Francine Gerhard-Krait, Joé Laroche, Jean Ulrich and Agata Krasny-Pacini .....	26
<i>Structuring Clinical Notes of Italian ST-elevation Myocardial Infarction Patients</i> Vittorio Torri, Sara Mazzucato, Stefano Dalmiani, Umberto Paradossi, Claudio Passino, Sara Moccia, Silvestro Micera and Francesca Ieva .....	37
<i>Towards AI-supported Health Communication in Plain Language: Evaluating Intralingual Machine Translation of Medical Texts</i> Silvana Deilen, Ekaterina Lapshinova-Koltunski, Sergio Hernández Garrido, Christiane Maaß, Julian Hörner, Vanessa Theel and Sophie Ziemer .....	44
<i>Large Language Models as Drug Information Providers for Patients</i> Luca Giordano and Maria Pia di Buono .....	54
<i>Towards Generation of Personalised Health Intervention Messages</i> Clara Wan Ching Ho and Volha Petukhova .....	64
<i>Analysing Emotions in Cancer Narratives: A Corpus-Driven Approach</i> Daisy Monika Lal, Paul Rayson, Sheila A. Payne and Yufeng Liu .....	73
<i>Study of Medical Text Reading and Comprehension through Eye-Tracking Fixations</i> Oksana Ivchenko and Natalia Grabar .....	84
<i>A Neuro-Symbolic Approach to Monitoring Salt Content in Food</i> Anuja Tayal, Barbara Di Eugenio, Devika Salunke, Andrew D. Boyd, Carolyn A. Dickens, Eulalia P. Abril, Olga Garcia-Bedoya and Paula G. Allen-Meares .....	93
<i>On Simplification of Discharge Summaries in Serbian: Facing the Challenges</i> Anđelka Zečević, Milica Čulafić and Stefan Stojković .....	104
<i>Medical-FLAVORS: A Figurative Language and Vocabulary Open Repository for Spanish in the Medical Domain</i> Lucia Pitarch, Emma Angles-Herrero, Yufeng Liu, Daisy Monika Lal, Jorge Gracia, Paul Rayson and Judith Rietjens .....	109



<i>Generating Synthetic Documents with Clinical Keywords: A Privacy-Sensitive Methodology</i> Simon Meoni, Éric De la Clergerie and Théo Ryffel.....	115
<i>Building Certified Medical Chatbots: Overcoming Unstructured Data Limitations with Modular RAG</i> Leonardo Sanna, Patrizio Bellan, Simone Magnolini, Marina Segala, Saba Ghanbari Haez, Monica Consolandi and Mauro Dragoni.....	124
<i>Towards Using Automatically Enhanced Knowledge Graphs to Aid Temporal Relation Extraction</i> Timotej Knez and Slavko Žitnik.....	131
<i>Experiments in Automated Generation of Discharge Summaries in Italian</i> Lorenzo Ruinelli, Amos Colombo, Mathilde Rochat, Sotirios Georgios Popeskou, Andrea Franchini, Sandra Mitrović, Oscar William Lithgow, Joseph Cornelius and Fabio Rinaldi.....	137
<i>Evaluating LLMs for Temporal Entity Extraction from Pediatric Clinical Text in Rare Diseases Context</i> Judith Jeyafreeda Andrew, Marc Vincent, Anita Burgun and Nicolas Garcelon.....	145
<i>Generating Distributable Surrogate Corpus for Medical Multi-label Classification</i> Seiji Shimizu, Shuntaro Yada, Shoko Wakamiya and Eiji Aramaki.....	153
<i>CliniRes: Publicly Available Mapping of Clinical Lexical Resources</i> Elena Zotova, Montse Cuadros and German Rigau.....	163
<i>MedDialog-FR: A French Version of the MedDialog Corpus for Multi-label Classification and Response Generation Related to Women's Intimate Health</i> Xingyu Liu, Vincent Segonne, Aidan Mannion, Didier Schwab, Lorraine Goeuriot and François Portet.....	173
<i>Exploring the Suitability of Transformer Models to Analyse Mental Health Peer Support Forum Data for a Realist Evaluation</i> Matthew Coole, Paul Rayson, Zoe Glossop, Fiona Lobban, Paul Marshall and John Vidler.....	184
<i>Revisiting the MIMIC-IV Benchmark: Experiments Using Language Models for Electronic Health Records</i> Jesus Lovon-Melgarejo, Thouria Ben-Haddi, Jules Di Scala, Jose G. Moreno and Lynda Tamine.....	189
<i>Unraveling Clinical Insights: A Lightweight and Interpretable Approach for Multimodal and Multilingual Knowledge Integration</i> Kanimozhi Uma and Marie-Francine Moens.....	197
<i>Automated Question-Answer Generation for Evaluating RAG-based Chatbots</i> Juan José González Torres, Mihai Bogdan Bîndilă, Sebastiaan Hofstee, Daniel Szondy, Quang-Hung Nguyen, Shenghui Wang and Gwenn Englebienne.....	204
<i>Speech Accommodation in Health-Care Interactions: Evidence Using a Mixed-Reality Platform</i> Rose Baker, Susan C. Bobb, Dai'Sha Dowson, Elisha Eanes, Makyah McNeill, Hannah Ragsdale, Audrey Eaves, Joseph G. Lee and Kathrin Rothermich.....	215

<i>Enhancing Consumer Health Question Reformulation: Chain-of-Thought Prompting Integrating Focus, Type, and User Knowledge Level</i>	
Jooyeon Lee, Luan Huy Pham and Özlem Uzuner .....	220
<i>Exploring the Challenges of Behaviour Change Language Classification: A Study on Semi-Supervised Learning and the Impact of Pseudo-Labelled Data</i>	
Selina Meyer, Marcos Fernandez-Pichel, David Elsweiler and David E. Losada .....	229
<i>Development of a Benchmark Corpus for Medical Device Adverse Event Detection</i>	
Susmitha Wunnava, David A. Harris, Florence T. Bourgeois and Timothy A. Miller ....	240
<i>Using BART to Automatically Generate Discharge Summaries from Swedish Clinical Text</i>	
Nils Berg and Hercules Dalianis .....	246
<i>Biomedical Entity Linking for Dutch: Fine-tuning a Self-alignment BERT Model on an Automatically Generated Wikipedia Corpus</i>	
Fons Hartendorp, Tom Seinen, Erik van Mulligen and Suzan Verberne .....	253
<i>Unveiling Voices: Identification of Concerns in a Social Media Breast Cancer Cohort via Natural Language Processing</i>	
Swati Rajwal, Avinash Kumar Pandey, Zhishuo Han and Abeed Sarker .....	264
<i>Intent Detection and Entity Extraction from Biomedical Literature</i>	
Ankan Mullick, Mukur Gupta and Pawan Goyal .....	271

# Workshop Program

**Monday May 20, 2024**

**09:00–09:05**      **Opening remarks**

**09:05–10:30**      **Session 1: Communicating with patients**

**09:05–09:35**      ***Invited talk – Barbara Di Eugenio: Engaging the Patient in Healthcare: Summarization and Interaction***

**09:35–09:55**      *Improving Sign Language Production in the Healthcare Domain Using UMLS and Multi-task Learning*  
Jonathan David Mutal, Raphael Rubino, Pierrette Bouillon, Bastien David, Johanna Gerlach and Irene Strasly

**09:55–10:15**      *It's Difficult to Be Neutral – Human and LLM-based Sentiment Annotation of Patient Comments*  
Petter Mæhlum, David Samuel, Rebecka Maria Norman, Elma Jelin, Øyvind Andresen Bjertnæs, Lilja Øvrelid and Erik Velldal

**10:15–10:30**      ***Poster boosters***

**10:30–11:00**      ***Coffee break***

**11:00–13:00**      **Session 2: Patients' language and care**

**11:00–11:30**      ***Invited talk – Natalia Grabar: Linguistic Foundations of the Simplification and its Current State***

**11:30–11:50**      *Simulating Diverse Patient Populations Using Patient Vignettes and Large Language Models*  
Daniel Reichenpfader and Kerstin Denecke

**11:50–12:10**      *Annotating Emotions in Acquired Brain Injury Patients' Narratives*  
Salomé Klein, Amalia Todirascu, Hélène Vassiliadou, Marie Kuppelin, Jeffrey Becart, Thalassio Briand, Clara Coridon, Francine Gerhard-Krait, Joé Laroche, Jean Ulrich and Agata Krasny-Pacini

**Monday May 20, 2024 (continued)**

12:10–12:30 *Structuring Clinical Notes of Italian ST-elevation Myocardial Infarction Patients*  
Vittorio Torri, Sara Mazzucato, Stefano Dalmiani, Umberto Paradossi, Claudio Passino, Sara Moccia, Silvestro Micera and Francesca Ieva

12:30–13:00 **Poster boosters**

13:00–14:30 **Lunch**

14:30–16:30 **Poster session (parallel)**

*Towards AI-supported Health Communication in Plain Language: Evaluating Intralingual Machine Translation of Medical Texts*

Silvana Deilen, Ekaterina Lapshinova-Koltunski, Sergio Hernández Garrido, Christiane Maaß, Julian Hörner, Vanessa Theel and Sophie Ziemer

*Large Language Models as Drug Information Providers for Patients*

Luca Giordano and Maria Pia di Buono

*Towards Generation of Personalised Health Intervention Messages*

Clara Wan Ching Ho and Volha Petukhova

*Analysing Emotions in Cancer Narratives: A Corpus-Driven Approach*

Daisy Monika Lal, Paul Rayson, Sheila A. Payne and Yufeng Liu

*Study of Medical Text Reading and Comprehension through Eye-Tracking Fixations*

Oksana Ivchenko and Natalia Grabar

*A Neuro-Symbolic Approach to Monitoring Salt Content in Food*

Anuja Tayal, Barbara Di Eugenio, Devika Salunke, Andrew D. Boyd, Carolyn A. Dickens, Eulalia P. Abril, Olga Garcia-Bedoya and Paula G. Allen-Meares

*On Simplification of Discharge Summaries in Serbian: Facing the Challenges*

Anđelka Zečević, Milica Čulafić and Stefan Stojković

*Medical-FLAVORS: A Figurative Language and Vocabulary Open Repository for Spanish in the Medical Domain*

Lucia Pitarch, Emma Angles-Herrero, Yufeng Liu, Daisy Monika Lal, Jorge Gracia, Paul Rayson and Judith Rietjens

**Monday May 20, 2024 (continued)**

*Generating Synthetic Documents with Clinical Keywords: A Privacy-Sensitive Methodology*

Simon Meoni, Éric De la Clergerie and Théo Ryffel

*Building Certified Medical Chatbots: Overcoming Unstructured Data Limitations with Modular RAG*

Leonardo Sanna, Patrizio Bellan, Simone Magnolini, Marina Segala, Saba Ghanbari Haez, Monica Consolandi and Mauro Dragoni

*Towards Using Automatically Enhanced Knowledge Graphs to Aid Temporal Relation Extraction*

Timotej Knez and Slavko Žitnik

*Experiments in Automated Generation of Discharge Summaries in Italian*

Lorenzo Ruinelli, Amos Colombo, Mathilde Rochat, Sotirios Georgios Popeskou, Andrea Franchini, Sandra Mitrović, Oscar William Lithgow, Joseph Cornelius and Fabio Rinaldi

*Evaluating LLMs for Temporal Entity Extraction from Pediatric Clinical Text in Rare Diseases Context*

Judith Jeyafreeda Andrew, Marc Vincent, Anita Burgun and Nicolas Garcelon

*Generating Distributable Surrogate Corpus for Medical Multi-label Classification*

Seiji Shimizu, Shuntaro Yada, Shoko Wakamiya and Eiji Aramaki

*CliniRes: Publicly Available Mapping of Clinical Lexical Resources*

Elena Zotova, Montse Cuadros and German Rigau

*MedDialog-FR: A French Version of the MedDialog Corpus for Multi-label Classification and Response Generation Related to Women's Intimate Health*

Xingyu Liu, Vincent Segonne, Aidan Mannion, Didier Schwab, Lorraine Goeuriot and François Portet

*Exploring the Suitability of Transformer Models to Analyse Mental Health Peer Support Forum Data for a Realist Evaluation*

Matthew Coole, Paul Rayson, Zoe Glossop, Fiona Lobban, Paul Marshall and John Vidler

**Monday May 20, 2024 (continued)**

**14:30–16:30 Virtual poster session (parallel)**

*Revisiting the MIMIC-IV Benchmark: Experiments Using Language Models for Electronic Health Records*

Jesus Lovon-Melgarejo, Thouria Ben-Haddi, Jules Di Scala, Jose G. Moreno and Lynda Tamine Lechani

*Unraveling Clinical Insights: A Lightweight and Interpretable Approach for Multimodal and Multilingual Knowledge Integration*

Kanimozhi Uma and Marie-Francine Moens

*Automated Question-Answer Generation for Evaluating RAG-based Chatbots*

Juan José González Torres, Mihai Bogdan Bîndilă, Sebastiaan Hofstee, Daniel Szondy, Quang-Hung Nguyen, Shenghui Wang and Gwenn Englebienne

*Speech Accommodation in Health-Care Interactions: Evidence Using a Mixed-Reality Platform*

Rose Baker, Susan C. Bobb, Dai'Sha Dowson, Elisha Eanes, Makyah McNeill, Hannah Ragsdale, Audrey Eaves, Joseph G. Lee and Kathrin Rothermich

*Enhancing Consumer Health Question Reformulation: Chain-of-Thought Prompting Integrating Focus, Type, and User Knowledge Level*

Jooyeon Lee, Luan Huy Pham and Özlem Uzuner

*Exploring the Challenges of Behaviour Change Language Classification: A Study on Semi-Supervised Learning and the Impact of Pseudo-Labelled Data*

Selina Meyer, Marcos Fernandez-Pichel, David Elsweiler and David E. Losada

*Using BART to Automatically Generate Discharge Summaries from Swedish Clinical Text*

Nils Berg and Hercules Dalianis

*Development of a Benchmark Corpus for Medical Device Adverse Event Detection*

Susmitha Wunnava, David A. Harris, Florence T. Bourgeois and Timothy A. Miller

**16:00–16:30 Coffee break**

**Monday May 20, 2024 (continued)**

**16:30–18:00      Session 3: Social media and literature**

**16:30–17:00      *Invited talk – Abeed Sarker: Learning and Educating via NLP of Social Media: the Use Case for Substance Use and Overdose in the United States***

**17:00–17:20      *Biomedical Entity Linking for Dutch: Fine-tuning a Self-alignment BERT Model on an Automatically Generated Wikipedia Corpus***  
Fons Hartendorp, Tom Seinen, Erik van Mulligen and Suzan Verberne

**17:20–17:40      *Unveiling Voices: Identification of Concerns in a Social Media Breast Cancer Cohort via Natural Language Processing***  
Swati Rajwal, Avinash Kumar Pandey, Zhishuo Han and Abeed Sarker

**17:40–18:00      *Intent Detection and Entity Extraction from Biomedical Literature***  
Ankan Mullick, Mukur Gupta and Pawan Goyal

**18:00–18:05      Closing remarks**

# Improving Sign Language Production in the Healthcare Domain Using UMLS and Multi-Task Learning

Jonathan Mutal, Raphael Rubino, Pierrette Bouillon, Bastien David, Johanna Gerlach, Irene Strasly

TIM/FTI, University of Geneva, 1205 Geneva, Switzerland  
{firstname.lastname}@unige.ch

## Abstract

This paper presents a study on Swiss-French sign language production in the medical domain. In emergency care settings, a lack of clear communication can interfere with accurate delivery of health related services. For patients communicating with sign language, equal access to healthcare remains an issue. While previous work has explored producing sign language gloss from a source text, we propose to extend this approach to produce a multichannel sign language output given a written French input. Furthermore, we extend our approach with a multi-task framework allowing us to include the Unified Medical Language System (UMLS) in our model. Results show that the introduction of UMLS in the training data improves model accuracy by 13.64 points.

**Keywords:** sign language production, UMLS, multi-task learning, medical dialog

## 1. Introduction

In emergency care settings, there is a crucial need for automated translation tools. Emergency services often have to take care of patients who have no language in common with staff, which negatively impacts both healthcare quality and associated costs (Meischke et al., 2013). A lack of clear communication can interfere with the prompt and accurate delivery of care (Turner et al., 2019) increasing the risk of erroneous diagnoses and serious consequences (Flores et al., 2003). This is particularly true for deaf people accessing healthcare services (Ji et al., 2023).

According to Kerremans et al. (2018), various bridging solutions are currently used by medical services. They mention the use of professional or ad hoc interpreters, as well as plain language, gestures, communication technologies, and visual supports such as images or pictographs. In particular, in emergency settings where interpreters are not always available, there is a growing interest in the use of translation tools to improve accessibility (Turner et al., 2019).

In this paper, we aim at developing text to Sign Language (SL) translation models, from French to Swiss-French sign language (LSF-CH), for the medical domain. The main goal of such systems is to facilitate the communication with deaf and hard-of-hearing patients in emergency settings. Due to the lack of parallel resources to train such translation models, we propose to leverage data in a relevant domain based on the Unified Medical Language System (UMLS) (Lindberg, 1990). We train translation models, combining UMLS-based data and SL as targets and French written text as source, by applying a multi-task learning approach introduced

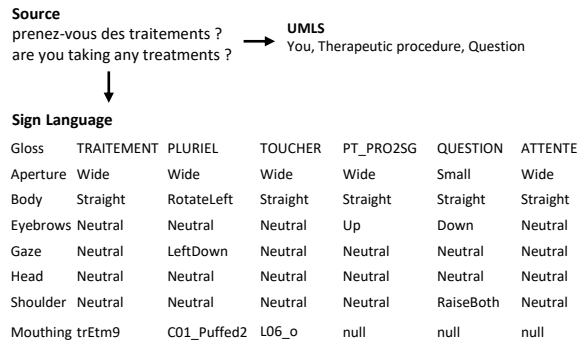


Figure 1: Example of proposed approach for multi-task training of UMLS and SL Translation.

originally for multilingual Neural Machine Translation (NMT) (Johnson et al., 2017).

The main motivation behind applying multi-task learning stems from the following research question: does a multi-task system trained on both UMLS and SL improve SL production in the medical domain compared to a mono-task system? Our hypothesis is that UMLS-based data, which is easy to create and expand due to its language independence, can be seen as a semantic pivot and can improve coverage for a low-resource target language such as LSF-CH.

The remainder of this paper is organised as follows. In Section 2, we introduce the background work and describe our approach for Sign Language Production (SLP). The methodology employed in our experiments is described in Section 3, followed by the experiments and results in Section 4. Finally, we provide an analysis of the results in Section 5 before presenting a few conclusions in Section 6.



## 2. Sign Language Production

There are three main approaches to SLP: hand-crafted animation, motion capturing and synthesis from written notation (Esselink et al., 2024). Our work focuses on synthesis from G-SiGML. G-SiGML is an XML-based representation of the physical form of signs based on Hamburg Notation System for Sign Languages (HamNoSys, Hanke, 2002). It describes both the manual (hand) and non-manual (body) features of the sign, named channels. The SiGML format allows to animate avatars. The production of animations from SiGML was first presented by Kennaway (2003) and is used in the JASigning platform (Elliott et al., 2010). Recently, it has attracted new interests, with methods to automatically convert video into SiGML (Skobov and Lepage, 2020), conversion tools into BML (Behaviour Markup Language) and integration with the new EVA avatar (Ubieto et al., 2024). Synthesis from written notation has several advantages for our context, in particular it allows fully-fledged animation of any signed form that can be described through the associated notation, without requiring video corpora or expensive equipment. Several experiments have been conducted on translating corpora to SiGML using Statistical Machine Translation and more recently using NMT (Brouer and Benabbou, 2021). However, most of them were limited to the gloss-based translation (Ebling, 2016).

In this work, we frame SLP as a machine translation task, where French serves as the source language and generates a sign table as output, as shown in Figure 1. The table represents the parallel channels of the SL output (manual activities – described as a sequence of “glosses” –, gaze, head movements, mouth movements, etc.) (Rayner et al., 2016). The table is used to generate SL in the G-SiGML format which in turn allows to animate the avatar. Creating this sign table requires both human expertise and time. Experts must have a comprehensive understanding of SL and be familiar with the formal structure of SL tables and the vocabulary. Our work aims at relieving the burden of creating new sign tables by training a joint UMLS and SL model.

## 3. Methodology

In this Section, we describe the mono and multi-task approaches employed in this paper, as well as the data used in our experiments.

### 3.1. Approaches

Two approaches were employed in our experiments, a mono-task system (noted *Mono*), trained on SL only as target, and a multi-task system (noted *Multi*),

combining UMLS and SL as targets. For the latter approach, we added a special token at the beginning of source sentences specifying which target to produce, either UMLS or SL (Johnson et al., 2017). Our rationale for this approach is to leverage parameter sharing in the decoder, aiming to enhance SLP performances, while increasing the amount of source data in French. As a comparison point, we also trained mono-task and multi-task models using the gloss channel only as target, instead of the full sign table.

### 3.2. Data

Training data for UMLS and SL are synthetic data generated from two different Synchronous Context-Free Grammars (SCFG, Aho and Ullman, 1969) which link French sentences to UMLS and sign tables (Bouillon et al., 2021).

**UMLS Data.** The UMLS grammar (Mutal et al., 2022) aims at generating parallel data which consists in French sentences (medical questions and instructions) aligned with their corresponding semantic UMLS gloss. The semantic gloss consists in an ordered sequence of concepts, combining UMLS concepts such as findings, diagnostic procedures, etc. with non-UMLS functional concepts (“You” in the example in Figure 1) or utterance modes (“Question”). The grammar has more than 3,000 rules, which expand into more than 15,000 unique UMLS sequences. These UMLS sequences are mapped to hundreds of French sentences.

**SL data.** The SL grammar generates parallel data that includes French sentences (medical questions and instructions) aligned with the corresponding SL table in LSF-CH. The sign tables were created based on human SL videos (Strasly et al., 2023). First, human video translations were created for a selected subset of sentences to develop SL reference translations for the medical terms and structures. This first set of human videos was then used as reference to productively create a larger corpus of G-SiGML from the grammar. The parallel corpus <sup>1</sup> with the human videos and their corresponding G-SiGML was used to test the comprehensibility of avatar videos in the medical domain in comparison with human videos (David et al., 2022).

## 4. Experiments and Results

This Section presents the experimental setup, including the corpora used in our experiments, the training procedure for the NMT models and the results obtained.

---

<sup>1</sup>Available at <https://yareta.unige.ch/archives/e93920a5-e5b8-47de-9979-d1fc594c068d>

Data	#Sents	#Vocab			
		FR	UMLS	SL	Gloss
UMLS	586k	4.3k	1.6k	-	-
SL	1.7m	1.0k	-	1.1k	678
Inter	5.2k	1.0k	809	1.5k	966

Table 1: Number of segments and vocabulary sizes (in thousands, denoted as “k”, or millions, denoted as “m”) for sign language (SL), UMLS-based data (UMLS), and the intersection (Inter). The vocabulary size is indicated for the source (FR) and for each target, namely UMLS-based data (UMLS), sign language tables (SL), and gloss from the sign table (Gloss).

#### 4.1. Experimental Setup

For our experiments, we used the grammars presented in Section 3.2 to generate two datasets, namely a dataset for French-SL and a dataset for French-UMLS. Prior to training the NMT models, punctuation marks on the source side were removed to be consistent between the two datasets. We transformed the SL tables into flattened sequences of column items. For the UMLS-based data, we added commas between the semantic concepts. To evaluate our models, we extracted 5,192 segments from the intersection of these two datasets. This portion of the corpus accurately represents the coverage we aim to enhance in SL translation. Finally, we extracted 3,000 segments for the validation set. Table 1 provides the segment and vocabulary counts for each dataset.

#### 4.2. Training Procedure

All the models presented in this paper are encoder-decoder models based on the Transformer architecture (Vaswani et al., 2017). We trained models from scratch with the *Marian* toolkit (Junczys-Dowmunt et al., 2018) using default parameters, except for the learning rate.<sup>2</sup> Models were trained until convergence monitored by the BLEU metric (Papineni et al., 2002) calculated on the validation set, with a patience value set to 10 (i.e. early stopping after 10 consecutive non improving validation steps). In the case of the multi-task approach, the two validation sets were used to keep the best performing models on each task.<sup>3</sup> The vocabulary size was equivalent to that of the target vocabulary for the decoder and 4,000 tokens for the encoder. The source side of the data was tokenized using BPE (Sennrich, 2017)

<sup>2</sup>The learning rate was searched within the following values:  $\{5e^{-6}, 2e^{-5}, 3e^{-5}, 3.5e^{-5}, 4e^{-5}, 3e^{-4}, 4.5e^{-4}\}$

<sup>3</sup>The models converged with high BLEU scores on the validation data, reaching 96pts BLEU for sign language.

Task	Model	BLEU $\uparrow$	chrF $\uparrow$	TER $\downarrow$	Acc $\uparrow$
SLP	<i>Mono</i>	80.43	86.47	16.45	30.41
	<i>Multi</i>	84.13*	88.61*	14.72*	44.05
Gloss	<i>Mono</i>	73.53	79.83	22.37	41.56
	<i>Multi</i>	87.09*	89.40*	13.35*	77.75

Table 2: BLEU, chrF, TER and SL table accuracy for system outputs on the test sets. Scores with \* are significantly better than previous rows with  $p < 0.01$ , calculated using paired approximate randomization with 10,000 trials.

implemented in the Sentencepiece toolkit (Kudo and Richardson, 2018), while the target sequences was divided based on spaces. We conducted all experiments employing three random seeds and averaging the results measured by the automatic metrics. This approach is intended to reduce the variability of results inherent to individual models randomly initialized.

Due to the size difference between the parallel SL and UMLS-based corpora, we over-sampled the latter 3 times to reach the size of the former. The final evaluation of our models was conducted using the following metrics: BLEU, chrF (Popović, 2015) and TER (Snover et al., 2006)<sup>4</sup>. We used paired approximate randomization with 10,000 trials to test the statistical significance of results (Riezler and Maxwell, 2005). We also measured SL table accuracy, which was calculated by comparing the SL table produced by our models to the gold reference, in order to determine how many generated full SL tables were identical to the reference.

#### 4.3. Results

Table 2 presents the test data results for all channels (SLP) and for the gloss channel only. For all channels, the model trained with UMLS (*Multi*) outperformed the model trained solely with SL (*Mono*) by 3.7pts BLEU and 13.64pts SL table accuracy. In comparison to models trained solely with the gloss channel, we observed a greater improvement with *Multi over Mono* of 13.56pts BLEU and 36.19pts accuracy. These results also show that generating the gloss channel is an easier task compared to producing the whole sign table.

Table 3 presents accuracy results by channel. We observed that the *Multi* model consistently outperformed the *Mono* model across all channels, in particular for the gloss and head channels by

<sup>4</sup>BLEU, chrF and TER were computed using the SacreBLEU 2.3.1 version of the library (Post, 2018). The signatures are: nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no

Model	Gloss/Manual	Aperture	Body	Eyebrows	Gaze	Head	Shoulder	Mouthing
<i>Mono</i>	37.54	46.48	45.80	43.07	46.21	37.31	44.34	41.30
<i>Multi</i>	52.41	54.66	54.43	50.75	52.21	49.35	50.04	46.05
<i>Gain</i>	14.87	8.18	8.63	7.68	6.00	12.04	5.70	4.75

Table 3: Accuracy for each model on the different SL channels: Gloss, Aperture, Body, Eyebrows, Gaze, Head, Shoulder and Mouthing.

14.87pts and 12.04pts increase respectively in terms of accuracy.

These results suggest that introducing UMLS in the training data is beneficial for the coverage of SL. To understand the gains of the multi-task over the mono-task on the SL task, we will delve into an analysis in the next section.

## 5. Qualitative Analysis

In this section, we perform a lexical analysis, followed by an analysis of semantic patterns, important for the domain. Finally, we comment on the non-manual channels.

### 5.1. Lexical Analysis

We compare the output of the *Mono* and *Multi* models focusing on gloss items, extracting differences at the lexical level when *Multi* output is correct while *Mono* output is incorrect. The main lexical improvement brought by the addition of UMLS during training is related to temporal markers such as *jour* (day), *aujourd'hui* (today), etc. The mono-task model fails at producing correct gloss items for these temporal terms in 800+ segments of the test set. Another large set of lexical elements correctly produced by *Multi* is related to medical terms, such as *psychose* (psychosis), *diarrhée* (diarrhea), etc. Mistakes made by *Mono* for these terms are critical as they may carry health or safety implications.

### 5.2. Pattern Analysis

The multi-task system systematically outperforms the mono-task for important patterns related to medical instructions, for example “I will prescribe you [treatment]” or “I will do an exam [scanner, radio, etc.] of [body part]”. In the mono-task version, all the translations of the pattern “I will prescribe you [...]” contain the extra gloss element PT\_PRO2SG (you, agent or patient), used for example in questions (“Do you have pain”) (see Figure 2).

### 5.3. Non-Manual Channel Analysis

The gain in BLEU for *Multi* at the level of non-manual channels is related to important SL features

source: je vais vous prescrire de l'aspirine  
*Mono*: PT\_PRO2SG ASPIRINE POUR-TOI PT\_POSS1SG PRESCRIRE ATTENTE  
*Multi*: ASPIRINE POUR-TOI PT\_POSS1SG PRESCRIRE ATTENTE  
reference: ASPIRINE POUR-TOI PT\_POSS1SG PRESCRIRE ATTENTE

Figure 2: Example of different translations in the *Mono* and *Multi* MT.

in the medical domain, for example sentiment intensification or emphasis on specific manual sign. The mono-task system has the tendency to overproduce a neutral position of the body, while the multi-task produces more variation. For instance, in “depuis combien d’années prenez-vous de l’aspirine cardio” (For how many years have you been taking cardio aspirin?), “Rotateleft” indicates that the emphasis is put on the sign for the medication (Gloss: MEDICAMENT) which becomes more visible due to rotation of the signer’s body (see Figure 3).

Gloss: MEDICAMENT COEUR\_PT\_PRO2SG TOUCHER DEPUIS ANNEE\_PL COMBIEN QUESTION ATTENTE  
Body: RotateLeft TiltBack Straight Straight TiltLeft Straight TiltForward Straight Straight

Figure 3: Example of translation in the *Multi* MT.

## 6. Conclusion

This paper presented a multi-task learning approach to translate text into sign language enhanced using domain relevant data. To the best of our knowledge, this is the first work on NMT for multi-channel sign language production in Swiss-French. Empirical results show that the introduction of UMLS-based data for training improves the generation of SL globally in terms of accuracy. In particular, the additional data improve lexical and syntactic coverage, and also have a positive impact on the non-manual channels. These results suggest that the creation and incorporation of additional UMLS data could further enhance the performance of sign language production.

Further work will explore neural architectures with dedicated decoders for SL channels, leveraging large pre-trained models as well. As a direct extension of our work, we will apply our approach to other languages, such as Simple English. Animations produced with the model outputs are currently being evaluated by deaf people.

## 7. Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments. This work is part of the PROPICTO project, funded by the Swiss National Science Foundation (N°197864) and the French National Research Agency (ANR-20-CE93-0005). All experiments were conducted on the University of Geneva computing cluster HPC *Baobab* and *Yggdrasil*.

## 8. Bibliographical References

- Alfred V. Aho and Jeffrey D. Ullman. 1969. [Syntax directed translations and the pushdown assembler](#). *Journal of Computer and System Sciences*, 3(1):37–56.
- Pierrette Bouillon, Johanna Gerlach, Jonathan Mutal, Nikos Tsourakis, and Hervé Spechbach. 2021. [A speech-enabled fixed-phrase translator for healthcare accessibility](#). In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 135–142, Online. Association for Computational Linguistics.
- Mourad Brouer and Abderrahim Benabbou. 2021. [Atlaslang nmt: Arabic text language into arabic sign language neural machine translation](#). *Journal of King Saud University-Computer and Information Sciences*, 33(9):1121–1131.
- Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. [Neural sign language translation](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7784–7793. IEEE.
- Bastien David, Jonathan David Mutal, Irene Strasly, Pierrette Bouillon, and Hervé Spechbach. 2022. [Babeldr, un système de traduction du discours médical vers l’animation virtuelle signée](#). In *Handicap 2022 - 12e conférence de l’IFRATH sur les technologies d’assistance*, pages 46–51, Paris. IFRATH.
- Mathieu De Coster and Joni Dambre. 2022. [Leveraging frozen pretrained written language models for neural sign language translation](#). *Information*, 13(5):220.
- Sarah Ebling. 2016. [Automatic Translation from German to Synthesized Swiss German Sign Language](#). Ph.D. thesis, [object Object].
- Ralph Elliott, Javier Bueno, Richard Kennaway, and John Glauert. 2010. [Towards the integration of synthetic SL animation with avatars into corpus annotation tools](#). In *Proceedings of the LREC2010 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, pages 84–87, Valletta, Malta. European Language Resources Association (ELRA).
- Lyke Esselink, Floris Roelofsen, Jakub Dotlačil, Shani Mende-Gillings, Maartje De Meulder, Nienke Sijm, and Anika Smeijers. 2024. [Exploring automatic text-to-sign translation in a health-care setting](#). *Universal Access in the Information Society*, 23(1):35–57.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Man-deep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. [Beyond english-centric multilingual machine translation](#). *Journal of Machine Learning Research*, 22(107):1–48.
- Glenn Flores, M. Barton Laws, Sandra J. Mayo, Barry Zuckerman, Milagros Abreu, Leonardo Medina, and Eric J. Hardt. 2003. [Errors in medical interpretation and their potential clinical consequences in pediatric encounters](#). *Pediatrics*, 111(1):6–14.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. [Don’t stop pretraining: Adapt language models to domains and tasks](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Thomas Hanke. 2002. [Hamnosys in a sign language generation context](#). In Rolf Schulmeister and Heimo Reinitzer, editors, *Progress in Sign Language Research / Fortschritte in der Gebärdensprachforschung. In Honor of Siegmund Prillwitz / Festschrift für Siegmund Prillwitz*, pages 249–266. Signum.
- Meng Ji, Pierrette Bouillon, and Mark Seligman. 2023. [Translation Technology in Accessible Health Communication](#), 1 edition. Cambridge University Press.
- Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. [Google’s multilingual neural machine translation system: Enabling zero-shot translation](#). *Transactions of the Association for Computational Linguistics*, 5:339–351.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann,

- Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. [Marian: Fast neural machine translation in C++](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.
- Rupinder Kaur and Parteek Kumar. 2014. Hamosys generation system for sign language. In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2727–2734. IEEE.
- Richard Kennaway. 2003. Experience with and requirements for a gesture description language for synthetic animation. In *International Gesture Workshop*, pages 300–311. Springer.
- Koen Kerremans, Laurent-Philippe De Ryck, Vanessa De Tobel, Rudi Janssens, Pascal Riloff, and Marianne Scheppers. 2018. [Bridging the communication gap in multilingual service encounters: A brussels case study](#). *The European Legacy*, 23(7–8):757–772.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Huije Lee, Jung-Ho Kim, Eui Jun Hwang, Jae-woo Kim, and Jong C Park. 2023. Leveraging large language models with vocabulary sharing for sign language translation. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 1–5. IEEE.
- C Lindberg. 1990. The unified medical language system (umls) of the national library of medicine. *Journal (American Medical Record Association)*, 61(5):40–42.
- Olga Lozynska, Maksym Davydov, Volodymyr Pasichnyk, and Nataliia Veretennikova. 2019. Rule-based machine translation into ukrainian sign language using concept dictionary. In *ICTERI*, pages 191–201.
- Hendrika W Meischke, Rebecca E Calhoun, Mei-Po Yip, Shin-Ping Tu, and Ian S Painter. 2013. The effect of language barriers on dispatching ems response. *Prehospital Emergency Care*, 17(4):475–480.
- Taro Miyazaki, Yusuke Morita, and Masanori Sano. 2020. [Machine translation from spoken language to sign language using pre-trained language model as encoder](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 139–144, Marseille, France. European Language Resources Association (ELRA).
- Boris Mocialov, Helen Hastie, and Graham Turner. 2018. [Transfer learning for British Sign Language modelling](#). In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 101–110, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jonathan Mutal, Pierrette Bouillon, Magali Norré, Johanna Gerlach, and Lucia Ormaechea Grijalba. 2022. [A neural machine translation approach to translate text to pictographs in a medical speech translation system - the BabelDr use case](#). In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 252–263, Orlando, USA. Association for Machine Translation in the Americas.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Maja Popović. 2015. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Manny Rayner, Pierrette Bouillon, Sarah Ebling, Johanna Gerlach, Irene Strasly, and Nikos Tsourakis. 2016. [An open web platform for rule-based speech-to-sign translation](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 162–168, Berlin, Germany. Association for Computational Linguistics.
- Manny Rayner, Nikos Tsourakis, and Johanna Gerlach. 2017. Lightweight spoken utterance classification with cfg, tf-idf and dynamic programming. In *Statistical Language and Speech Processing*, pages 143–154, Cham. Springer International Publishing.

- Stefan Riezler and John T. Maxwell. 2005. [On some pitfalls in automatic evaluation and significance testing for MT](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.
- Barbara C. Schouten, Antoon Cox, Gözde Duran, Koen Kerremans, Leyla Köseoğlu Banning, Ali Lahdidioui, Maria Van Den Muijsenbergh, Sanne Schinkel, Hande Sungur, Jeanine Suurmond, Rena Zendedel, and Demi Krystallidou. 2020. [Mitigating language and cultural barriers in healthcare communication: Toward a holistic approach](#). *Patient Education and Counseling*, 103(12):2604–2608.
- Rico Sennrich. 2017. [How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, page 376–382, Valencia, Spain. Association for Computational Linguistics.
- Victor Skobov and Yves Lepage. 2020. [Video-to-HamNoSys automated annotation system](#). In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, pages 209–216, Marseille, France. European Language Resources Association (ELRA).
- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.
- Irene Strasly, Pierrette Bouillon, Bastien David, and Hervé Spechbach. 2023. *Healthcare Accessibility for the Deaf*, 1 edition. Cambridge University Press.
- Shengeng Tang, Richang Hong, Dan Guo, and Meng Wang. 2022. Gloss semantic-enhanced network with online back-translation for sign language production. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5630–5638.
- Anne M Turner, Yong K Choi, Kristin Dew, Ming-Tse Tsai, Alyssa L Bosold, Shuyang Wu, Donahue Smith, and Hendrika Meischke. 2019. [Evaluating the usefulness of translation technologies for emergency response communication: A scenario-based study](#). *JMIR Public Health and Surveillance*, 5(1):e11171.
- V. Ubieto, J. Pozo, E. Valls, B. Cabrero-Daniel, and J. Blat. 2024. Sign language synthesis: Current signing avatar systems and representation. In Andy Way, Loraine Leeson, and Dimitar Shterionov, editors, *Sign Language Machine Translation*. Springer. To appear.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Kayo Yin and Jesse Read. 2020. Better sign language translation with stmc-transformer. *arXiv preprint arXiv:2004.00588*.
- Dele Zhu, Vera Czehmann, and Eleftherios Avramidis. 2023. [Neural machine translation methods for translating text to sign language glosses](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12523–12541, Toronto, Canada. Association for Computational Linguistics.

# It's Difficult to be Neutral – Human and LLM-based Sentiment Annotation of Patient Comments

Petter Mæhlum<sup>1</sup>, David Samuel<sup>1</sup>, Rebecka Maria Norman<sup>2</sup>, Elma Jelin<sup>2</sup>,  
Øyvind Andresen Bjertnæs<sup>2</sup>, Lilja Øvrelid<sup>1</sup>, Erik Velldal<sup>1</sup>

<sup>1</sup>University of Oslo, <sup>2</sup>Norwegian Institute of Public Health,

<sup>1</sup>{pettemae, davisamu, liljao, erikve}@ifi.uio.no,

<sup>2</sup>{rebecka.norman, elma.jelin,oyvindandresen.bjertnaes}@fhi.no,

## Abstract

Sentiment analysis is an important tool for aggregating patient voices, in order to provide targeted improvements in healthcare services. A prerequisite for this is the availability of in-domain data annotated for sentiment. This article documents an effort to add sentiment annotations to free-text comments in patient surveys collected by the Norwegian Institute of Public Health (NIPH). However, annotation can be a time-consuming and resource-intensive process, particularly when it requires domain expertise. We therefore also evaluate a possible alternative to human annotation, using large language models (LLMs) as annotators. We perform an extensive evaluation of the approach for two openly available pretrained LLMs for Norwegian, experimenting with different configurations of prompts and in-context learning, comparing their performance to human annotators. We find that even for zero-shot runs, models perform well above the baseline for binary sentiment, but still cannot compete with human annotators on the full dataset.

**Keywords:** patient feedback, sentiment analysis, generative models, in-context learning, annotation

## 1. Introduction

The Norwegian government has a long tradition of collecting data on patient experiences in the form of surveys, and recently this has also included of unstructured free-text comments. The application of sentiment analysis (SA) to these texts is expected to provide valuable information on patient experiences, which can then be used to improve care at both district and national levels.

This paper documents a large-scale annotation effort to add comment- and sentence-level polarity to patient feedback, representing a collaboration between NLP researchers and health professionals. Specifically, we target patient comments on experiences with General Practitioners and Special Mental Healthcare. In addition to presenting the annotation guidelines and an analysis of the resulting dataset, we also include experimental results on augmenting the human annotations with predictions by pretrained large language models (LLMs). Using two recently released generative LLMs for Norwegian based on the T5 and Mistral architectures (Raffel et al., 2020; Jiang et al., 2023), we present results for different prompts combined with zero- and few-shot learning. We also compare and discuss the differences in error types made by human annotators and the models.

Due to privacy concerns and the sensitive nature of patient feedback, the underlying text material can unfortunately not be openly distributed, but we publish the prompts, the class distributions and the annotation guidelines.<sup>1</sup>

<sup>1</sup><https://github.com/lrgoslo/Sentiment-Annotation-of-Patient-Comments/>

## 2. Background and Motivation

**Importance of patient feedback** Systematic reviews of the literature show that positive patient experiences are associated with better patient safety, better effectiveness, higher levels of adherence, and lower healthcare utilization (Anhang Price et al., 2014; Doyle et al., 2013).

An important patient-oriented tool at the national level is the national system for measuring patient experiences. The purpose of the system is to systematically measure patient experiences with health services, as a basis for quality improvement, management of health services, free patient choice, and public accountability. To underpin these goals, quantitative results from surveys are produced and presented at different health care levels, e.g., results for hospitals, for health regions, and results at the national level.

**NIPH Surveys** The Norwegian Institute of Public Health (NIPH) has conducted many national patient experience surveys. All surveys include one or more open-ended questions in which patients are encouraged to write about their experiences with the health service, which is equivalent to tens of thousands of comments each year. These comments present a rich data source on health service evaluation (Grob et al., 2019b; Rivas et al., 2019), but are mostly unused due to the time and resources needed to analyze them. NIPH's current approach is to conduct manual content analysis of a random sample of 500 comments in each survey and report main findings at the national level alongside quantitative results. Furthermore, providers

might get access to the data for their patients, but most providers lack the competence, systems and resources to analyze qualitative data. This means thousands of free text comments from each survey are excluded from further analysis and consequently also from provider-level reports. This exclusion is problematic from an ethical point of view, but also because these types of data at lower levels are highly valued by providers (Riiskjær et al., 2012) and are well-suited for use in quality improvement (Grob et al., 2019b). Free-text comments from surveys have the potential to nuance the quantitative data. For instance, a substantial proportion of patients with the highest quantitative scores describe negative experiences in free-text comments (Iversen et al., 2014), indicating that these could be used for differentiating patients at the higher end of the scale. Thus, one can expect added value of quantitative indicators at the provider-level based on the qualitative feedback of text comments.

Therefore, there is a clear need for an innovative and highly efficient method for analyzing large amounts of patient comments. In this paper we describe the first steps towards the automatic analyses of Norwegian free-text comments from patients. Feedback reports with results of free text analysis at the provider-level will make them more relevant and actionable for clinicians and managers who want to improve quality (Grob et al., 2019b; Riiskjær et al., 2012), thus possibly also strengthening the patient’s voice in quality improvement. Alleviating the workload and costs associated with annotating data for these systems constitutes an important step in this direction.

### 3. Previous work

**NLP for Patient Feedback** Khanbhai et al. (2021) present a systematic review on the application of NLP and machine learning techniques to patient experience feedback. It shows that 80% of the surveyed studies applied language analysis techniques to patient feedback from social media sites followed by structured surveys. These studies include work based on both supervised and unsupervised learning for text and sentiment analysis (SA). To provide data for supervised SA, previous work relies heavily on manual classification of a subset of data by themes and sentiment (Alemi et al., 2012; Bahja and Lycett, 2016; Doing-Harris et al., 2017; Greaves et al., 2013; Hawkins et al., 2016; Huppertz and Otto, 2018; Wagland et al., 2016; Jiménez Zafra et al., 2017). In previous work, comments extracted from social media have also been analyzed using an unsupervised approach; however, free-text comments from surveys are typically analyzed using supervised machine learning (Khanbhai et al., 2021). Khanbhai et al. (2021)

discusses that comprehensive reading of all comments within the dataset remains the ‘gold standard’ method for analyzing free-text comments, and that this is currently the only way to ensure that all relevant comments are coded and analyzed, demonstrating that language analysis using an ML approach is only as good as the dataset used to inform it. Other studies recently published in this area are all examples of how NLP and SA can be used to make the information more accessible and usable in various quality improvement initiatives, for example, using dashboards, pipelines, and visualization (Alexander et al., 2022; Khanbhai et al., 2022; Rohde et al., 2022; van Buchem et al., 2022).

**Norwegian Sentiment Analysis** To our best knowledge, there has been no previous work on sentiment analysis (SA) for free-text patient feedback, or for any user-generated text in Norwegian. The bulk of previous work on Norwegian SA in general relies on the NoReC dataset of multi-domain reviews collected from various news sources (Vellidal et al., 2018). Based on an annotated subset, Øvrelid et al. (2020) have published a fine-grained SA corpus (NoReC<sub>fine</sub>) along with annotation guidelines, on which we partly base our manual annotation effort, further described in Section 4. NoReC<sub>fine</sub> contains annotations for fine-grained sentiment: annotating the opinion holder, target and polarity. In addition to this, previous work has focused on entity-level aggregation of SA annotation (Rønningstad et al., 2022), and also on improving existing models using data augmentation based on a masked language model (Kolesnichenko et al., 2023). It’s worth noting that, being written by professional reviewers, the documents in NoReC do not contain many of the features that are typical of user-generated data, as in our patient comments, apart from the obvious differences in domain.

**Zero/Few-Shot Evaluation** In this work, the main focus will be on decoder-based, or generative, language models and the evaluation of their capabilities as sentiment annotators. These types of models have been shown to perform well in zero-shot or few-shot settings (Brown et al., 2020; Wei et al., 2022) with limited annotated data. Given the possibility of accessing these models via natural language prompts, they are arguably also easier to use than traditional models, especially in consideration of health professionals who may not have programming experience. We further focus exclusively on openly available models that can be run locally and do not risk leaking of sensitive data via an API to a proprietary service. This is crucial in our data setting, where surveys are considered sensitive data.



## 4. Annotation

**Scope and Sources** Our data comes from free-text comments from the NIPH patient experience surveys described above. The surveys cover various different domains, but we focus on two subsets of data: experiences with General Practitioners (GPs) and evaluation of Special Mental Healthcare (SMH). The data from these surveys has been the focus of earlier research (Kjøllesdal et al., 2020; Iversen et al., 2022), but the use of machine learning methods on the free-text comments is new. While the domain (health-related) and the genre (user-generated text) in these surveys are quite different from the professional reviews found in the existing NoReC corpus, we adopt a similar annotation setup as that used in the NoReC<sub>fine</sub> (Øvrelid et al., 2020) annotation effort. The original annotations were done at both comment- and sentence-level, with a three-way intensity scale, together with positive and negative sentiment, as well as explicit mention of neutral sentences, indicating that there is no expression of sentiment. Rather than using the full space of distinct labels allowed by this annotation scheme, in the experiments reported later we will only use four, corresponding to positive, negative, mixed, and neutral, ignoring the intensities.

**Annotators** Annotation was performed by seven researchers in the health service. A set of annotation guidelines was devised based on NoReC<sub>fine</sub>, with some adaptations to the active domain. Certain aspects of NoReC<sub>fine</sub>, such as the distinction between various types of sentiment, and the precise delimitation of holders, targets and polar expressions were not carried over. The seven researchers annotated in rounds of 50 comments each. In addition to this, annotators received small sets of 20 comments for quality control and Inter-Annotator Agreement (IAA) calculations. These datasets were annotated without discussions, but the results were discussed in order to resolve any ambiguity or potential issues.

### 4.1. Annotation Procedure

The sentiment annotation was performed at both the comment-level and the sentence-level. The annotators marked polarity (positive, negative) and intensity (slight, standard, strong). The original texts were sentence-segmented and tokenized using Stanza (Qi et al., 2020). Manual inspection of the resulting data shows that the tool provides accurate segmentation; however, cases such as emoticons (‘: -D’, etc.) are sometimes split erroneously. Sentences containing both positive and negative sentiment were annotated separately for both polarities. Sentences containing no polarity were marked as neutral.

Type	Total	SMH	GP
Comments	2 250	1 050	1 200
Sentences per comment	3.4	3.5	3.3
Words per sentence	14.8	16.0	13.8

Table 1: Number of comments, and average number of sentences per comment and words per sentence. We see that feedback in psychiatric care tends to be longer in both measures.

Comments were generally annotated using the same set of guidelines as for sentences, but comment-level polarity was scored based on a general impression of the comment as a whole, not just an aggregate of the sentiment of the sentences it contains. Annotators placed special weight on how the actual service (GP, SMH) was evaluated when assigning labels at the comment-level. Basic comment statistics are reported in Table 1.

**Examples** The comments in the dataset vary in terms of how sentiment is expressed, and sentiment expressions can contain direct evaluations, as we see in Example (1), where we find the strong positive adjective *fantastisk* ‘fantastic’ describing hospital employees, as well as the adjective *fin* ‘nice’ describing the patient’s stay.

- (1) *Fantastiske ansatte som har gjort fantastiske employees who have made oppholdet så fint . stay.the so nice .*  
‘Wonderful employees that have made my stay so nice.’

In Example (2) on the other hand, we see a more indirect evaluation of the treatment. The suggestion that a video consultation is useful is interpreted as a slightly negative evaluation of the health service in question.

- (2) *Kanskje kunne videokonsultasjon være til Maybe could video.consultation be to nytte ved en slik livssituasjon ? use by a such life.situation ?*  
‘Perhaps a video consultation could be of use in this kind of life situation?’

The data also contain several neutral examples – typically patients reporting on their own health situation, as in Example (3), or general descriptions.

- (3) *I tillegg bruker jeg øyedråper mot In addition use I eye.drops against høyt trykk ( grønn stær ) . high pressure ( glaucoma ) .*  
‘In addition, I use eye drops for high pressure (glaucoma).’

**Further Discussions** The annotators had weekly meetings both with each other, and with researchers working on sentiment analysis, to discuss problematic cases and to update the guidelines so that they better reflect the choices taken. Some of the issues discussed include to what extent descriptive sentences can indicate the patients’ opinion, and to what extent context should be taken into account. In cases where sentiment would be ambiguous without context, annotators were asked to use the full context of the comment, and also their knowledge of the field. In some cases, apparently negative words could indicate positive sentiment, as in cases where patients note that they get diagnosed with a serious illness, perhaps indicating that it was good that it was actually diagnosed, or that it reflects well on the GP who diagnosed them correctly, and not focusing on the negative aspect of having a disease in itself.

## 4.2. Style and Variation

The data is notably different from the existing resources for Norwegian sentiment analysis. Being user-generated language, without editing, it contains some variation, both grammatical and stylistic. Certain terms pertaining to the domains, such as *fastlege* ‘general practitioner’, and *opphold* ‘stay’ are naturally very frequent. We also observe a tendency for certain polar expressions to be very frequent, with examples including *forneøyd* ‘pleased’ being among the top 20 most frequent words in the training set. A similar tendency is not observed in the NoReC<sub>fine</sub> data, where there is more variation and a weaker tendency for certain polar expressions to dominate. However, the texts notably do not contain much medical jargon, although some terms related to diseases and treatments. We believe the texts to be different from typical *medical domain* writing. The texts further reflect the society to some degree: both official Norwegian written norms are represented, Bokmål and Nynorsk. There is also some English in the dataset. Finally, some comments show indications of lexical and syntactic patterns associated with learner language, and with spoken language in general. For example, subjectless sentences are relatively common, as in [Example \(4\)](#), where the subject of the verbs *har* and *kan* has been elided.

- (4) *Har vert innlagt før og kan trygt  
Have been admitted before and can safely  
si at her burde noen gripe tak .  
say that here should some grab hold .  
‘[I] have been admitted before and can safely say  
that someone should address the problems here.’*

	POS	NEG	NEUT	MIX
Train	1 396	1 753	476	220
Test	1 396	1 755	477	220
Total	2 792	3 508	953	440

Table 2: Class distribution in the dataset for Positive, Negative, Neutral and Mixed sentences. The high similarity in counts is due to all classes being weighted when creating the splits, making sure both splits were as similar as possible.

## 4.3. Data Preparation and Splits

Although the full dataset is annotated with information about intensity, we decided not to include these attributes for the experiments in the current study due to the added complexity. Sentences containing slight, standard or strong polarity are thus labeled as only belonging to one of the two polarity classes, POS or NEG. Sentences containing both of these, in any polarity constellation, are labeled as MIX, while sentences with no polarity are labeled NEUT. The distribution of these four classes in the dataset is reported in [Table 2](#). As we can see, the class distribution is well balanced across the training and test data, which were split 50–50, motivated by the need to ensure the test set is large enough and the fact that we do not perform any training. The split was done making sure that both datasets contained a similar number of labels for each class. The sentences were randomly sampled for each class. We mainly focus our experiments on this four-class setup, because we consider this to correspond most closely to a realistic use-case and because distinguishing sentiment-containing data from neutral sentences is an important part of naturally occurring data. However, we also report results for a reduced 2-class version, discarding the MIX and NEUT examples.

## 4.4. Human Inter-Annotator Agreement

**First Round** Initially, two test rounds of annotation were performed, in order to evaluate the guidelines, the experience of the annotator, and their agreement. Following an initial set of test annotations, smaller sets of 20 comments were annotated with regular intervals to judge the annotators’ progress. Although the results from the first of these sets showed a large variance between annotators, with kappa agreement scores for annotations including intensity varying from very low at 0.21 to as high as 0.76. We note that much of the disagreements stem from the annotations of intensity. However, when considering to what extent annotators

	A1	A2	A3	A4	A5	A6
A1	1.0	0.96	0.95	0.95	0.96	0.95
A2	0.96	1.0	0.95	0.92	0.95	0.92
A3	0.95	0.95	1.0	0.95	0.95	0.93
A4	0.95	0.92	0.95	1.0	0.92	0.93
A5	0.96	0.95	0.95	0.92	1.0	0.92
A6	0.95	0.92	0.93	0.93	0.92	1.0

Table 3: Inter annotator agreement scores for positive, negative and neutral sentiment from the second round of IAA-annotations. One annotator who annotated for the project did not partake in these specific rounds.

agree on the two polarity classes and neutral sentences, ignoring intensity, the lowest kappa score was 0.62, and the highest 0.93.

**Second Round** Following the results from the first of the 20-comment annotations, the annotators discussed various specific cases of disagreement. Topics of discussion included to what extent certain expressions exhibit sentiment, and annotators were asked to give justification for their annotations, which were then discussed before consensus was made, if possible. After a series of discussions, a new IAA dataset was annotated to show to what extent these discussions affected the agreement among annotators, giving the results in Table 3, not including intensity, where we see that scores have improved markedly, showing that humans agree on which sentences are positive, negative and neutral. There is notably more disagreement for the negative polarity than for the positive. We also observe that there are often more than one annotator disagreeing for negative polarity, indicating greater variation and uncertainty for negative labels. One annotator annotated for the project, but did not partake in these two IAA-rounds.

## 5. LLM-Based Annotation

For free-text comment analysis, finding human annotators can be difficult. Not only are there few to choose from who have the necessary background knowledge to interpret the context correctly, annotation work can be expensive and time-consuming, placing additional economic strain on health services researchers or other personnel annotating data. We want to explore whether newer models could aid in this effort. Therefore, we compare the zero-shot and few-shot performance of two Norwegian LLMs with the annotations of six healthcare professionals.

### 5.1. Language Models

In particular, we evaluate predictions of one relatively small but instruction-finetuned model, CHATNORT5, and of a larger model that has not been finetuned on any downstream tasks, NORMISTRAL.

**CHATNORT5** This model is an instruction-finetuned version of `nort5-large`, an 808-million-parameter Norwegian encoder-decoder language model (Samuel et al., 2023).<sup>2</sup> By itself, NorT5 is pretrained on masked language modeling (Raffel et al., 2020), therefore, we further finetune it on instructions (via causal language modeling), to turn it into a generative model capable of predicting sentiment labels in zero-shot or few-shot settings. To make the evaluation more realistic, we train on a general set of instructions, not specifically focusing on sentiment analysis. We use a collection of Alpaca-like datasets (Taori et al., 2023) and translate them from English to Norwegian Bokmål with OPUS-MT (Tiedemann and Thottingal, 2020).<sup>3</sup> In total, we have translated 287k conversations and then finetuned the model for one epoch. One conversation consists of multiple query-response turns, and the model is trained to produce a gold response (using its decoder part) given all previous turns (provided to the encoder part).

**NORMISTRAL** In addition to the instruction-tuned model, we also test an openly available Norwegian language model called `normistral-7b-warm`.<sup>4</sup> This model has been continually pretrained from the Mistral 7B model (Jiang et al., 2023) and has been shown to perform well in zero-shot and few-shot evaluations, even without instruction-tuning.

### 5.2. Likelihood Scores

As using the generated output from a causal LLM can lead to difficulties in mapping to the gold classes, we only consider a fixed set of possible responses for each prompt. We follow Brown et al. (2020) and formulate sentiment analysis as a task of choosing one prompt completion out of a limited number of other possible completions, based on

<sup>2</sup><https://huggingface.co/ltg/nort5-large>

<sup>3</sup>UltraChat (Ding et al., 2023): <https://huggingface.co/datasets/stingning/ultrachat>, ShareGPT: <https://huggingface.co/datasets/philschmid/sharegpt-raw>, WizardLM: [https://huggingface.co/datasets/WizardLM/WizardLM\\_evol\\_instruct\\_V2\\_196k](https://huggingface.co/datasets/WizardLM/WizardLM_evol_instruct_V2_196k) and SODA (Kim et al., 2022): <https://huggingface.co/datasets/allenai/soda>.

<sup>4</sup><https://huggingface.co/norallm/normistral-7b-warm>

their *likelihood scores*. Both evaluated language models can output  $P(s_i | s_{0:i})$ , the estimated probability of producing a subword  $s_i$  given the previous subwords  $s_{0:i} = s_0 \dots s_{i-1}$ . We use this ability to test three ways of calculating a likelihood score of a completion  $c = (c_0 \dots c_n)$  given a query  $q$ :

1.  $\mathcal{L}_1(c|q) = \sum_{c_i \in c} \log P(c_i | q, c_{0:i})$ ,
2.  $\mathcal{L}_2(c|q) = 1/n \cdot \sum_{c_i \in c} \log P(c_i | q, c_{0:i})$ ,
3.  $\mathcal{L}_3(c|q) = 1/n_{\text{char}} \cdot \sum_{c_i \in c} \log P(c_i | q, c_{0:i})$ .

The first formula calculates the actual estimated log-probability of  $c$  given  $q$ ; however, in practice, this formulation tends to overestimate the likelihood of short sequences. Therefore, we try to normalize the likelihood by the length of completion  $c$  – the second formula normalizes by its number of subwords,  $n$ , and the third formula by its number of characters,  $n_{\text{char}}$ .

### 5.3. Prompting

As there is little earlier work on prompts for sentiment analysis for Norwegian, we based our initial prompts on the existing sentiment-related prompts in the FLAN dataset (Wei et al., 2022). The FLAN dataset contains four datasets: IMDB (Maas et al., 2011), Sent140 (Go et al., 2009), SST-2 Socher et al. (2013) and Yelp.<sup>5</sup> FLAN includes several sets of English SA prompts for each of these, and they were manually translated into Norwegian.

**Prompt Variation** The prompts from FLAN contain variation in terms of multiple choice variation, differences in formality, as well as different near synonyms, and the words used to refer to the text itself (*the preceding, this*). We aimed at keeping some variation, but discarded multiple choice questions and informal variants, and did not experiment with synonyms.

**Prompt Filtering** Among all Norwegian translations of prompts, only natural-sounding sentences were considered. As we calculate likelihoods based on a certain reply given a prompt, we also wanted to keep the number of possible replies low. Some sentences were discarded due to requiring very different replies. We also wanted to be able to compare the prompts with each other, and therefore excluded sentences that would force us to expand our number of prompts drastically. FLAN-prompts not related to sentiment classification were excluded. The resulting base prompts are shown in Table 4.

<sup>5</sup><https://course20.fast.ai/datasets.html>

**Prompt Expansion** These base prompts were then expanded to create 48 prompts. Each resulting prompt is given a 4-number code based on the kind of modification it received. The first number (1-9) indicates the base prompt from Table 4. The second number indicates whether the test set sentence comes after (1) or before (2) the prompt. The third number indicates whether the question has no mentions of any of the four classes (0), mention of positive and negative (2), or all four (4). The final number indicates whether the word *positiv* comes before (2) or after (1) the word *negativ*. We give an example of prompt 8-2-4-2 in Example (5). The id means that it is based on prompt 8, has the target sentence before the prompt, has 4 classes, and has *negativ* before *positiv*.

- (5) *Oppfatter du denne setningen som Consider you this sentence.the as positiv, negativ, blandet eller nøytral?"*  
positive, negative, mixed or neutral?  
'Do you consider this sentence positive, negative, mixed or neutral?'

**Possible Replies to Prompts** Each prompt is combined with a limited set of possible answers. Much of the variation in these replies comes from the two main classes of answers, one with the word *setningen* 'the sentence' and one containing *sentimentet* 'the sentiment', which require masculine (*setningen*) and neuter (*sentimentet*) agreement, respectively. We introduce versions of replies that have a pronoun (*den*, 'it') instead, and finally versions where only the class is mentioned in the reply. In total, there are 16 alternatives per prompt, of which two have no difference depending on grammatical gender, giving 30 different replies. Each possible reply was associated with its suitable prompt data. A prompt file containing all possible prompts with all possible replies for each prompt was the basis of our experiments.

Examples 6, 7 and 8 show how the answers to a prompt vary in syntactic structure and content. Example 6 shows a full sentence referring back to the sentence (or sentiment), while in 7 it is substituted by a pronoun, and finally in 8 both the pronoun and the verb is elided to provide a minimal answer.

- (6) *Setningen er positiv .*  
the.sentence is positive .  
'The sentence is positive.'
- (7) *Den er positiv .*  
it is positive .  
'It is positive.'
- (8) *Positiv .*  
positive .  
'Positive.'

ID	Norwegian	English
1	Hvordan er sentimentet til denne setn.?	How is the sentiment of this sent.?
2	Hva er sentimentet til denne setn.?	What is the sentiment of this sent.?
3	Hvordan vil du beskrive sentimentet til denne setn.?	How would you describe the sentiment of this sent.?
4	Beskriv sentimentet i denne setn..	Describe the sentiment in this sent..
5	Ville du sagt at denne setn. er positiv eller negativ?	Would you say that this sent. is positive or negative?
6	Vil du si at denne setn. er positiv eller negativ?	Would you say that this sent. is positive or negative?
7	Er sentimentet i denne setn. positivt eller negativt?	Is the sentiment in this sent. positive or negative?
8	Oppfatter du denne setn. som positiv eller negativ?	Do you see this sent. as positive or negative?
9	Er denne setn. positiv eller negativ?	Is this sent. positive or negative?

Table 4: The 9 base prompts, and their English translations. Note that the translations here are back-translations. sent.=sentence, setn. =setningen.

## 6. LLM Experiments and Results

We performed experiments for both zero-shot and few-shot set-ups for both CHATNORT5 and NORMISTRAL. We look at both a 4-class representation and a reduced binary representation, which corresponds roughly to cases where we would expect low and high agreement, respectively. To evaluate the binary dataset, we simply do not evaluate the model output on the neutral and mixed labels, and we limit the evaluation by only investigating the predicted likelihoods for the replies mapping to the classes POS and NEG. The results are compared to a simple bag-of-words Naive Bayes model baseline.

**Naive Bayes Baseline** We set up a simple Naive Bayes model using the Natural Language ToolKit (NLTK) Python library (Bird et al., 2009), using the entire vocabulary of the train set, removing the 20 most common words. With this baseline, we achieve a macro  $F_1$  score of 41.0 for the four-class problem, higher than the random baseline of  $F_1$  22.0 For the binary setup we achieve a quite high  $F_1$  score of 79.0 compared to the baseline of 50.0

**Experimental Setup** Given the document with all 48 prompts and 16 alternatives per prompt, we estimated the likelihoods for each sentence in the test set. For a given sentence in the test set, there are  $16 \times 3$  likelihood measures. For each of the three different likelihood scores, we selected the maximum across these 16, mapped the reply alternatives to one of the four classes, and treated that as the predicted value of that sentence.

In the 2-class setup, only responses that map to the binary classes were iterated through. We could then calculate the macro  $F_1$  score across the dataset for each likelihood, for each prompt. We use the likelihood method that gives the overall best  $F_1$  score for all prompts in the test set to evaluate which prompts we use for the few shot setup. We then select the best-performing prompts from each model.

	$\mathcal{L}$	CHATNORT5	NORMISTRAL
4-class	$\mathcal{L}_1$	39.9	<b>9.4</b>
	$\mathcal{L}_2$	40.6	2.8
	$\mathcal{L}_3$	<b>42.4</b>	2.7
2-class	$\mathcal{L}_1$	88.7	84.8
	$\mathcal{L}_2$	89.2	89.0
	$\mathcal{L}_3$	<b>89.3</b>	<b>89.1</b>

Table 5: Zero-shot results. The highest  $F_1$  scores (among different prompts) for the 4-class and 2-class evaluation using the CHATNORT5 and NORMISTRAL models, and the three formulations of likelihood scores.

### 6.1. Zero-Shot Runs

In the zero-shot setup, we provide the model with a prompt, and calculate the likelihood for the 16 possible answers, using the likelihood estimates described above. Results for zero-shot runs with the two models, for both the 4-class and 2-class results, are reported in Table 5. We observe notable differences both between the two models, and between the binary and 4-classes. The CHATNORT5 model performs much better in the 4-class setting, but this is mainly due to high scores for the negative and positive classes. We find that neutral and mixed are difficult for both models.

**CHATNORT5** We see that for CHATNORT5, it is the third likelihood that gives the best results, both in the 4-class and 2-class setup. The differences between the three values are not large, and we observe that the difference between the various prompts are far more marked. The two best overall prompts were 7-2-2 and 7-1-2, and the best 4-class prompts were 7-2-4-2 and 8-2-4. 7-2-2 was also the best binary prompt, along with 5-1-2. We see that three of these are based on prompt 7, which is originally binary but expanded to 4-class in

7-2-4-2. Regarding placement of the input sentence, in 19 of 24 pairs of sentences, having the sentence in front gives a higher macro  $F_1$ . In general, there is a problem that many prompts lead to few predictions of the neutral and mixed classes. We also see that, in terms of difference between the likelihoods, invariably, if a prompt predicts overall more positive sentiment for one likelihood, it does so for all likelihoods. It seems that prompts that work well for the positive and negative classes outperform other prompts even if they predict more neutral and mixed classes.

**NORMISTRAL** For NORMISTRAL we see that all prompts almost exclusively predict the MIXED class. Here, the difference is larger between the three likelihood scores. While the likelihood score  $\mathcal{L}_1$  gives the best 4-class score, the best binary score is obtained again by  $\mathcal{L}_3$ . While  $\mathcal{L}_3$  has no predictions outside the mixed class,  $\mathcal{L}_2$  has a single prediction outside, but  $\mathcal{L}_1$  has 201 positive, 416 negative, and 5612 neutral predictions. The weak 4-class results are somewhat surprising, and can indicate either that the current prompts are largely inadequate for this model or that the model does not understand more nuanced sentiment. However, in the binary setup, the opposite is true. We approach very high  $F_1$  scores, suggesting that weak 4-class scores are a result of inadequate prompting. Interestingly, the best-performing sentences are four-class prompts: 9-1-4 and 6-1-4.

Due to low and largely similar results for the 4-class setup, we select only four prompts for NORMISTRAL, exclusively from the binary setup, and also include prompts 2-1-0 and 6-1-2-2.

## 6.2. Few-Shot Runs

Having run zero-shot runs for both models, we use the best performing prompts in a four-shot setting. This setup consists of four pairs of query–response examples, one from each class, given to the model as context, before the test sentence we want to make a prediction for. Each example was randomly sampled from the training set, and selected separately for each sentence in the dataset. These examples are all taken from the training set. The best scores for the three likelihoods are reported in Table 6.

**Few-Shot with CHATNORT5** The results from the few-shot runs with CHATNORT5 are reported in Table 7. We observe that the  $F_1$  score drops due to the 7-based prompts almost exclusively predicting the mixed class, while the 8-based prompt favors the neutral class, but is more balanced. In the binary setup, however, we see very high scores, almost beating the zero-shot results.

	$\mathcal{L}$	CHATNORT5	NORMISTRAL
4-class	$\mathcal{L}_1$	28.6	<b>12.2</b>
	$\mathcal{L}_2$	28.6	2.7
	$\mathcal{L}_3$	28.6	2.7
2-class	$\mathcal{L}_1$	89.1	<b>84.9</b>
	$\mathcal{L}_2$	<b>89.3</b>	83.9
	$\mathcal{L}_3$	83.6	83.9

Table 6: Few-shot results. The highest  $F_1$  scores for the 4-class and 2-class, as for the zero-shot results.

Prompt ID	4-class	2-class
7-2-2	3.8 (0.2)	89.0 (0.3)
7-1-2	3.2 (0.1)	85.2 (0.3)
7-2-4-2	7.2 (0.2)	87.8 (0.4)
8-2-4	26.1 (0.3)	83.0 (0.3)
5-1-2	28.3 (0.2)	62.0 (0.4)

Table 7: Mean  $F_1$  and standard deviation for the few-shot experiments using CHATNORT5.

**Few-Shot with NORMISTRAL** The results of the four prompts from the NORMISTRAL zero-shot run are reported in Table 8. We see that NORMISTRAL also struggles with making reliable predictions in the 4-class setup, but performs well on POS and NEG, albeit not as well as CHATNORT5.

**Comparison with Baseline** We observe that for the 4-class setup, CHATNORT5 achieves similar scores to the baseline model in the zero-shot runs, while NORMISTRAL achieves notably lower scores for all runs using the 4-class setup. However, for binary sentiment, both generative models achieve higher scores than the baseline in most cases, both for zero-shot and few-shot. The low scores in the 4-class setup are surprising for both models, and we hope to investigate this in later experiments.

Prompt ID	4-class	2-class
2-1-0	6.5 (0.4)	84.4 (0.1)
6-1-4	11.8 (0.2)	76.4 (0.5)
6-1-2-2	11.8 (0.2)	72.3 (0.3)
9-1-4	11.0 (0.1)	69.5 (0.5)

Table 8: Mean  $F_1$  and standard deviations for the few-shot runs using NORMISTRAL.

Prompt ID	A1	A2	A4	A6	A7
2-1-0 (2)	0.69	0.00	0.70	0.62	0.70
7-2-2 (4)	0.46	0.32	0.52	0.49	0.55
7-2-2 (2)	0.79	0.00	0.81	0.76	0.83
9-1-4 (2)	0.81	0.00	0.77	0.75	0.83

Table 9: Annotation agreement between the best predictions and the human annotators. Annotators A3 and A5 are not represented due to lack of applicable data in the test set. Prompt 7-2-2 is tested both in the 2-class and the 4-class setup. Annotator A7 did not partake in the previous IAA rounds.

### 6.3. Comparison with Human Annotators

Due to the inability of any prompt to reliably predict the NEUT and MIX classes, we selected prompts based on the best 2-class results, along with the best four-class and binary for zero shot CHAT-NORT5. This gives us three prompts: 7-2-2, 2-1-0 and 9-1-4.

**Model-annotator IAA** We compared these three combinations of prompts with models, comparing them individually with each annotator, which have the results presented in Table 9. We treat all documents annotated by an annotator as representing that annotator, and calculate kappa scores between the model and the human annotators, like we did with the human annotators. For the full 4-class problem, we see that agreement is low, but still not as low as some of the project-initial disagreements.

**Human versus Model** While we found the  $F_1$  scores to much higher for the 2-class setup, treating the model outputs as annotations still gives us lower scores than that expected from humans familiar with the task. The obvious issue is that we struggle to get our models to reliably distinguish between neutral and polar sentences. Inspection shows that a prevalent error for both prompt 2-1-0 and 7-2-2 is to mistake positive sentences for negative, while prompt 9-1-4 has more cases where the model treats positive as negative. The most common mistake in the 4-class setup for 7-2-2 is that non-mixed sentences are classified as mixed.

## 7. Conclusion

This paper has described how free-text comments in patient surveys collected by the Norwegian Institute of Public Health have been annotated with information about sentiment. Specifically, our data comprise patient comments in Norwegian on ex-

periences with General Practitioners and Special Mental Healthcare, which we have annotated with positive/negative polarity (including intensity) on both the comment- and sentence-level. In addition to describing the annotation guidelines and presenting an analysis of the resulting dataset, we also include experimental results on augmenting the human annotations with predictions by two different open-source pretrained large language models (LLMs); CHATNORT5 and NORMISTRAL. We report results for both zero- and few-shot settings for several different prompting configurations. We find that the predictions of the LLMs are sensitive to the particular prompt used, and that the best configuration depends on the specific model. Moreover, while we find that both models perform well for the simple binary cases where sentences are either positive or negative, they both struggle with neutral and mixed-polarity examples. Our error analysis shows that the predictions of the LMMs used in this study are still inferior in quality to the human annotations for our dataset.

## 8. Limitations

**Annotator representations** Due to work load limitations, we were not able to provide an even distribution of data across human annotators. This makes the claims on some of the annotators hold less than for others.

**Intensity** While we would have liked to include intensity scores, this will have to be the subject for later research. While interesting due to being a source of disagreement in humans, and we believe that differences in the treatment of intensity might reveal further differences between humans and models, it requires more space than what we could dedicate in this paper.

**Variation** We note that there is linguistic variation in the dataset, but addressing this is outside the scope of our paper. We hope to be able to return to this to be able to better assess how user language might affect how patients' voices are analyzed using systems often trained and evaluated on normative and edited language.

## 9. Ethical Considerations

While it might be possible to get similar or even better results with certain commercial models, there are several reasons why we opt for open-source Norwegian models. First of all, these models can be run locally, also without using APIs that would require sending data to servers not cleared for storage of our data, and do not pose any conflict in

terms of privacy or potential data leakage. Secondly, models trained on data from the same area as the patients might lead to less likelihood of cultural bias affecting judgements. Finally, the models' training data are open, and it is therefore possible to investigate biases and potential problems should they arise.

## 10. Acknowledgements

We would like to thank Karoline Aasgaard, Kirsten Danielsen, Lina Harvold Ellingsen-Dalskau, Hilde Hestad Iversen and Inger Opedal Paulsrud for their annotation work. This work was supported by the Norwegian Research Council, through the project *Styrking av pasientstemmen i helsetjenesteevaluering: maskinl ring p  fritekstkommentarer fra sp rreunders kelser og nettkilder*. 'Strengthening the patient voice in health service evaluation: machine learning on free-text comments from surveys and online sources', project number 103707. The computations were performed on resources provided through Sigma2 – the national research infrastructure provider for High-Performance Computing and large-scale data storage in Norway.

## 11. Bibliographical References

- Farrokh Alemi, Manabu Torii, Laura Clementz, and David C. Aron. 2012. [Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments](#). *Quality Management in Healthcare*, 21(1).
- George Alexander, Mohammed Bahja, and Gibran Farook Butt. 2022. [Automating large-scale health care service feedback analysis: Sentiment analysis and topic modeling study](#). *JMIR Med Inform*, 10(4):e29385.
- Rebecca Anhang Price, Marc N Elliott, Alan M Zaslavsky, Ron D Hays, William G Lehrman, Lise Rybowski, Susan Edgman-Levitan, and Paul D Cleary. 2014. Examining the role of patient experience surveys in measuring health care quality. *Med Care Res Rev*, 71(5):522–554.
- Mohammed Bahja and Mark Lycett. 2016. Identifying patient experience from online resources via sentiment analysis and topic modelling. In *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, Shanghai, China.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.
- Kristina Doing-Harris, Danielle L Mowery, Chrissy Daniels, Wendy W Chapman, and Mike Conway. 2017. Understanding patient satisfaction with received healthcare services: A natural language processing approach. *AMIA Annu Symp Proc*, 2016:524–533.
- Cathal Doyle, Laura Lennox, and Derek Bell. 2013. [A systematic review of evidence on the links between patient experience and clinical safety and effectiveness](#). *BMJ Open*, 3(1).
- Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, CS224N project report, Stanford.
- Felix Greaves, Daniel Ramirez-Cano, Christopher Millett, Ara Darzi, and Liam Donaldson. 2013. [Use of sentiment analysis for capturing patient experience from free-text comments posted online](#). *J Med Internet Res*, 15(11):e239.
- Rachel Grob, Mark Schlesinger, Lacey Barre, Naomi Bardach, Tara Lagu, Dale Shaller, Andrew Parker, Steven Martino, Melissa Finucane, Jennifer Cerully, and Alina Palimaru. 2019a. [What words convey: The potential for patient narratives to inform quality improvement](#). *The Milbank Quarterly*, 97:176–227.
- Rachel Grob, Mark Schlesinger, Lacey Rose Barre, Naomi Bardach, Tara Lagu, Dale Shaller, Andrew M Parker, Steven C Martino, Melissa L Finucane, Jennifer L Cerully, and Alina Palimaru. 2019b. [What words convey: The potential for patient narratives to inform quality improvement](#). *The Milbank Quarterly*, 97(1):176–227.
- Jared B Hawkins, John S Brownstein, Gaurav Tuli, Tessa Runels, Katherine Broecker, Elaine O Nsoesie, David J McIver, Ronen Rozenblum, Adam Wright, Florence T Bourgeois, and Felix Greaves. 2016. [Measuring patient-perceived quality of care in us hospitals using twitter](#). *BMJ Quality & Safety*, 25(6):404–413.



- John W. Huppertz and Peter Otto. 2018. [Predicting hcahps scores from hospitals' social media pages: A sentiment analysis](#). *Health Care Management Review*, 43(4).
- Hilde Iversen, Øyvind Bjertnæs, and Kjersti Skudal. 2014. [Patient evaluation of hospital outcomes: An analysis of open-ended comments from extreme clusters in a national survey](#). *BMJ open*, 4:e004848.
- Hilde Hestad Iversen, Mona Haugum, and Øyvind Bjertnæs. 2022. Reliability and validity of the psychiatric inpatient patient experience questionnaire – continuous electronic measurement (PIPEQ-CEM). *BMC Health Services Research*, 22(1):897.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Salud M. Jim nez Zafra, Mar a Teresa Mart n Valdivia, Isa Maks, and Rub n Izquierdo Bevi . 2017. Analysis of patient satisfaction in Dutch and Spanish online reviews.
- Mustafa Khanbhai, Patrick Anyadi, Joshua Symons, Kelsey Flott, Ara Darzi, and Erik Mayer. 2021. [Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review](#). *BMJ Health & Care Informatics*, 28(1).
- Mustafa Khanbhai, Leigh Warren, Joshua Symons, Kelsey Flott, Stephanie Harrison-White, Dave Manton, Ara Darzi, and Erik Mayer. 2022. [Using natural language processing to understand, facilitate and maintain continuity in patient experience across transitions of care](#). *International Journal of Medical Informatics*, 157:104642.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *ArXiv*, abs/2212.10465.
- Marte Kj llesdal, Thor Indseth, Hilde Hestad Iversen, and Øyvind Bjertnæs. 2020. Patient experiences with general practice in norway: a comparison of immigrant groups and the majority population following a national survey. *BMC Health Serv Res*, 20(1):1106.
- Larisa Kolesnichenko, Erik Velldal, and Lilja  vrelid. 2023. [Word substitution with masked language models as data augmentation for sentiment analysis](#). In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 42–47, T rshavn, the Faroe Islands. Association for Computational Linguistics.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. [Learning word vectors for sentiment analysis](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.
- Petter M ehlum, Jeremy Barnes, Lilja  vrelid, and Erik Velldal. 2019. [Annotating evaluative sentences for sentiment analysis: a dataset for Norwegian](#). In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 121–130, Turku, Finland. Link ping University Electronic Press.
- Lilja  vrelid, Petter M ehlum, Jeremy Barnes, and Erik Velldal. 2020. [A fine-grained sentiment dataset for Norwegian](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5025–5033, Marseille, France. European Language Resources Association.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).
- Erik Riiskj r, Jette Ammentorp, and Poul-Erik Kofoed. 2012. The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from hospital perspective. *International Journal for Quality in Health Care*, 24(5):509–516.
- Carol Rivas, Daria Tkacz, Laurence Antao, Emmanouil Mentzakis, Margaret Gordon, Sydney Anstee, and Richard Giordano. 2019. Automated analysis of free-text comments and dashboard representations in patient experience surveys: a multimethod co-design study. *Health Services and Delivery Research*, (7.23).

- Daniela Rohde, Mona Isazad Mashinchi, Nina Rizun, Dritjon Gruda, Conor Foley, Rachel Flynn, and Adegboyega Ojo. 2022. [Generating actionable insights from free-text care experience survey data using qualitative and computational text analysis: A study protocol](#). *HRB Open Research*.
- Egil Rønningstad, Erik Velldal, and Lilja Øvrelid. 2022. [Entity-level sentiment analysis \(ELSA\): An exploratory task survey](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6773–6783, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. [NorBench – a benchmark for Norwegian language models](#). In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7.
- Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.
- Marieke M. van Buchem, Olaf M. Neve, Ilse M. J. Kant, Ewout W. Steyerberg, Hileen Boosman, and Erik F. Hensen. 2022. [Analyzing patient experiences using natural language processing: development and validation of the artificial intelligence patient reported experience measure \(ai-prem\)](#). *BMC Medical Informatics and Decision Making*, 22(1):183.
- Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. [NoReC: The Norwegian review corpus](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Richard Wagland, Alejandra Recio-Saucedo, Michael Simon, Michael Bracher, Katherine Hunt, Claire Foster, Amy Downing, Adam Glaser, and Jessica Corner. 2016. [Development and testing of a text-mining approach to analyse patients’ comments on their experiences of colorectal cancer care](#). *BMJ Quality & Safety*, 25(8):604–614.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners](#). In *Proceedings of the 10th International Conference on Learning Representations*, online.

# Simulating Diverse Patient Populations Using Patient Vignettes and Large Language Models

Kerstin Denecke  and Daniel Reichenpfader 

Institute for Patient-centered Digital Health  
Bern University of Applied Sciences, Biel, Switzerland  
{daniel.reichenpfader, kerstin.denecke}@bfh.ch

## Abstract

Ensuring equitable access to digital therapeutics (DTx) is essential to avoid healthcare inequalities in an era of increasing digitization. This requires DTx to be tested with users from diverse populations, which is often not realistic due to time and resource constraints. In this paper, we propose the use of large language models (LLMs) to simulate diverse patients. Specifically, we manually create a patient vignette that characterizes a specific population group. Variations of this vignette are used for role-prompting a commercial LLM, GPT-4, instructing the LLM to take on the role described in the patient vignette and act accordingly. We investigate if the LLM stays in its given role. To do this, we simulate a medical anamnesis interview with the role-prompted LLM and analyze its responses for compliance, coherence, correctness, containment, and clarification. Our results show that GPT-4 generates compliant, coherent and clinically valid responses, including information that is not explicitly stated in the provided patient vignette.

**Keywords:** Large Language Models, Inclusive Design, Accessibility, Patient Vignettes, Simulation

## 1. Introduction

Digital therapeutics (DTx) promise to transform patient care and outcomes (Dang et al., 2020). As these digital interventions become more widespread, it is important to ensure that their design is inclusive and accessible to diverse user groups (Rivera-Romero et al., 2022). The principle of inclusivity not only enhances the usability of DTx across different demographic groups, but also underpins the effectiveness and equity of DTx. However, testing a DTx with a broad spectrum of patients is not only time consuming, but also requires significant financial resources, limiting the scope and frequency of these essential evaluations. Furthermore, the recruitment process is inherently susceptible to selection bias, skewing the sample and potentially missing critical user needs and preferences which undermines the goal of inclusive design. Beyond, the participation of vulnerable groups often requires adaptations in the testing procedure (Peute et al., 2022).

Given these limitations, there is a need for innovative methods that can simulate a wide range of patient populations. Specifically, this article aims to explore the potential of Large Language Models (LLMs) as a tool for simulating various user groups based on patient vignettes. If LLMs are a reliable method to simulate patient populations, they could contribute to the development of more inclusive and effective DTx relying on verbal communication, such as chatbots or conversational agents.

A vignette is a short, carefully written description of a person or situation (Schoenberg and Ravdal, 2000). They are a useful tool for health education, evaluating health professionals, conducting

health research (Evans et al., 2015), and evaluating symptom checkers (Ben-Shabat et al., 2022). Benoit already investigated the ability of LLMs to generate and rewrite vignettes (Benoit, 2023). In contrast to their work, we are not interested in developing a text vignette using an LLM, but in using an LLM to simulate the patient characterized by a vignette. Campillos-Llanos et al. already created a system that simulates patients, but it is based on terminology-rich resources instead of LLMs (Campillos-Llanos et al., 2020). We assume that LLMs might have the potential to simplify the development of such a system, having recently demonstrated human-level performance on various tasks, e.g. for medical question answering (Singhal et al., 2023) or for provision of medical information (Cocci et al., 2024). LLMs can be instructed to follow a certain role (Kong et al., 2023) such as the role of a teacher, physician etc. This approach to role-prompting will be used in this paper. Specifically, we will consider the following research questions:

- Which aspects are needed to accurately simulate a patient?
- Do LLMs stay in the role defined by a vignette and answer accordingly? Do LLMs instructed to follow a role provide meaningful information that is not explicitly contained in the patient vignette?

This paper reports on the methodology and validation results based on a single patient vignette.

## 2. Methods

In order to answer the above-mentioned research questions, we follow a 5-step process, see [Figure 1](#). First, relevant aspects needed for creating patient vignettes are collected based on a selective literature review. These aspects comprise, e.g., demographic data, past medical history and current symptoms or medical problems. Second, an example for each aspect of a patient vignette is drafted.

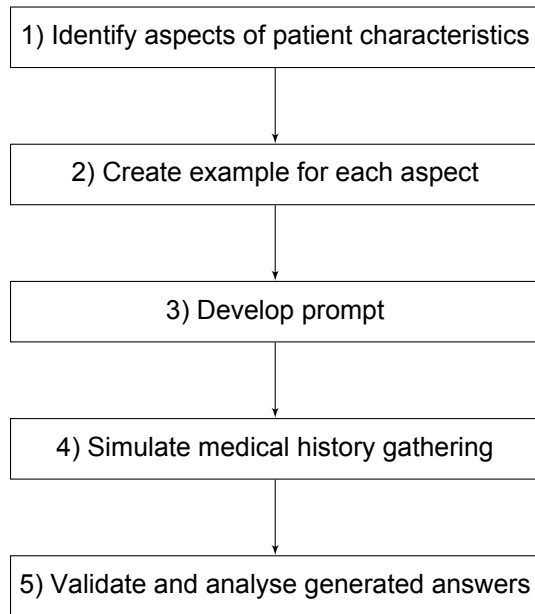


Figure 1: Methodology

Next, a prompt template is developed that instructs the LLM used for validation to impersonate the personality described within the patient vignette. The prompt development process is based on the work of Wang et al. ([Wang et al., 2024](#)) and the recommendations of OpenAI ([OpenAI, 2024](#)).

For the validation process, GPT-4 by OpenAI is used due to its accessibility and unprecedented performance. Reproducibility is ensured by implementing a Python application that executes the validation. Using the patient vignette and the prompt template from the previous steps, we will simulate a medical anamnesis interview between a physician and a patient: For this simulation, the patient is impersonated by the LLM while the medical history questions are manually entered into the system. For each interaction, the previous interactions are appended to a conversation history and included in the model input. For conducting the medical interview, we define a script based on the recommendations to conduct a medical history interview suggested by Fießl et Middeke ([Fießl and Middeke, 2022](#)).

We carry out two ablation studies, removing specific parts of the patient vignette (e.g. secondary information), in order to investigate whether the LLM is able to infer information which is not explicitly stated in the vignette. After completion of the interview, each turn of conversation is assessed according to the following assessment categories. The assessment is performed by both authors. Additionally, a qualitative analysis of generated responses is carried out.

- Compliance: The model output complies with the defined patient vignette.
- Coherence: The model output coheres with previous outputs.
- Correctness: The model output is clinically meaningful and realistic.
- Containment: The model output is explicitly contained in the patient vignette.
- Clarification: The model output contains a question asking for clarification.

## 3. Results

We make all results as well as the source code publicly available as a Git repository via [Zenodo](#) (doi:10.5281/zenodo.10889465). The total costs for prompt template development and the validation of three variants amounted to USD 3,58.

### 3.1. Patient Vignette Development

Based on six sources, we identified 16 dimensions to be included in a patient vignette, see [Table 1](#). We distinguish two categories of information: Primary information is directly asked by the health professional. Secondary information is usually not asked directly, but might have a major impact on communication: For example, Clack et al. investigate personality differences between clinicians and patients and their implications on the patient-clinician relationship. Their findings indicate that different types of personality can cause miscommunication during the consultation process ([Clack et al., 2004](#)). Redelmeier et al. review the OCEAN taxonomy, an evidence-based model to understand personalities, and state that spontaneous impressions formed by clinicians could induce incorrect clinical judgements ([Redelmeier et al., 2021](#)). Pérez-Stalbe and El-Toukhy identify factors associated with poor patient-clinician communication ([Pérez-Stalbe and El-Toukhy, 2018](#)). Bartz et al. review the role of factors related to sex and gender in healthcare ([Bartz et al., 2020](#)). Chipidza et al. give recommendations on how to evaluate and treat angry patients ([Chipidza et al., 2016](#)).

Dimension	Source
<b>Primary information</b>	
Current symptoms Past medical history Current medication Triggering factors Psychosocial aspects Family anamnesis Occupational anamnesis	(Füeßl and Middeke, 2022)
<b>Secondary information</b>	
Personality traits	(Redelmeier et al., 2021; Clack et al., 2004)
Communication style	(Clack et al., 2004)
Health literacy	(Pérez-Stable and El-Toukhy, 2018)
Race, geographic location and country of origin	(Pérez-Stable and El-Toukhy, 2018)
Sex and gender	(Bartz et al., 2020)
Emotion	(Chipidza et al., 2016)
Language proficiency	(Pérez-Stable and El-Toukhy, 2018)
Digital literacy	(Pérez-Stable and El-Toukhy, 2018)
Socioeconomic status	(Pérez-Stable and El-Toukhy, 2018)

Table 1: Dimensions of information contained in a patient vignette

Below, we show excerpts from the developed patient vignette for the dimensions *current symptoms*, *past medical history*, *health literacy* and *emotional state*. For the complete vignette, we refer to the Git repository.

- Current symptoms: *You have a headache and a fever*
- Past medical history: *You have a history of migraines*
- Health literacy: *You are very knowledgeable about your condition*
- Emotional state: *You are feeling anxious and depressed*

The following task prompt was developed iteratively and additionally self-improved by asking GPT-4 for optimisation: *Imagine that you are in the shoes of a patient during a medical consultation. You are about to engage in a detailed conversation with a healthcare provider who is taking your medical history, also known as an anamnesis. Below, you will find specific information*

*about your health, lifestyle, and medical background. Use this information to respond accurately and thoughtfully to the healthcare provider's inquiries. Remember, your role is to embody the patient's experience, drawing from the details provided. Your responses should reflect the depth and nuances of the concerns, experiences, and medical history of a real patient. <Dimensions are inserted here>. As the consultation wraps up, remember to stay true to the character and information you have been given. If the healthcare provider asks for details not explicitly mentioned, use your imagination to provide realistic and considerate answers that align with the character's background and current health scenario. Should any question seem unclear or unfamiliar based on your role as the patient, don't hesitate to ask for further clarification, just as a real patient might seek to understand their healthcare provider's inquiries fully.*

### 3.2. Anamnesis Simulation and Validation of Role-prompted LLM

The anamnesis simulation consisted of eleven questions posed to the LLM in total, see below:

1. Tell me more about your symptoms.
2. Can you give me more details regarding your headache?
3. Tell me more regarding its localization and spread.
4. Tell me more about its quality.
5. Tell me more about its severity.
6. Are you currently taking any medication?
7. Have you noticed any factors that trigger your symptoms?
8. Do you currently face difficult situations in your life?
9. Are there any diseases that run in your family?
10. What is your occupation?
11. Are you taking the pill?

In total, three variations of the patient vignette were investigated: As baseline, the complete patient vignette was used. For the first ablation study, we only kept primary information according to Table 1 and removed all secondary information. For the second ablation study, only current symptoms were retained as primary information and all secondary information was retained. The results of all three variants are summarised in Table 2. Across all three simulation variants, GPT-4 generated answers that complied with each vignette,

Variant	Compliance	Coherence	Correctness	Containment	Clarification	Average word count of model answers
Baseline	100 %	100 %	100 %	64 % (n=7/11)	0 %	53
Ablation 1	100 %	100 %	100 %	45 % (n=5/11)	0 %	41
Ablation 2	100 %	100 %	100 %	9 % (n=1/11)	0 %	78

Table 2: Validation results: For each variant, the same eleven questions were posed to the role-prompted model.

that cohered with previous answers given and that were clinically meaningful and realistic. It becomes apparent that the model makes up large proportions of its output, realistically adding information to the provided vignette. This effect is strongest in ablation 2; ten of the eleven generated responses contained information that was not included in the patient vignette. The model did not ask for clarification. Interestingly, the model used a scale from one to ten to answer the question about the intensity of symptoms. We can also see that the model tends to negate specific questions, e.g. regarding the use of oral contraception, instead of making up an answer. For example, in ablation 1 the model gave the following answer: *No, I am not currently taking any form of contraceptive pill. Other than the Tylenol for my headaches, I'm not on any other medication.* However, in case the model adds information to the provided vignette, it shows coherence when doing so: For example, the model mentioned the use of ibuprofen twice during the anamnesis simulation (ablation 2). On average, GPT-4 generates the longest answers with only minimal primary information (ablation 2) and the shortest answers when omitting secondary information (ablation 1).

#### 4. Discussion and Outlook

In this paper, we show a first approach to simulating various patient populations based on manually drafted patient vignettes. We identified 16 dimensions to be included in a patient vignette. GPT-4 generates compliant, coherent and clinically valid responses and succeeds in adding additional information not contained in the patient vignette. The role-prompted LLM comprehensively described the headache that was mentioned as symptom in the vignette. While the vignette only contained the term "headache" and, in case the medical history was included, the term "migraine", the description of pain was very detailed, even including a rating of the pain on a scale. In this sense, we can conclude that the LLM acted

well in its defined role. However, it remains open to study whether these extensions and elaborations of the symptoms are biased or follow certain stereotypes. Furthermore, it is still unknown whether more complicated vignettes reflecting complex clinical cases can still be accurately simulated.

Furthermore, it is interesting that the generated answers are longer when less primary information is provided in the vignette. Thus, when the LLM lacks a clear guidance, it fills the gaps as requested in our prompt, but with a higher risk of losing its role and adding information that does not fit accordingly. In none of the three variants, the LLM asked for clarification, although the prompt suggested this. A reason might be that the questions for medical history taking were pretty simple. However, other researchers have already found that LLMs are unable to ask for clarification and, therefore, to play a proactive role (Deng et al., 2023).

This paper reports work in progress and thus has some limitations: We conducted this study with only one patient vignette that was created by a medical informatician without clinical validation. Similarly, the assessment of generated answers was carried out by both authors who have a background in medical informatics, but no medical training. The literature considered for identifying the aspects considered in the vignette was collected in a selective non-systematic literature research and did not use a consensus-based approach. In future work, when developing more vignettes, we will follow the recommendations for vignette content provided by Evans et al. (Evans et al., 2015). Instead of inventing patient histories, synthetic patient data could be used (Guillaume et al., 2023). Furthermore, our approach is based on GPT-4, a commercial LLM. Future research might focus on investigating whether similar results can be achieved with open source LLMs such as BioMistral, a set of LLMs based on Mistral being further pre-trained on texts from PubMed Central (Labrak et al., 2024).

We highlight additional open research topics: LLMs might deny impersonating specific patient vignettes due to the practice of model alignment, where undesired or harmful behaviour is reduced during the training process. Also, mimicking certain personality features might be impossible (e.g. becoming aggressive). In this way, the approach will fail to properly simulate a patient. Furthermore, the patient vignette used for the three variants was rather short. It must be noted that the length of the vignette as well as the simulated conversation are directly proportional to model cost. This is because the costs of the commercial model are calculated on the basis of the input and output length. For each request, the entire conversation history is attached as model input, accumulating over time.

We envision two use cases for the application of the methodology tested in this paper, including educational purposes and evaluation of DTx. Simulations are used to train health professionals to act appropriately in critical situations or, generally, in patient interactions. A frequently chosen approach is to hire actors who simulate patients. With our approach, patients could be simulated by a role-prompted LLM, augmented by text-to-speech generation. The interaction could take place between the LLM and the health professional in training. For such a use case, it is less important that all the information provided is correct in a clinical sense (patients might also be inconsistent in their statements). It is more important that the main characteristics of the role are maintained, i.e. the health literacy level or cognitive abilities. Our evaluation corresponds to the general principles of simulation-based learning (Herold-Majumdar et al., 2023). In these settings, the simulated interaction takes place and is analysed afterwards. It still has to be assessed whether our approach is effective for such educational purposes.

Another potential application area is using the role-prompting-based simulation to evaluate DTx that are centred on communication. For example, conversational agents could be tested with such simulated patients. This would allow challenging the DTx with a diversity of user inputs, in different language capabilities, health literacy levels, etc. For this scenario, it still has to be clarified how role-prompted LLMs react to ambiguous or unclear input. To support this, we plan to develop a patient vignette generator where the different characteristics can be selected from a predefined list and the clinical validity of the generated patient vignette can be ensured. This vignette can then directly be used for role-prompting in an LLM. We conclude that there is potential in using LLMs together with patient vignettes to simulate interactions with pa-

tients. A more in-depth analysis is required to systematically identify potentials and limitations.

## 5. Bibliographical References

- Deborah Bartz, Tanuja Chitnis, Ursula B. Kaiser, Janet W. Rich-Edwards, Kathryn M. Rexrode, Page B. Pennell, Jill M. Goldstein, Mary Angela O'Neal, Meryl LeBoff, Maya Behn, Ellen W. Seely, Hadine Joffe, and JoAnn E. Manson. 2020. [Clinical Advances in Sex- and Gender-Informed Medicine to Improve the Health of All: A Review](#). *JAMA Internal Medicine*, 180(4):574–583.
- Niv Ben-Shabat, Gal Sharvit, Ben Meimis, Daniel Ben Joya, Ariel Sloma, David Kiderman, Aviv Shabat, Avishai M Tsur, Abdulla Watad, and Howard Amital. 2022. [Assessing data gathering of chatbot based symptom checkers—a clinical vignettes study](#). *International Journal of Medical Informatics*, 168:104897.
- James RA Benoit. 2023. [ChatGPT for Clinical Vignette Generation, Revision, and Evaluation](#). *MedRxiv*, pages 2023–02.
- Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. [Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation](#). *Natural Language Engineering*, 26(2):183–220.
- Fallon Chipidza, Rachel S. Wallwork, Traci N. Adams, and Theodore A. Stern. 2016. [Evaluation and Treatment of the Angry Patient](#). *The Primary Care Companion for CNS Disorders*, 18(3):10.4088/PCC.16f01951.
- Gillian B Clack, Judy Allen, Derek Cooper, and John O Head. 2004. [Personality differences between doctors and their patients: Implications for the teaching of communication skills](#). *Medical Education*, 38(2):177–186.
- Andrea Cocci, Marta Pezzoli, Mattia Lo Re, Giorgio Ivan Russo, Maria Giovanna Asmundo, Mikkel Fode, Giovanni Cacciamani, Sebastiano Cimino, Andrea Minervini, and Emil Durukan. 2024. [Quality of information and appropriateness of ChatGPT outputs for urology patients](#). *Prostate Cancer and Prostatic Diseases*, 27(1):103–108.
- Amit Dang, Dimple Arora, and Pawan Rane. 2020. [Role of digital therapeutics and the changing future of healthcare](#). *Journal of Family Medicine and Primary Care*, 9(5):2207–2213.

- Kerstin Denecke, Richard May, and Octavio Rivera-Romero. 2024. Transformer models in healthcare: A survey and thematic analysis of potentials, shortcomings and risks. *Journal of Medical Systems*, 48(1):23.
- Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621, Singapore. Association for Computational Linguistics.
- Spencer C Evans, Michael C Roberts, Jared W Keeley, Jennifer B Blossom, Christina M Amaro, Andrea M Garcia, Cathleen Odar Stough, Kimberly S Canter, Rebeca Robles, and Geoffrey M Reed. 2015. Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in icd-11 field studies. *International journal of clinical and health psychology*, 15(2):160–170.
- Hermann Fießl and Martin Middeke. 2022. Bestandteile der Anamnese. In *Duale Reihe Anamnese Und Klinische Untersuchung*, 7. edition. Thieme, Stuttgart.
- Marjory Gordon. 2022. *Pflegeassessment Notes*, 2. edition. Hogrefe, Bern.
- Morgan Guillaudeux, Olivia Rousseau, Julien Petot, Zineb Bennis, Charles-Axel Dein, Thomas Goronflot, Nicolas Vince, Sophie Limou, Matilde Karakachoff, Matthieu Wargny, et al. 2023. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digital Medicine*, 6(1):37.
- Astrid Dorothea Herold-Majumdar, Selina Baumann, Kathrin Hofman, Julia Kämmer, Debora Küllsen, and Valentina Müller. 2023. *Klinisches Simulationlernen in Der Pflege: Die Skills-Lab-Methode*, 1. edition. Hogrefe, Bern.
- Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. *BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains*.
- Siddika S Mulchan, Megan Miller, Christopher B Theriault, William T Zempsky, and Adam Hirsh. 2022. A systematic approach to developing virtual patient vignettes for pediatric health equity research. *Health Equity*, 6(1):862–872.
- OpenAI. 2024. Prompt Engineering. <https://platform.openai.com>.
- Eliseo J. Pérez-Stable and Sherine El-Toukhy. 2018. Communicating with diverse patients: How patient and clinician factors affect disparities. *Patient Education and Counseling*, 101(12):2186–2194.
- Linda W Peute, Gaby-Anne Wildenbos, Thomas Engelsma, Blake J Lesselroth, Valentina Lichtner, Helen Monkman, David Neal, Lex Van Velsen, Monique W Jaspers, and Romaric Marcilly. 2022. Overcoming challenges to inclusive user-based testing of health information technology with vulnerable older adults: Recommendations from a human factors engineering expert inquiry. *Yearbook of medical informatics*, 31(01):074–081.
- Donald A. Redelmeier, Umberin Najeeb, and Edward E. Etchells. 2021. Understanding Patient Personality in Medical Care: Five-Factor Model. *Journal of General Internal Medicine*, 36(7):2111–2114.
- Octavio Rivera-Romero, Elia Gabarron, Talya Miron-Shatz, Carolyn Petersen, and Kerstin Denecke. 2022. Social media, digital health literacy, and digital ethics in the light of health equity. *Yearbook of Medical Informatics*, 31(01):082–087.
- Nancy E Schoenberg and Hege Ravdal. 2000. Using vignettes in awareness and attitudinal research. *International journal of social research methodology*, 3(1):63–74.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine*, 7(1):1–9.



# Annotating Emotions in Acquired Brain Injury Patients' Narratives

Salomé Klein<sup>1</sup>, Amalia Todirascu<sup>1</sup>, Hélène Vassiliadou<sup>1</sup>, Marie Kuppelin<sup>2</sup>,  
Joffrey Becart<sup>1</sup>, Thalassio Briand<sup>1</sup>, Clara Coridon<sup>1</sup>, Francine Gerhard-Krait<sup>1</sup>, Joé  
Laroche<sup>1</sup>, Jean Ulrich<sup>1</sup>, Agata Krasny-Pacini<sup>2</sup>

University of Strasbourg, <sup>1</sup>UR 1339/LiLPa & FRLC / <sup>2</sup>INSERM UMR-S 1329

Le Portique, 14, rue René Descartes, 67084 Strasbourg Cedex (BP 80010)

{salklein, todiras, vassili, joffrey.becart, briandt, gerhard, laroche}@unistra.fr, {marie.kuppelin,  
clara.coridon, jean.ulrich}@etu.unistra.fr, agata.krasny-pacini@upecam.assurance-maladie.fr

## Abstract

In this article, we aim to measure the patients' progress in recognizing and naming emotions by capturing a variety of phenomena that express emotion in discourse. To do so, we introduce an emotion annotation scheme adapted for Acquired Brain Injury (ABI) patients' narratives. We draw on recent research outcomes in line with linguistic and psychological theories of emotion in the development of French resources for Natural Language Processing (NLP). From this perspective and following Battistelli et al. (2022) guidelines, our protocol considers several means of expressing emotions, including prototypical expressions as well as implicit means. Its originality lies on the methodology adopted for its creation, as we combined, adapted, and tested several previous annotation schemes to create a tool tailored to our spoken clinical French corpus and its unique characteristics and challenges.

**Keywords:** Emotion, Annotation, Acquired Brain Injury, Spoken French

## 1. Introduction

Corpora enriched with emotional information become increasingly important, especially in clinical linguistics. In the field of Natural Language Processing (NLP), finding a consensus on emotional categories and their expression as well as developing an automatic emotion annotation system is a widely recognized challenge (see for instance EmotionX challenge in Hsu and Ku, 2018 or, for French, the DEFT challenges 2015; 2018). In automatic emotion detection, most of the existing corpora are written. Spoken French and particularly patients' narratives have been little explored, due to the lack of available data (Amblard et al., 2020). Patients' narratives represent valuable data for doctors, linguists and NLP researchers, as they can be used for diagnosis purposes, to evaluate the effectiveness of a therapy or to detect imminent signs of a crisis. These corpora represent real challenges for NLP systems due to phenomena related to the oral channel such as disfluencies, repetitions, hesitations. Various expressions of the emotions can be found in these narratives and NLP systems fail to identify them, because of the large variability of their expressions: simple words, multi-word expressions, entire sentences.

Our corpus is composed of patients' narratives suffering from Acquired Brain Injury (ABI), who present difficulties naming their emotions. In this article, we aim to measure the patients' progress in recognizing and naming emotions by capturing a variety of phenomena that express emotions in discourse. To do so, we introduce an emotion annotation scheme to improve understanding of ABI patients' narratives. We exploit recent research outcomes in the development of French NLP resources for emotion annotation (Etienne

et al., 2022; Troiano et al., 2022; Cortal et al., 2023; Etienne, 2023), in line with linguistic and psychological theories of emotions. Following Battistelli's et al. (2022) guidelines, our protocol (§ 4) considers several means of expressing emotions, including prototypical expressions (e.g. 'I'm angry') and implicit means. Its originality lies on the methodology adopted for its creation. We combined, adapted, and tested several previous annotation schemes (§ 3.2) to create a new one adapted to our spoken clinical French corpus and its unique characteristics and challenges. The narrative corpus shows the patients' difficulties, and the need to encode the subtle and non-standard ways they use to convey them. Our scheme enables a lightweight and flexible annotation that codes the specific features of emotional expressions in spoken language, such as lexical repetitions, hesitations, non-standard idioms, and emotion expressions diluted over several sentences.

The paper is organized as follows. First, we present the context of our research (§ 2), followed by the psycholinguistic and NLP research projects used to create our annotation scheme (§ 3). We then describe our corpus, our annotation scheme, and our methodology (§ 4). Finally, we discuss the annotation campaign, the results obtained, and the drawbacks of our method (§ 5). We conclude with the future perspectives, including some solutions to the obstacles encountered (§ 6).

## 2. Context of the Study

Our study takes place within the participatory research project GRoupeRegulationEMOtion for people with acquired brain injury (GREMO; NCT 05 39 34 92 *Regulating Emotions and Behaviors After Brain Injury*) involving clinical psychology,

linguistics and NLP. GREMO is an intensive emotion regulation intervention comprising group sessions and individual psychotherapy based on Dialectical Behavior Therapy (DBT) developed by Linehan (2015). Patients with an ABI may suffer from emotional dysregulation for which evidence lacks to establish adapted treatments. Moreover, patients' alexithymia (affect disorder) and lack of insight biases the usual measures of patient-reported outcomes. One of the aims of the GREMO group is to help patients (re)learn emotional regulation skills and to associate physiological and behavioral changes, such as increased heart rate, screaming, and sweating, with specific emotional states like anger and fear. They also learn to link these changes with stereotypical situations, such as being lied to, yelled at, or abandoned.

The overall goal of the project is to explore a new objective measure of emotional regulation through linguistic and pragmatic markers, that could be used to test the effectiveness of DBT in ABI (Kuppelin and Krasny-Pacini, 2023). In addition to the other standard outcome measures, 45 patients with chronic ABI were voice-recorded for 40-90 minutes semi-directive interviews. The interviewer asks the patient to recount emotionally-charged memories or to describe the emotions they feel when shown an emotionally-inducing or a neutral image, at 3 time points 5-month apart: at the beginning of the baseline (T0, 5 months before the therapy), immediately before therapy (T1, to explore retest effects and stability of response) and immediately after therapy (T2, to explore for gains due to a 5-month intensive DBT). All patients are fluent but have pathological scores on the clinical evaluation scales: they have impaired cognitive functions and suffer from emotional regulation difficulties.

### 3. Emotion in Psycholinguistics and in NLP: A Brief Overview

We do not aim to provide an exhaustive overview of the literature on emotions and the distinctions between *emotion*, *feeling*, *affect* or *sensation* and *opinion* that vary greatly from one author to another. In the psychiatric field, Sifneos (1996: 138) recommends "for clarifications purposes, that in future studies 'affect' should be used as a general term to include 'emotion' with its somatic components and 'feeling' with its subjective experiential aspects." Some linguists suggest, on the contrary, to avoid this too conventional and technical term which does not reflect the difference between inner feelings (i.e. contingents) and behaviors or attitudes that are subject to external judgement (Polguère, 2013). In NLP, the label of *emotion* seems to prevail and is used as a generic term including opinions and attitudes (Suttles and Ide, 2013; Bostan and Klinger, 2018).

It is crucial to consider these theoretical aspects which inevitably influence annotators' decisions. Speaker's intuition regarding the difference between an attitude and an emotion strictly speaking highlights the difficulty of defining semantically vague concepts in everyday words. While we aspire obtaining a fine-grained robust description model of emotions applicable to various corpora, the paradox is that the greater the number of categories added, the lower the inter-annotator agreement (see Öhman, 2020).

#### 3.1 Emotion in Psycholinguistics

It is well known that there is no widely accepted and satisfactorily used classification of emotions. The psychologist Ekman (1992) considers that emotions are discrete, categorizable units which can be reduced to a finite number of primary or primitive emotions (fear, sadness, disgust, joy, surprise and anger). Other classifications reckon with the dimensional viewpoints, advocated by Wund (1903) and Russel (1980) in particular, who see emotions as belonging to positive/negative axes, and of high or low intensity (see Galati and Sini, 1998). This also ties in with the Appraisal Theory (Lazarus, 1991; Frijda, 2007; Ellsworth, 2013) where emotions involve an evaluation that has caused a given reaction.

Linguistically speaking, emotions have been studied in terms of their explicit expression, using idioms and terms that directly describe emotional states (e.g. 'I feel bad', 'I'm happy'; Anscombe, 1995; Flaux and Van de Velde, 2000). Emotion is understood in the sense of a physio-psychological manifestation in an affected 'place'. Nonetheless, statements conveying emotional information are not limited to those that explicitly name an affect. Micheli (2014) for instance distinguishes between explicit and implicit modes of expressing emotions. Wharton and de Saussure (2022: 670) also insist on considering "expressions that are irreducible to purely conceptual or propositional meanings" such as interjections. In the same vein, Etienne and Battistelli (2021) note two modes of expressions which do not rely on emotional labels: those arising from situations typically associated with a particular emotion, such as a funeral, a party or get slapped, and those being cues or consequences of the emotion expressed in the discourse, such as interjections, exclamatory statements and descriptions of behaviors associated with emotions. To sum up, emotions can be expressed explicitly through emotion labels, while implicit modes of expressions are inferential.

In our study, we consider that emotions stand out from other affective phenomena such as sensations, feelings, moods because of their *praxis* which underlies their categorization in language (see also the primitive "WANT to do" in Wierzbicka, 1992 or "the urge to act" in the GREMO program). The concept of *emotion* in the

repertoire of everyday language is defined by typical situations to which we react with certain affective manifestations that may be followed by typical behaviors: emotions are in a sense a type of behavior, or at least are strongly linked to behaviors/attitudes that they motivate in response to practical scenarios.

### 3.2 Emotion Annotation in NLP: Important Findings

Due to the difficulties of classifying emotions, authors of automatic emotion detection systems were from the very beginning aware of the fact that emotion annotation is not a simple emotion tag assignment. In other words, it seems impossible to just apply any annotation guidelines randomly. Existing annotation schemes consider different categories (private states, beliefs, thoughts, opinions, emotions), different emotion taxonomies, and procedures in one or two steps (word level and/or sentence level annotations) that may include or not intensity and polarity. And, most importantly, they may consider or not negation and modality. Finally, it is not always clear if context elements play a role in annotators' decisions nor if implicit means are processed.

Dimensional viewpoints and primary emotions are the most common for the modeling of lexical resources such as "Affects Lexicon" (Augustyn and Tutin, 2009), "Polarimots" (Gala and Brun, 2012) or "Diko" (Lafourcade et al., 2015). These resources generally use the Ekman's classification, the most widely accepted in NLP, which requires clear and simple categories as stated above. The use of a categorial classification is here the easiest way to implement a model for automatic emotion detection tools (Bhaumik et al., 2023; Cortal et al., 2023). Other systems create corpora taking into account the dimensional issues of the emotions proposed by the appraisal theory (Troiano et al., 2023). Besides Ekman's primary emotions, some annotation schemes add "complex emotions" (combinations of primary ones) following Turner's sociological taxonomy (2007): e.g. 'pride' as the combination of 'happiness' and 'fear' (Aman and Szpakowicz, 2007; Chen et al., 2009; Etienne and Battistelli, 2021). Plutchik's (1980) taxonomy identifies eight basic emotions: joy, fear, disgust, anger, sadness, surprise, confidence, and anticipation. This taxonomy is included in some annotation guidelines, such as in Giouli et al. (2014). But it is important to note that the use of Ekman's six primary emotions or Plutchik's taxonomy does not limit annotations to six or eight labels. Most annotation schemes associate each of them with relevant keywords (e.g. 'anger' can be associated with words such as *hate*, *dislike*, and *disgust*). Other models include several tags from the outset: Vidrascu (2007), based on Kappas et al. (1991) work, considers 20 labels of discrete emotions, Augustyn (2015) opts for 41

preestablished emotional categories, Abdul-Mageed and Ungar (2017) or Demszky et al. (2020) built a large dataset for the 24 labels of the Plutchik's wheel.

Another crucial point in emotion annotation tasks is polarity and intensity. Most of annotation schemes distinguish between positive, negative, none for polarity and low, medium, high (sometimes extreme) intensity or measured on a scale from 1 (low) to 5 (extreme).

Emotion annotation might sometimes be divided into 2 subtasks: a) emotion detection, which distinguishes between emotional and neutral content, generally at the utterance level and b) emotion classification, which assigns an emotion tag to a word (Aman and Szpakowicz, 2007; Chen et al., 2009; Giouli et al., 2014). Most projects opt though for a word-level, chunk or clause-level annotation (Vidrascu, 2007; Wiebe et al., 2005; Augustyn, 2015; Roman et al., 2015). Finally, sentences containing negation or modality are usually filtered out. As far as the distinction implicit/explicit is concerned, explicit mentions are preferred, pragmatic considerations being context dependent and thus difficult to apply by the NLP systems.

To sum up, there is no unified emotion classification in the existing NLP annotation projects, resources, or tools. Units' delimitation is also variable: word-, chunk- or sentence-level are used to annotate emotions and the adjacent properties. Some of them consider intensity or polarity, but usually negation, modality or pragmatic aspects are neglected (except for Grabar and Dumonet, 2015).

## 4. Annotation Scheme

We opted for creating our own guidelines by combining and adapting existing ones (§ 3.2 and 5.2) to focus on specific aspects of our study. The aim is not to annotate the emotions felt by the patient, nor the interviewer's discourse. Our goal is to capture the variety of ways in which a patient expresses emotions, including when they are attributed to others, denied, or modalized. The annotation is made, beyond the sentence, by trying to limit the influence of the general context and has two layers. The first is an utterance-level annotation, which allows annotators to distinguish between sentences that convey emotion from those that do not (Yes/no answers). This level also encodes information on polarity, intensity, and emotional categories, which are detailed below:

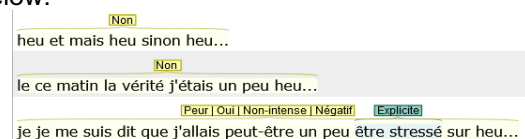


Figure 1: Sentence Layer (yellow) in INCEPTION (Klie et al., 2018)

The last sentence in Figure 1 is annotated for emotion (i.e., as conveying an emotion or not, ‘Oui/Yes’), intensity (‘non-intense’), and polarity (‘Negative’). The emotion category (‘Peur’/‘Fear’) is also tagged.

The second layer identifies the salient emotional expression in the labeled sentences (see Fig. 1 ‘être stressé’ ‘to be nervous’). Emotions are encoded as either explicit or implicit, depending on whether they are directly mentioned or suggested by cues in the discourse. Implicit marks of emotional expression can be further specified as either ‘suggested’ or ‘manifested’ (see below). The diagram in Figure 2 illustrates the different levels and types of annotation:

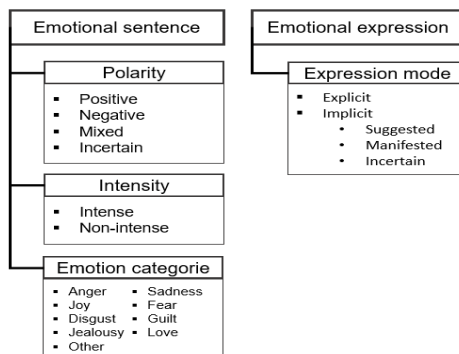


Figure 2: Annotation scheme

When a sentence expresses an emotion, it is then tagged with more specific modalities: polarity (positive, negative, mixed, uncertain), intensity (intense or non-intense), and emotional category. Sentence-level annotation of polarity considers the overall evaluation of the sentence and negation markers invert it. This layer also includes intensity, which can be intrinsic to the linguistic expression used (e.g. ‘terrified’ signals a high level of fear, ‘enraged’ a high level of anger), conveyed with modifiers (‘very angry’) or syntactic phenomena such as repetition and accumulation. The polarity and intensity features are encoded at the sentence-level to capture a range of intensifying phenomena without annotating them explicitly, thus reducing the cognitive cost for the annotator.

The properties of emotional category correspond to a fine-grained annotation of the emotion expressed by the tagged sentence. Eight emotional categories were chosen based on the DBT therapeutic manual given to patients: *anger*, *sadness*, *joy*, *fear*, *disgust*, *guilt*, *jealousy*, *love*. These emotions, their characteristics, triggers, and consequences are discussed in detail during therapy sessions. Five of them (the first listed) correspond to Ekman’s primary emotions. *Surprise* was not kept because it is considered too fleeting. The remaining three are complex, social emotions (§ 3.2). An ‘Other’ choice was also included, permitting the annotator to add new

categories, bringing the total number of categories to nine.

Moreover, the second layer distinguishes between the explicit and implicit way of expressing emotion. Explicit expressions directly designate the psychological states involved using nouns, verbs, adjectives, locutions, and multi-word expressions that refer directly to an emotion. The annotator can use a list of keywords corresponding to the emotional category based on EMOTAIX lexicon (Piolat and Bannour, 2009). To determine whether an expression falls under the explicit category of a designated emotion, we apply the lexicographical criteria cited in Etienne and Battistelli (2021): the expression should be found into a list of keywords associated with an emotion.

Although annotators were encouraged to choose from the list of emotion labels, an open-ended field was also available in case they felt it necessary to add a label (Devillers et al., 2002). We knew that there was a risk of irrelevant answers, but we thought that it was worth trying to account for the possible diversity of emotions based on speakers’ intuition. We will briefly discuss this point in section 6.2.

If the emotional triggers do not meet the above conditions, they are annotated as implicit expressions. There are two types of implicit expressions. First, ‘suggested emotions’ (Etienne and Battistelli, 2021; see also Micheli’s 2014 ‘supported’ emotions) refer to a situation, an event, or a detail which is a socio-culturally accepted reason for the source of an emotion (for example, a funeral is seen as a situation generating sadness). Then, as discussed in Etienne and Battistelli (2021), ‘displayed’ and ‘behavioral’ emotions, are expressed implicitly through actions or descriptions of actions resulting from an emotion. This kind of emotion is called in our scheme ‘manifested’ (Fig. 2). In this case, the inference is made in the opposite direction from the suggested emotion. This allows for the annotation of phenomena such as laughing, crying, or clenching one’s teeth in response to an emotional state. If a fine-grained inferential label of the implied emotion cannot be attached to either category, the annotator can tag the expression as ‘Uncertain’, allowing for the identification of more borderline expressions.

To avoid contextual influences, annotated sentences were taken out of context. Annotators were asked to start annotating transcripts from the end of the transcription to the beginning to limit subjective biases. Additionally, the number of labels for all properties has been reduced to facilitate annotators’ decisions and homogenize annotations. Decision trees were also provided:

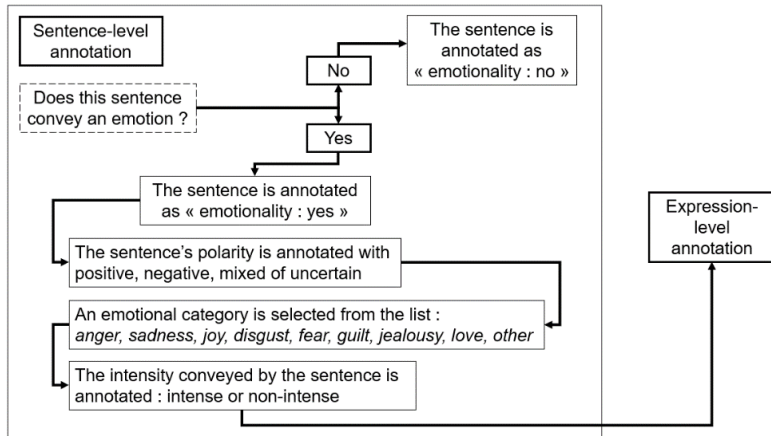


Figure 3: Decision tree for the sentence-level annotation

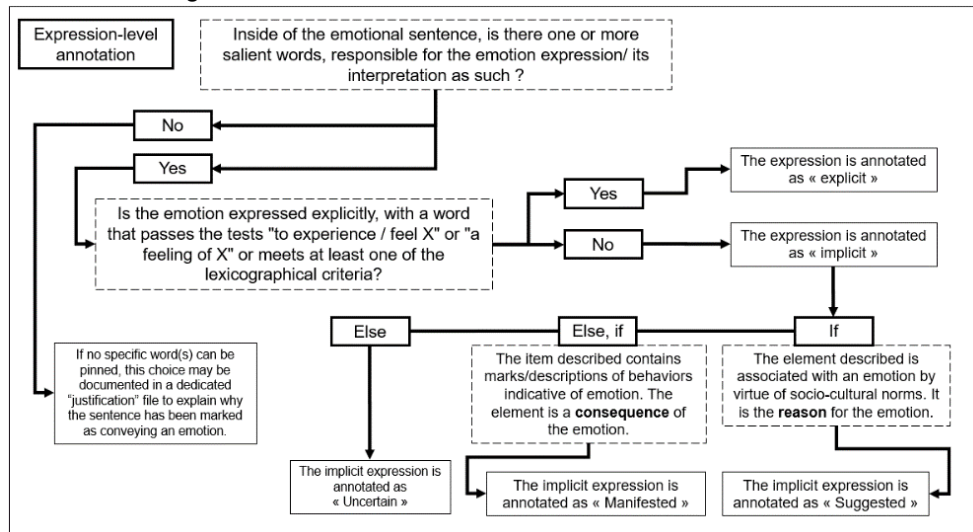


Figure 4: Decision tree for the expression-level annotation

## 5. Application of the Annotation Guidelines

The corpus was annotated by nine expert linguists. All annotators were trained to use the tool INCePTION (Klie et al., 2018) with which the campaign was conducted. Nine of the ten transcripts were annotated by three different annotators, making it easier to resolve any discrepancies in the annotations. The remaining transcript was annotated by all nine annotators and served as a control annotation to assess more accurately inter-annotators' variability, their understanding and proper application of the guidelines. The corpus was annotated in a blind manner, without knowledge of the patient or the recording time. The annotation campaign lasted 3 weeks. In general, the Emotionality task (Yes/no) was easier to perform than deciding on the categories of emotions (§ 5.2).

### 5.1 Corpus Description

The corpus contains transcriptions of interview recordings conducted with patients at the three stages of the protocol (§ 2). The recordings were transcribed following specific XML transcription guidelines that included tags for turns of speech

and disfluency phenomena such as hesitation and repetition. The corpus was then segmented into sentence units using the Whisper automatic voice recognition tool developed by OpenAI (Radford et al., 2022). This tool segments recognized sentences based on prosodic and syntactic parameters. Whisper was solely used for the segmentation step, as it did not perform well on the transcription task due to the unique features related to ABI patients' speech, such as stuttering, long pauses, and non-standard speech rate. The corpus was segmented into sentences by comparing Whisper's output with our manual transcriptions and by incorporating them where the tool added full stops. It was manually corrected when necessary.

To begin the annotation test phase, we first trained the annotators on a transcript extract. We then selected 10 transcripts, totaling 7 hours, 41 minutes and 1 second of recording time. The patients' speech consists of 58,625 tokens. The interviewers' speech is not meant to be annotated as already mentioned. The 10 transcripts include 5 pairs of recordings produced at two different times by the same person. Of these 5 pairs, 2 are control pairs, allowing us to compare

speeches at T0 and T1. The remaining 3 pairs are GREMO pairs of T1 and T2 recordings, i.e., pre- and post-therapy.

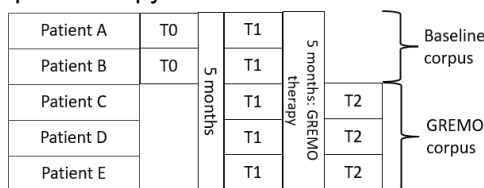


Figure 5: Corpus composition

This distribution of the transcriptions enables us to compare the annotations before and after therapy (§ 2).

## 5.2 Annotation Results

Cohen's Kappa inter-annotator agreement (Cohen, 1960) for the 'Emotionality' (Yes/No) property, which indicates whether a sentence includes emotional information or not, is ranged from 0.4 to 0.74 among all the pairs of annotators, with an average of 0.60. Thus, most annotators agreed on the emotional expression of a segment. At this stage of the study, our primary focus was to compare the ability to consistently recognize an 'emotional' sentence. Nevertheless, we can mention the following scores for the other properties annotated. The 'Polarity' feature ranges from 0.38 to 0.68 among annotator pairs, while the 'Emotional Category' ranges from 0.22 to 0.62. By way of comparison, we can mention Kim and Klingers's (2018) findings who reported agreement ranging from 0.06 to 0.40 for the annotation of 8 emotions out of 1115 sentences.

### Emotionality.

patient code	recording time	sentence number	emotional sentence	proportion of emotional
A	T1	670	260	38,81%
A	T2	477	141	29,56%
B	T1	406	179	44,09%
B	T2	324	158	48,77%
C	T1	291	131	45,02%
C	T2	396	150	37,88%
D	T0	120	27	22,50%
D	T1	291	55	18,90%
E	T0	434	180	41,47%
E	T1	473	208	43,97%
Total		3882	1489	

Figure 6: Overall results for the Emotionality feature

More than 33% of sentences uttered by the patient were annotated as containing an emotional expression. This was expected as the narration of emotional memories naturally leads to emotion expressions. Variability in the proportion of emotional sentence (ranging from 18,9% to 48,8%) reflects differences in narrative strategies among individuals. We hypothesize that this could also serve as a reliable marker for alexithymia, which will be explored in future work.

*Polarity.* Out of the 1489 annotated sentences, 445 (29.9%) were positive, 938 (62.9%) were negative, 65 (4.4%) were of mixed polarity, and 41 (2.8%) uncertain. The sentences marked as 'Uncertain' were mostly related to contexts of incomprehension (e.g. 'I may have felt a lot of emotions, but I can't remember them now') or general discussions about emotions.

*Intensity.* A total of 332 sentences (22%) of the 1489 annotated sentences in the corpus were considered as intense. It is worth noting that, for 4 out of the 5 patients, the number of sentences judged as carrying intense emotional information increased between the 1<sup>st</sup> and 2<sup>nd</sup> corpus recordings (T0-T1 or T1-T2; see section 2).

*Emotion categories.* The emotional categories assigned to the sentences in our corpus expressing emotions are distributed as follows:

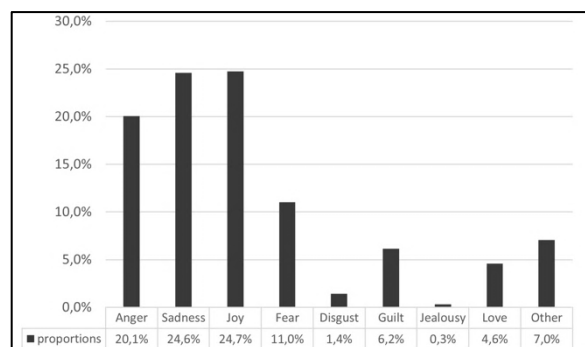


Figure 7: Distribution of the emotional categories

The categories of *disgust* and *jealousy* were rarely used for annotation, while the 'Other' category accounts for more than one sentence out of every 20. This indicates an imbalance between the emotional categories taught to patients during therapy and the concepts expressed during narration, which will be discussed in section 6.2.

*Emotion expression.* Regarding the second level of annotation for the trigger expression of emotional interpretation, we annotated 2054 expressions within the 1489 emotional sentences in our corpus. On average, there were 1.36 emotional expressions per sentence. Four out of five patients showed an increase in the proportion of emotional expressions per sentence between the first and second recording. This might suggest a higher concentration of emotional terms in their narrative, indicating an improvement in the structuring and delivery of the emotional message after the therapy.

*Expression modes.* Out of the 2054 annotations, 804 were explicit expressions of emotions, which accounts for approximately 40%. The remaining 1250 expressions were categorized as 70% suggested (886) and 30% manifested (364).

Although they learn to articulate their emotions more, as their speech is not artificial, it is governed by general pragmatic principles: for instance, it is redundant to say that a funeral is sad.

### 5.3 Related Work: Comparison with other Emotion Annotation Guidelines

Etienne's and Battistelli's guidelines (2021) suggest annotating several elements such as the experiencer, the cause and the consequence of the emotion, the affect relationship or emotional passage schema. Additionally, the guidelines offer to annotate explicit and implicit emotions (and their subtypes).

These guidelines have been selected as the basis of our annotation scheme due to the distinction explicit/implicit emotions and the variety of subtypes of implicit emotions. However, unlike Etienne et Battistelli (2021), we do not annotate the experiencer, the cause, the consequence or the affect relations, due to the spoken clinical nature of our corpus and the specific purpose of our annotation scheme. Besides, contrary to Etienne and Battistelli (2021), polarity and intensity were added at the sentence level to make it easier to take into account the negation and the intensifiers at a higher-level annotation.

Thus, three labels were added for polarity: positive, negative and uncertain. The latter was used to identify contexts in which emotion is mentioned but the narrative is not axiologically marked. Polarity tagging was based on the results reported by Wiebe et al. (2005), who include positive, negative, other and none, and by Vidrascu (2007) who tags polarity as positive, negative and unknown (see Bostan and Klinger, 2018 for an overview). Valence ("the pleasantness of the stimulus"; Warriner et al., 2013) is important when we deal with more complex emotional expressions (e.g. 'conforter'/'to comfort' suggests sadness but a positive polarity). Additionally, annotating polarity helped us capture all cases of denied or modalized emotion. By doing so, we proceeded in the opposite way to Etienne and Battistelli (2021): for a sentence like "Paul n'est pas heureux/Paul is not happy", the authors annotate 'happy' in the 'joy' category, whilst we tag the whole sentence in negative polarity (negation of a positive emotion) and annotate it as 'sadness'. Along with polarity, the intensity of the emotion was tagged in a simpler way than other guides, that is intense or not intense (Augustyn, 2015 and Roman et al., 2015 distinguish two levels of intensity: medium/high and low/non-low). This approach, which does not just rely on the lexicon, allowed for strong charged expressions to be retrieved: e.g. 'Je n'y arrive pas'/'I can't manage it', 'ce n'est pas facile'/'it's not easy', 'c'est trop pour moi'/'it's too much for me', 'Je ne peux plus le faire'/'I can't do

it anymore', 'ce n'est pas la peine'/'there is no point (in doing...)'.

Our two-step annotation is inspired by Giouli et al. (2014). The authors annotate in emotion both at the utterance-level (for an entire sentence, emotion: yes/no) and at the word or the multi-word expression-level (emotion tags). The aim is to obtain a corpus annotated in several levels, with the wider context of what we called the 'sentence\_emo', and the finer context of the expression (word or multi-word) conveying an emotion (Aman and Szpakowicz, 2007). The double level makes it possible to create a corpus of negative examples, with sentences containing no expression of emotion. In the context of a spoken corpus, automatic segmentation is based on pauses. Annotation at the utterance-level allowed us thus to remove truncated and incomplete segments from the annotation.

Finally, contrary to Etienne and Battistelli (2021), our annotation is intended to be context-free. A major contribution of our work relies on the guidelines given to annotators: as mentioned before, annotators were asked to consider each sentence separately, preferably in disorder, without taking the context into account for the analysis. As the patients' transcripts narrate memories of their lives at the first person, which are sometimes difficult and emotionally charged, it is easy for the annotator to fall into empathy and annotate contexts that are too broad because of a possible identification with the patient.

Our approach can also increase our knowledge of many lexicalized turns of phrase specific to oral expression.

## 6. Discussion

### 6.1 Results

This pilot annotation was conducted on a reduced corpus and will undergo further iterations to measure the effects of the therapy. Despite the small size of the corpus, many overall trends can be discussed.

*Emotion categories.* Joy, sadness, and anger are the dominant emotional categories (70% of the annotations). The categories of anger, sadness, fear, disgust, guilt, and jealousy represent 63.6% of the emotional labels used, which is consistent with the negative polarity distribution of our corpus at 62.9%. The remaining discrepancy is due to words like 'conforter/to comfort' or phrases like 'Je ne me suis pas fâché/I didn't get angry', in which the polarity is reversed whilst the emotional tag is not. This distribution of categories was expected, given the nature of our corpus. The themes of accident, disability, and difficulty are central in patients' discourse and usually generate negative emotions.

*Expression modes.* Approximately 60% of the emotional triggers are implicit. More than two-thirds represent suggested emotions, which are prototypical situations associated with the feeling of an emotion. Unlike Etienne et al. (2022), our categories are not equally distributed. The 'Manifested' label combines Etienne and Battistelli's 'Displayed' and 'Behavioral' categories, but it is our lowest frequent mode of expression (17%). Patients mainly use narration of situations to justify an emotion (suggested emotions at 43%) and explicitly express their feelings (40%). This may be an attempt to gain empathy from the listener. The patient may tend to focus on emotional triggers and feelings rather than actions taken in response to those emotions, such as yelling, kicking, or storming out of a room. These manifested emotions may cause the patient to feel exposed, guilty, and out of control.

## 6.2 Difficulties and Adjustments

This first annotation campaign was discussed during a feedback meeting intended to talk about the difficulties encountered by the annotators and to improve the scheme.

One of the main problems was the segmentation of units expressing emotions, especially due to the properties of spoken data, where an emotional expression can be interrupted by hesitations, repetitions, and revisions. Stricter criteria were therefore introduced, including the annotation of the light verb together with the noun of the emotion ('avoir peur/'be afraid', 'me faire peur/'scare me'), the non-annotation of verb auxiliaries and of specifiers (determiners).

The issue of annotating pragmatic markers like phatic elements has also arisen as a specific problem in our spoken speech corpora. The question is whether conversation rituals, such as 'pardon/sorry' or 'pas de problème/no problem' produced in dialogue contexts should be annotated or not as emotion expressions. We decided that if the segment is explicitly a phatic marker and not in a sentence containing an emotional expression, it should not be annotated. Indeed, words with positive or negative connotations, such as 'problem', can be found in non-emotional sentences. In the sentence 'moi qui aime bien le vélo j'ai été faire un tour' ('I like cycling, so I went for a ride'), 'aime bien' ('to like') is a positive subjective predicate, but the sentence expresses an opinion rather than an emotion and therefore it is not annotated.

One of the issues we anticipated (§ 4) was the presence of a free field for adding emotional expressions. The annotators added 32 labels, some of which had interesting aspects, but many of them were also redundant. The categories *jealousy* and *disgust* were largely underused due to misunderstanding of the labels. We changed them to 'covetousness' and 'lassitude' to make

their specificities more comprehensible to annotators. The inter-annotator agreement on Emotion Categories is quite low, due to the large number of emotions and the difficulty of choosing only one label for the segment. One annotator also used the 'Other' category to make double emotion annotations. This modification allows for the specification of 'Mixed' polarities, and it was decided to keep the possibility of double annotation, as in Etienne and Battistelli (2021). This will enable the coding of more complex and nuanced emotions, such as 'dismay', and 'anguish'. The average inter-coder agreement shows that the polarity is a demanding task, mainly because of the difficulty to capture it when negation appears.

Finally, annotators frequently encountered difficulties extracting annotations from the context due to their empathy overtaking them and despite following the instructions to annotate backwards from the text. To address this issue, a unique identifier will be added to each sentence, allowing them to be presented in a random order and limiting the contextual effect.

## 7. Conclusion and Further Work

The annotation scheme presented in this paper is a valuable resource for annotating emotions in French patient narratives, which are currently scarce. It aims at annotating emotion at sentence-level and expression-level. We provide a lightweight and flexible scheme suited for annotating non-standard language constructions. Our approach synthesizes concepts from various previous works, in line with Etienne and Battistelli (2021) and Troiano et al. (2022), creating a cohesive scheme. It is theoretically motivated, as it considers both the linguistic studies of pragmatic and semantic expression of emotion (Plantin, 2011; Micheli, 2014; Wharton and de Saussure, 2022) and the psycholinguistic aspects of it, in the context of the DBT applied to ABI patients.

Since manual annotation is a tedious and time-consuming task, we intend to use the final annotated corpora to fine-tune a pre-trained language model for automatic annotation of our corpus. These annotations will be tested as linguistic markers of the success of the DBT therapy among the recorded patients, hopefully contributing to the relevance of linguistics in the clinical context. To the best of our knowledge, this work is one of the first of its kind in annotating transcripts of patients' narratives, and as such is a pilot contribution to annotating emotion expression in spoken French.

Finally, our corpus will be used in future work for deep learning approaches or distant supervision tasks as training and validation data.



## 8. Acknowledgements

This study was supported by LiLPa (UR 1339) and the FRLC (Fédération de Recherche en Langage et Communication) and uses data from the GREMO-LCA project supported by UGECAM Alsace and by the French Eastern Interregional Group of Clinical Research and Innovation (GIRCI Est; 2021). We wish to thank EMOI-TC patients for their participation in the protocol and for sharing their emotional and behavioral difficulties in the recording.

## 9. Bibliographical References

- Abdul-Mageed, M. and Ungar, L. (2017). EmoNet: Fine-grained emotion detection with gated recurrent neural networks. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1)* 718-728.
- Aman, S., and Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. In *Proceedings of the 10th International Conference Text, Speech and Dialogue*, p. 196-205.
- Amblard, M., Braud, C., Li, C., Demily, C., Franck, N., and Musiol, M. (2020). Investigation par méthodes d'apprentissage des spécificités langagières propres aux personnes avec schizophrénie (Investigating Learning Methods Applied to Language Specificity of Persons with Schizophrenia). In C. Benzitoun, C. Braud, L. Huber, D. Langlois, S. Ouni, S. Pogodalla, and S. Schneider (Eds.), *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2: Traitement Automatique des Langues Naturelles* (p. 12-26). ATALA et AFCP.
- Anscombre, J.-C. (1995). Morphologie et représentation événementielle: Le cas des noms de sentiment et d'attitude. *Langue française*, 105(1):40-54.
- Augustyn, M. (2015). *Annotations des marques de la subjectivité langagière: Discours rapporté, passages entre guillemets et lexique des affects. Manuel de codage. Consulté à l'adresse <https://lidilem.univ-grenoble-alpes.fr/ressources/corpus/marques-subjectivite>*
- Augustyn, M., and Tutin, A. (2009). Constitution d'un corpus annoté autour du lexique des émotions: Collocations et fonctions lexicales. In D. Beck, J. Milićević, and A. Polguère (Eds.), *Actes de la Quatrième conférence internationale sur la Théorie Sens-Texte*. p.25-35.
- Battistelli, D., Étienne, A., and Lecorvé, G. (2022). L'émotion à un niveau textuel: La fonction structurante des émotions observée à partir d'annotations. *Discours*, 30.
- Bhaumik, A., Bernhardt, A., Katsios, G., Sa, N., and Strzalkowski, T. (2023). Adapting Emotion Detection to Analyze Influence Campaigns on Social Media. In J. Barnes, O. De Clercq, and R. Klinger (Eds.), *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis* (p. 441-451). Association for Computational Linguistics.
- Bostan, L.-A.-M., and Klinger, R. (2018). An Analysis of Annotated Corpora for Emotion Classification in Text. In E. M. Bender, L. Derczynski, and P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (p. 2104-2119). Association for Computational Linguistics.
- Chen, Y., Lee, S. Y. M., and Huang, C.-R. (2009). A cognitive-based annotation system for emotion computing. *Proceedings of the Third Linguistic Annotation Workshop on ACL-IJCNLP '09*, 1-9.
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37-46.
- Cortal, G., Finkel, A., Paroubek, P., and Ye, L. (2023). *Emotion Recognition based on Psychological Components in Guided Narratives for Emotion Regulation*. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL2023)* (p. 72-81), May 5, 2023, Association for Computational Linguistics
- Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (p. 4040-4054).
- Devillers, L., Rosset, S., Bonneau-Maynard, H., and Lamel, L. (2002). Annotations for Dynamic Diagnosis of the Dialog State. In M. González Rodríguez and C.P. Suarez Araujo (Eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA).
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4), 169-200.
- Ellsworth, P. C. (2013). Appraisal theory: Old and new questions. *Emotion Review*, 5(2), 125-131.

- Etienne, A. (2023). *Analyse automatique des émotions dans les textes : Contributions théoriques et applicatives dans le cadre de l'étude de la complexité des textes pour enfants* [Thèse de doctorat, Paris Nanterre].
- Etienne, A., and Battistelli, D. (2021). *Annotation manuelle des émotions dans des textes écrits avec la plateforme Glozz*. [Research Report]. MoDyCo, Université Paris Nanterre.
- Etienne, A., Battistelli, D., and Lecorvé, G. (2022). A (Psycho-)Linguistically Motivated Scheme for Annotating and Exploring Emotions in a Genre-Diverse Corpus. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (p. 603-612). European Language Resources Association.
- Flaux, N., and Van de Velde, D. (2000). *Les noms en français : Esquisse de classement*. Ophrys.
- Frijda, N. H. (2007). *The laws of emotion*. Lawrence Erlbaum Associates Publishers.
- Gala, N., and Brun, C. (2012). *Propagation de polarités dans des familles de mots : Impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions*. Actes de la conférence TALN (Traitement Automatique des Langues, Grenoble).
- Galati, D., and Sini, B. (1998). Les mots pour dire les émotions : Recherche sur la structure du lexique émotionnel italien. *Revue de Sémantique et Pragmatique*, 4, 139-161.
- Giouli, V., Fotopoulou, A., and Mouka, E. (2014). Annotating sentiment expressions for lexical resources. In P. Blumenthal, I. Novakova, and D. Siepmann (Eds.), *Les émotions dans le discours / Emotions in Discourse* (p. 281-296). Peter Lang D.
- Grabar, N., & Dumonet, L. (2015). Automatic computing of global emotional polarity in French health forum messages. In *Lecture Notes in Computer Science* (p. 243-248).
- Hsu, C.-C., and Ku, L.-W. (2018). SocialNLP 2018 EmotionX Challenge Overview: Recognizing Emotions in Dialogues. *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media* (p. 27-31).
- Kappas, A., Hess, U., and Scherer, K. R. (1991). Voice and emotion. In *Fundamentals of nonverbal behavior* (p. 200-238). Cambridge University Press.
- Kim, E., and Klinger, R. (2018). Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, (p.1345-1359). Association for Computational Linguistics.
- Klie, J.-C., Bugert, M., Boulosa, B., Eckart de Castilho, R., and Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In D. Zhao (Ed.), *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (p. 5-9). Association for Computational Linguistics.
- Kuppelin, M., and Krasny-Pacini, A. (2023). La thérapie comportementale dialectique et la régulation émotionnelle. In M. Kuppelin, I. Tavares, C. Jourdan, and A. Krasny-Pacini (Eds.), *Gérer les émotions en rééducation* (Sauramps medical).
- Lafourcade, M., Le Brun, N., and Joubert, A. (2015). Vous aimez ?... ou pas ? Likelt, un jeu pour construire une ressource lexicale de polarité. In J.-M. Lecarpentier and N. Lucas (Eds.), *Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*.
- Lazarus, R. S. (1991). Cognition and motivation in emotion. *American Psychologist*, 46(4), 352-367.
- Linehan, M. (2015). *DBT skills training manual* (Second edition). The Guilford Press.
- Micheli, R. (2014). *Les émotions dans les discours : Modèle d'analyse et perspectives empiriques*. De Boeck Duculot.
- Öhman, E. (2020). Emotion annotation: Rethinking emotion categorization. *CEUR Workshop Proceedings*, 2865 (p. 134-144).
- Piolat, A., and Bannour, R. (2009). EMOTAIX : un scénario de Tropes pour l'identification automatisée du lexique émotionnel et affectif. *L'année psychologique*, 109, 655-698.
- Plantin, C. (2011). *Les bonnes raisons des émotions*. Peter Lang CH.
- Plutchik, R., (1980). A General Psychoevolutionary Theory of Emotion. In R. Plutchik and H. Kellerman (Eds.), *Emotion, theory, research, and experience: Theories of emotions* (p. 3-33). Academic Press.
- Polguère, A. (2013). Les petits soucis ne poussent plus dans le champ lexical des sentiments. In F.H. Baider and G. Cislaru, *Cartographie des émotions : Propositions*

- linguistiques et sociolinguistiques* (p. 21-42). Presses Sorbonne Nouvelle.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision*.
- Roman, N. T., Piwek, P., Carvalho, A.M.B.R., and Alvares, A.R. (2015). Sentiment and Behaviour Annotation in a Corpus of Dialogue Summaries. *Journal of Universal Computer Science*, 21(4), 561-586.
- Russell, J.A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161-1178.
- Sifneos, P. E. (1996). Alexithymia: Past and present. *The American Journal of Psychiatry*, 153(Suppl.), 137-142.
- Suttles, J., and Ide, N. (2013). Distant supervision for emotion classification with discrete binary values. *International Conference on Intelligent Text Processing and Computational Linguistics*.
- Troiano, E., Oberlaender, L.A.M., Wegge, M., and Klinger, R. (2022). x-enVENT: A Corpus of Event Descriptions with Experiencer-specific Emotion and Appraisal Annotations. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (p. 1365-1375). European Language Resources Association.
- Troiano, E., Oberländer, L. and Klinger, R. (2023). Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, 49(1):1-72.
- Turner, J. H. (2007). *Human Emotions: A Sociological Theory*. Routledge.
- Vidrascu, L. (2007). *Analyse et détection des émotions verbales dans les interactions orales*. [Thèse de doctorat, Université Paris Sud].
- Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4), 1191-1207.
- Wharton, T., and Saussure, L. de. (2022). The pragmatics of emotion, argument and conflict. In G.L. Schiewer, J. Altarriba, and B.C. Ng (Éds.), *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science [HSK] 46/1* (p. 664-680). De Gruyter.
- Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation*, 39(2-3), 165-210.
- Wierzbicka, A. (1992). Defining emotion concepts. *Cognitive Science*, 16, 53-581.
- Wundt, W. (1903). Naturwissenschaft und Psychologie. In *Ethik: Eine Untersuchung der Thatsachen und Gesetze des sittlichen Lebens* (W. Engelmann).

# Structuring Clinical Notes of Italian ST-elevation Myocardial Infarction Patients

Vittorio Torri<sup>1</sup>, Sara Mazzucato<sup>2</sup>, Stefano Dalmiani<sup>3</sup>, Umberto Paradossi<sup>3</sup>,  
Claudio Passino<sup>3,4</sup>, Sara Moccia<sup>2</sup>, Silvestro Micera<sup>2,4,5</sup>, Francesca Ieva<sup>1,6</sup>

<sup>1</sup>MOX - Modelling and Scientific Computing Lab, Department of Mathematics,  
Politecnico di Milano, Italy  
{name.surname}@polimi.it

<sup>2</sup>Biorobotics Institute, Department of Excellence in Robotics and AI,  
Scuola Superiore Sant'Anna, Italy  
{name.surname}@santannapisa.it

<sup>3</sup>Fondazione Toscana Gabriele Monasterio, Italy  
{dalmiani, uparadossi, passino}@ftgm.it

<sup>4</sup>Health Science Interdisciplinary Center, Scuola Superiore Sant'Anna, Italy

<sup>5</sup>Bertarelli Foundation Chair in Translational Neural Engineering,  
Center for Neuroprosthetics and Institute of Bioengineering,  
Ecole Polytechnique Federale de Lausanne, Switzerland

<sup>6</sup>HDS - Health Data Science Centre, Human Technopole, Italy

## Abstract

In recent years, it has become common for patients to get full access to their Electronic Health Records (EHRs), thanks to the advancements in the EHRs systems of many healthcare providers. While this access empowers patients and doctors with comprehensive and real-time health information, it also introduces new challenges, in particular due to the unstructured nature of much of the information within EHRs. To address this, we propose a pipeline to structure anamneses, providing patients with a clear and concise overview of their health data and its longitudinal evolution, also allowing clinicians to focus more on patient care during consultations. In this paper, we present preliminary results on extracting structured information from EHRs of patients diagnosed with ST-Elevation Myocardial Infarction from an Italian hospital. Our pipeline exploits text classification models to extract relevant clinical variables, comparing rule-based, recurrent neural network and BERT-based models. While various approaches utilized ontologies or knowledge graphs for Italian data, our work represents the first attempt to develop this type of pipeline. The results for the extraction of most variables are satisfactory (f1-score > 0.80), with the exception of the most rare values of certain variables, for which we propose future research directions to investigate.

**Keywords:** Natural Language Processing, Clinical Notes, EHR Summarization, ST-elevation myocardial infarction

## 1. Introduction

In the past decades, the adoption of Electronic Health Records (EHRs) has become widespread among healthcare providers. In recent years EHRs have also granted direct access to patients, bypassing the need for a physician's mediation (Klein et al., 2016). This advancement offers patients numerous benefits, including immediate access to their latest exam results and allows them to review their medical history at their convenience, ultimately enhancing the relational aspect of care (Blease et al., 2020). However, a significant portion of EHRs is still in the format of unstructured documents, with only a fraction of their data available in structured formats (Rosenbloom et al., 2011; Tayefi et al., 2021). This lack of structure complicates the work of physicians since, despite their familiarity with such documents, they still require substantial time to extract pertinent information, consequently impacting their interactions with patients (Friedberg et al., 2014).

A potential solution lies in leveraging Natural Lan-

guage Processing (NLP), a field that has witnessed remarkable advancements in recent years. However, applying NLP models in the medical domain presents challenges due to the unique formats and terminology inherent in medical documents. Despite the attempts of the most recent models to achieve a certain level of multilingualism, the highest performances in NLP models predominantly continue to occur with English documents, which also serve as the primary focus for most benchmarks (Hedderich et al., 2021; Lai et al., 2023). The intersection of these two areas of complexity poses significant hurdles (Névéol et al., 2018).

In this study, we present preliminary findings on developing a pipeline to extract structured data from EHRs of patients diagnosed with ST-Elevation Myocardial Infarction (STEMI) at Fondazione Toscana Gabriele Monasterio (FTGM), an Italian hospital specialized in cardiology. Specifically, we analyze a dataset comprising 9275 clinical notes pertaining to 1730 patients, manually annotated by clinicians to identify the most relevant risk factors, comor-

bidities, and clinical characteristics associated with STEMI.

To the best of our knowledge, this represents the first attempt to extract such granular clinical details, as the site of STEMI or the presence and location of stenosis, from Italian EHRs. Since all these features are categorical, we develop and compare different approaches for their extraction, ranging from rule-based to recurrent neural networks and transformer-based models, showing that different types of features necessitate distinct models based on their complexity levels. Current results are promising and we believe that this pipeline will enhance patients' experiences, both directly as they access their EHRs and indirectly during consultations with physicians.

## 2. Related Works

The sole study that has undertaken a similar approach with Italian documents is (Viani et al., 2019) but notable distinctions exist between their work and ours. They analyzed 75 cardiology reports, focusing on event extraction and specific attributes associated with these events. Consequently, their task is modelled as entity extraction, followed by event classification into four categories (problem, test, treatment, and occurrence) and with respect to some attributes (DocTimeRel, Polarity, Modality, and Experiencer). While their work is significant, we note that our task delves into extracting more detailed structured information, thus presenting a higher level of complexity. Additionally, the substantially larger dataset in our possession provides greater confidence in the generalizability of our results. In (Viani et al., 2018) a similar task was attempted with an ontology-based approach, requiring an elevated effort in the development of the ontology. Similarly, out of the cardiology domain, (Agnello et al., 2021) and (D'Auria et al., 2023) used ontologies and knowledge graphs to extract and link entities from Italian clinical notes.

Out of the Italian domain, as far as our knowledge extends, there exist no specific applications of NLP dedicated to extracting structured information from documents of STEMI patients. Expanding our scope to the wider cardiology domain, we find a few studies on data extraction from EHRs. Some rely solely on rule-based systems (e.g., Patterson et al., 2017, Berman et al., 2021), while others leverage BERT-based models (e.g., Silva et al., 2020, Richter-Pechanski et al., 2021, Singh et al., 2022) or the MedCat tool (Shek et al., 2021). However, these studies only focus on comorbidities or numerical measurements. Alternatively, there are works related to utilizing NLP for cardiology patient classification (e.g., Afzal et al., 2017, Ambrosy et al., 2021, Zaman et al., 2021, Berman et al., 2023). Yet,

none of them align with our objective of identifying more granular clinical details, such as the site of STEMI or the presence and location of stenosis.

## 3. Material and Methods

### 3.1. Dataset

The dataset consists of 9275 EHRs obtained from STEMI patients at FTGM, a specialized cardiology hospital situated in Pisa, Italy. It covers a large period of time from May 2006 to April 2023, and encompasses records from 1730 patients. The experimental protocol was approved by the FTGM ethical committee. Informed consent was obtained from all participants according to the declaration of Helsinki. Each EHR corresponds to a note written by a physician after a specialized patient examination. On average, each patient has 4 records (with a first quartile of 2 and a third quartile of 7). The length of the notes ranges from 1040 to 2047 characters, with a median length of 1677 characters. Clinicians manually annotated these notes with a set of structured or semi-structured variables using Excel software, so the dataset serves as both training and test data for our pipeline.

Table 1 provides a summary of the variables extracted, delineating their types and values. It is worth mentioning that Coronary Artery Bypass Graft surgery history (CABG) and Percutaneous Transluminal Coronary Angioplasty history (PTCA) variables, along with the boolean indicator of non-culprit stenosis, exhibit a large number of missing values. This is primarily because it is not always feasible to assess the positive or negative values of these variables from the textual content of the note since in clinical documentation the primary focus is often on acute medical concerns rather than historical procedures like CABG or PTCA.

The location of the non-culprit stenosis is considered a semi-structured variable because it is reported as free text in the annotations, yielding 272 unique values. We manually consolidated these values into 12 distinct locations, acknowledging that each value might encompass multiple locations. These locations were further condensed into the same five groups used for the culprit vessel.

### 3.2. Methods

Our objective is to develop a pipeline able to extract these variables from clinical notes as they are frequently solely reported within free-form text within clinical practice. Due to the categorical nature of the clinical variables of interest and the semi-structured variable can be mapped to categories, we approach the problem of their extraction from the text of the clinical notes as a text classification problem. The workflow is reported in Figure 1. Consequently, our

Variable	Description	Type	Struct. Values (distribution)	% NA
SK	Smoker	S	Yes (41.8%) / No (58.2%)	1.6%
DB	Diabetes	S	Yes (19.5%) / No (80.5%)	1.8%
HC	Hypercholesterolemia	S	Yes (46.0%) / No (54.0%)	1.9%
HT	Hypertension	S	Yes (57.8%) / No (42.2%)	1.6%
CAD	Coronary Artery Disease Family history	S	Yes (33.6%) / No (66.4%)	1.6%
MI	Myocardial Infarction history	S	Yes (9.2%) / No (90.8%)	1.6%
CABG	Coronary Artery Bypass Graft surgery history	S	Yes (6.1%) / No (93.9%)	82.9%
PTCA	Percutaneous Transluminal Coronary Angioplasty history	S	Yes (22.3%) / No (77.7%)	79.3%
ECG	STEMI location from Diagnostic ECG	S	Front (42.0%) / DX (39.1%) / Lat (17.1%) / Post (0.8%) / Negative (1.1%)	1.8%
CV	Culprit vessel	S	DX (36.0%) / IVA-DA (42.4%) / CX (11.9%) / TC (1.0%) / Other (8.7%)	2.1%
SNC	Presence of a non-culprit stenosis	S	Yes (55.0%) / No (45.0%)	46.4%
LS	Non-culprit stenosis location*	SS	DX (26.7%) / IVA (39.9%) / CX (30.1%) / TC (4.7%) / Other (19.4%)	8.8%

Table 1: Summary information for the variables to be extracted from the notes. S = Structured, SS = Semi-structured.\*LS numbers are only applicable to records with a positive value for SNC; their percentages may exceed 100% due to possible multiple locations. For coronary arteries the corresponding English terms are: DX = RCA, IVA = LDA, CX = LCx, TC = LMCA

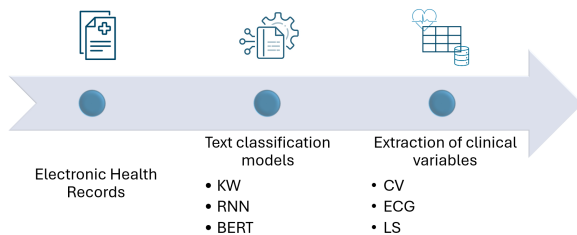


Figure 1: Schema of the proposed pipeline

pipeline consists of a series of text classification models, some of which also require an additional pre-processing of the text: (i) Keyword-based (KW), (ii) Recurrent Neural Network (RNN), (iii) Bidirectional Encoder Representations from Transformers-based (BERT).

The KW model examines the presence of pre-defined keywords and their synonyms within the text associated with each variable. To facilitate this analysis, the text undergoes lowercase conversion and lemmatization. We employ the *Spacy it\_core\_news\_sm* model for lemmatization <sup>1</sup>.

The RNN model architecture is composed of an embedding layer, followed by an Long Short-Term Memory layer and a fully connected layer, each integrated with dropout. Training the network involves utilizing the Adam optimizer with early stopping and cross-entropy as loss function. The hyper-

<sup>1</sup>[https://spacy.io/models/it#it\\_core\\_news\\_sm](https://spacy.io/models/it#it_core_news_sm)

parameters, including layer sizes, dropout rates, and learning rates, are selected via grid search. Preprocessing for this model involves lowercasing, lemmatization, and the removal of words with very low frequencies ( $< 0.5\%$ ), in addition to eliminating Italian stopwords defined in the *nltk* Python library (Bird et al., 2009) and punctuation.

The BERT-based model is a fine-tuned adaptation of the latest Italian version of BERT, Umberto (Tamburini et al., 2020). We further pre-trained the Umberto model on a publicly available corpus of clinical documents (Bernardo Magnini et al, 2020). Since this corpus includes documents in five languages, we automatically translated into Italian, using Google Translate, all documents that were in other languages. This additional pre-training allows the model to acquire knowledge of Italian medical terms. In our classification tasks, fine-tuning occurs solely in the last two layers of the transformer model, while the previous ones are kept frozen during training. The final layer consist of a fully connected layer with sigmoid activation function for binary variables and softmax for multiclass variables. For the location of the non-culprit stenosis, the final layer has multiple binary outputs, one for each potential location. The model is trained with the AdamW optimizer with early stopping and using cross-entropy as loss function. No preprocessing steps are needed since the BERT base model is pre-trained on data without any specific preprocessing.

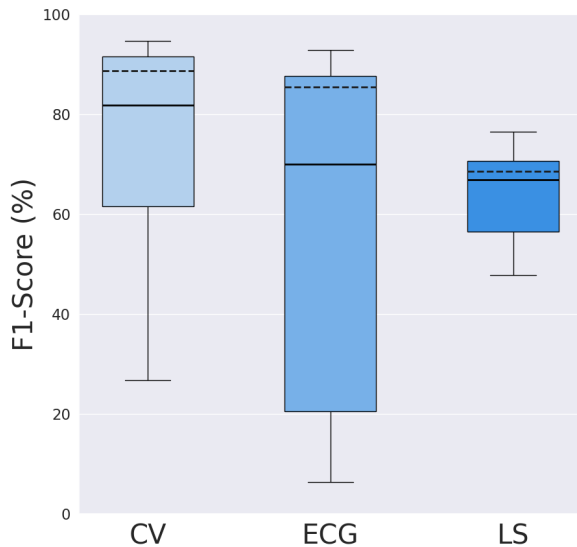


Figure 2: F1-score for the non-binary clinical variables over their possible values. Dashed lines correspond to weighted average.

The models are evaluated with 10-fold stratified cross-validation. Evaluation metrics include precision, recall, f1-score, and area under the receiver operating characteristic curve (AUC).

#### 4. Results

In Table 2, we report the results for each binary variable, comparing the metrics among the three models. In Figure 2, we outline the results for non-binary variables. We report boxplots of the f1-score distribution over the different values that can be assumed by these variables, and we highlight the weighted average. This is relevant as these variables exhibit highly unbalanced value distributions, resulting in outstanding performance for common values but potentially poor outcomes for rare ones. The least favourable results are associated with *Posterior* and *Negative* values of ECG and the *TC* value of both CV and LS. For these variables the best model is BERT.

#### 5. Discussion and Conclusions

The study presents a novel method for extracting structured information from Italian unstructured EHRs, with a focus on STEMI patients. By developing a pipeline and evaluating various classification models, our preliminary results show encouraging outcomes for many variables, with notable achievements such as an f1-score of 89.6% for HT. However, they are less satisfactory for the less frequent values of some variables, like CV, ECG and LS. It is not possible to have a precise comparison with previous works, due to the different variables and

type of data, but for the comorbidities variables our results are aligned with similar works reported in Section 2. Furthermore, our study contributes to the field by evaluating various NLP approaches for Italian data, representing the first attempt to develop such a pipeline in this language. Comparing the different approaches, BERT outperforms in variables requiring deeper contextual understanding and semantics as the case of SNC, whereas for simpler variables, RNN suffices. Specifically, in the case of CABG where the best model results KW, it is expressed very clearly with a few keywords, rendering the use of more complex methods unnecessary. To address the less frequent values, in future work, we propose testing generative models to generate additional training examples for these values. This might be helpful also to cover certain types of expressions that are less frequent and less clear to interpret, such as *"progresso intervento rivascolarizzazione miocardica mediante triplice BPAC"* (*"Previous myocardial revascularization surgery by triple CABG"*) which requires knowing that a triple CABG implies a stenosis on TC (LMCA), CX (LCx) and IVA (LAD). Another potential expansion to be explored is a joint model to extract multiple variables simultaneously, enhancing the efficiency of the pipeline. Additionally, the integration of explainability methods could provide insights into how the model makes decisions, improving its interpretability and trustworthiness. We also defer the comparison with open-source large language models to future work. A limitation of our study is its restriction to data sourced from a single center and from patients affected by a single specific disease (STEMI). To address this problem, validation on a different dataset would be beneficial to ensure the generalizability of our findings across diverse patient populations and healthcare settings. Despite the identified limitations, our pipeline holds significant utility for patients. By organizing the data embedded within their EHRs, we provide clinicians and patients with a more transparent comprehension of their health status and treatment possibilities. This not only encourages patient involvement in their healthcare decisions but also nurtures deeper interactions between patients and physicians during clinical encounters. Therefore, our research contributes to the continuous advancement of patient care and healthcare delivery through the innovative utilization of NLP technologies.

#### 6. Acknowledgements

The present research is part of the activities of "Dipartimento di Eccellenza 2023-2027". Work partially supported by Proximity Care Project and #NEXTGENERATIONEU (NGEU) and funded by

Metric	Model	SK	DB	HC	HT	CAD	MI	CABG	PTCA	SNC
P	KW	70.0 %	52.6 %	82.4 %	83.7 %	60.4 %	23.2 %	90.1 %	37.8 %	63.1 %
P	RNN	<b>90.1 %</b> (4.4 %)	71.4 % (2.1%)	<b>88.4 %</b> (2.7%)	87.8 % (3.4 %)	<b>91.6 %</b> (3.1 %)	67.5 % (1.5 %)	<b>93.4 %</b> (2.2 %)	<b>86.5 %</b> (4.3 %)	77.8 % (2.1 %)
P	BERT	86.0 % (1.4 %)	<b>80.4 %</b> (3.2%)	81.9 % (1.6 %)	<b>89.6 %</b> (1.0 %)	88.1 % (0.9 %)	<b>70.1 %</b> (3.1 %)	89.0 % (4.1 %)	84.9 % (5.2 %)	<b>84.8 %</b> (1.9 %)
R	KW	79.9 %	<b>79.9 %</b>	23.7 %	63.0 %	35.6 %	53.7 %	<b>84.5 %</b>	21.3 %	37.4 %
R	RNN	83.0 % (4.5 %)	78.9 % (2.6 %)	77.1 % (2.9 %)	<b>91.7 %</b> (6.6 %)	85.2 % (3.3 %)	<b>76.1 %</b> (2.6 %)	77.6 % (1.1 %)	<b>73.0 %</b> (5.2 %)	72.0 % (3.8 %)
R	BERT	<b>85.9 %</b> (1.4 %)	71.1 % (2.9 %)	<b>81.8 %</b> (1.5 %)	89.5 % (1.0 %)	<b>88.0 %</b> (0.8 %)	74.8 % (3.2 %)	82.0 % (2.0 %)	68.7 % (8.8 %)	<b>83.8 %</b> (5.2 %)
F1	KW	74.6 %	63.4 %	36.8 %	71.9 %	44.8 %	32.4 %	<b>87.2 %</b>	27.3 %	46.7 %
F1	RNN	<b>86.9 %</b> (4.5 %)	<b>73.4 %</b> (2.5 %)	<b>83.5 %</b> (3.0%)	<b>89.6 %</b> (9.6 %)	<b>88.3 %</b> (3.6 %)	<b>71.5 %</b> (3.0 %)	82.9 % (2.8 %)	<b>78.8 %</b> (4.3 %)	74.6 % (3.1 %)
F1	BERT	85.9 % (1.4 %)	71.2 % (2.8 %)	81.8 % (1.6 %)	89.5 % (1.0 %)	88.0 % (0.9 %)	71.3 % (2.8 %)	85.4 % (2.9 %)	75.6 % (6.0 %)	<b>84.0 %</b> (4.9 %)
AUC	RNN	85.0 % (4.5 %)	<b>78.9 %</b> (2.6 %)	<b>83.7 %</b> (2.9 %)	<b>87.2 %</b> (2.6 %)	<b>88.7 %</b> (3.4 %)	69.4 % (2.6 %)	<b>79.5 %</b> (5.8 %)	<b>75.2 %</b> (5.2 %)	75.3 % (3.9 %)
AUC	BERT	<b>92.7 %</b> (1.1 %)	77.0 % (1.1 %)	82.8 % (1.3 %)	86.8 % (1.5 %)	86.5 % (1.2 %)	<b>72.5 %</b> (0.7 %)	79.4 % (5.1 %)	73.3 % (2.8 %)	<b>82.0 %</b> (4.5 %)

Table 2: Results for the binary variables on 10-fold stratified cross-validation, reported as mean (std dev) for each model. P = Precision, R = Recall, F1 = F1-Score, AUC = Area Under the receiver operating characteristic Curve. Standard deviation is not reported for KW since there is no training set. AUC is not reported for KW since it does not output a probability. Best results for each metric and variable are highlighted in bold.

the Ministry of University and Research (MUR), National Recovery and Resilience Plan (NRRP), project BRIEF (IR0000036) - Biorobotics Research and Innovation Engineering Facilities (DN. 103 17.06.2022).

## 7. Bibliographical References

- Naveed Afzal, Sunghwan Sohn, Sara Abram, Christopher G Scott, Rajeev Chaudhry, Hongfang Liu, Iftikhar J Kullo, and Adelaide M Arruda-Olson. 2017. Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. *Journal of vascular surgery*, 65(6):1753–1761.
- Patrizia Agnello, Silvia Maria Ansaldi, Fabio Azalini, Giovanni Gangemi, Davide Piantella, Emanuele Rabosio, Letizia Tanca, et al. 2021. Extraction of medical concepts from Italian natural language descriptions. In *29th Italian Symposium on Advanced Database Systems, SEBD*, volume 2994, pages 275–282.
- Andrew P Ambrosy, Rishi V Parikh, Sue Hee Sung, Anand Narayanan, Rajeev Masson, Phuong-Quang Lam, Kevin Kheder, Alan Iwahashi, Alexander B Hardwick, Jesse K Fitzpatrick, et al. 2021. A natural language processing–based approach for identifying hospitalizations for worsening heart failure within an integrated health care delivery system. *JAMA Network Open*, 4(11):e2135152–e2135152.
- Adam N Berman, David W Biery, Curtis Ginder, Olivia L Hulme, Daniel Marcusa, Orly Leiva, Wanda Y Wu, Nicholas Cardin, Jon Hainer, Deepak L Bhatt, et al. 2021. Natural language processing for the assessment of cardiovascular disease comorbidities: The cardio-Canary comorbidity project. *Clinical Cardiology*, 44(9):1296–1304.
- Adam N Berman, Curtis Ginder, Zachary A Sporn, Varsha Tanguturi, Michael K Hidrue, Linnea B Shirkey, Yunong Zhao, Ron Blankstein, Alexander Turchin, and Jason H Wasfy. 2023. Natural Language Processing for the Ascertainment and Phenotyping of Left Ventricular Hypertrophy and Hypertrophic Cardiomyopathy on Echocardiogram Reports. *The American Journal of Cardiology*, 206:247–253.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."
- Charlotte Blease, Leonor Fernandez, Sigall K Bell,



- Tom Delbanco, and Catherine DesRoches. 2020. Empowering patients and reducing inequities: is there potential in sharing clinical notes? BMJ Quality & Safety, 29(10):1–2.
- Daniela D’Auria, Vincenzo Moscato, Marco Postiglione, Giuseppe Romito, and Giancarlo Sperlí. 2023. Improving graph embeddings via entity linking: a case study on Italian clinical notes. Intelligent Systems with Applications, 17:200161.
- Mark W Friedberg, Peggy G Chen, Kristin R Van Busum, Frances Aunon, Chau Pham, John Caloyer, Soeren Mattke, Emma Pitchforth, Denise D Quigley, Robert H Brook, et al. 2014. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. Rand health quarterly, 3(4).
- Paul Hager, Friederike Jungmann, Kunal Bhat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Robbie Holland, Rickmer Braren, Marcus Makowski, Georgios Kaisis, and Daniel Rueckert. 2024. [Evaluating and Mitigating Limitations of Large Language Models in Clinical Decision Making](#). medRxiv.
- Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2545–2568.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Jared W Klein, Sara L Jackson, Sigall K Bell, Melissa K Anselmo, Jan Walker, Tom Delbanco, and Joann G Elmore. 2016. Your patient is now reading your note: opportunities, problems, and prospects. The American journal of medicine, 129(10):1018–1021.
- Viet Dac Lai, Nghia Trung Ngo, Amir Pourn Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. arXiv preprint arXiv:2304.05613.
- Bernardo Magnini, Begona Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanolini. 2021. The E3C project: European clinical case corpus. Language, 1(L2):L3.
- Aurélie Névéol, Hercules Dalianis, Sumithra Velupillai, Guergana Savova, and Pierre Zweigenbaum. 2018. Clinical natural language processing in languages other than English: opportunities and challenges. Journal of biomedical semantics, 9(1):1–13.
- Olga V Patterson, Matthew S Freiberg, Melissa Skanderson, Samah J. Fodeh, Cynthia A Brandt, and Scott L DuVall. 2017. Unlocking echocardiogram measurements for heart disease research through natural language processing. BMC cardiovascular disorders, 17:1–11.
- Phillip Richter-Pechanski, Nicolas A Geis, Christina Kiriakou, Dominic M Schwab, and Christoph Dieterich. 2021. Automatic extraction of 12 cardiovascular concepts from German discharge letters using pre-trained language models. Digital Health, 7:20552076211057662.
- Joseph Root, Natalia V Oster, Sara L. Jackson, Roanne Mejilla, Jan Walker, and Joann G Elmore. 2016. Characteristics of patients who report confusion after reading their primary care clinic notes online. Health communication, 31(6):778–781.
- S Trent Rosenbloom, Joshua C Denny, Hua Xu, Nancy Lorenzi, William W Stead, and Kevin B Johnson. 2011. Data from clinical notes: a perspective on the tension between structure and flexible documentation. Journal of the American Medical Informatics Association, 18(2):181–186.
- Anoop D Shah, Emily Bailey, Tim Williams, Spiros Denaxas, Richard Dobson, and Harry Hemingway. 2019. [Natural language processing for disease phenotyping in UK primary care records for research: a pilot study in myocardial infarction and death](#). Journal of biomedical semantics, 10(Suppl 1):20. PMID: 31711543.
- Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, Venet Osmani, et al. 2019. Natural language processing of clinical notes on chronic diseases: systematic review. JMIR medical informatics, 7(2):e12239.
- Anthony Shek, Zhilin Jiang, James Teo, Joshua Au Yeung, Ajay Bhalla, Mark P Richardson, and Yee Mah. 2021. Machine learning-enabled multitrust audit of stroke comorbidities using natural language processing. European Journal of Neurology, 28(12):4090–4097.
- João Figueira Silva, João Rafael Almeida, and Sérgio Matos. 2020. Extraction of family history information from clinical notes: deep learning and heuristics approach. JMIR Medical Informatics, 8(12):e22898.

Pulkit Singh, Julian Haimovich, Christopher Reeder, Shaan Khurshid, Emily S Lau, Jonathan W Cunningham, Anthony Philippakis, Christopher D Anderson, Jennifer E Ho, Steven A Lubitz, et al. 2022. One clinician is all you need—cardiac magnetic resonance imaging measurement extraction: Deep learning algorithm development. JMIR Medical Informatics, 10(9):e38178.

Fabio Tamburini et al. 2020. How “BERTology” changed the state-of-the-art also for Italian NLP. In CEUR WORKSHOP PROCEEDINGS, volume 2769, pages 1–7. CEUR-WS.

Maryam Tayefi, Phuong Ngo, Taridzo Chomutare, Hercules Dalianis, Elisa Salvi, Andrius Burdionis, and Fred Godtliebsen. 2021. Challenges and opportunities beyond structured data in analysis of electronic health records. Wiley Interdisciplinary Reviews: Computational Statistics, 13(6):e1549.

Natalia Viani, Cristiana Larizza, Valentina Tibollo, Carlo Napolitano, Silvia G Priori, Riccardo Bellazzi, and Lucia Sacchi. 2018. Information extraction from Italian medical reports: An ontology-driven approach. International journal of medical informatics, 111:140–148.

Natalia Viani, Timothy A Miller, Carlo Napolitano, Silvia G Priori, Guergana K Savova, Riccardo Bellazzi, and Lucia Sacchi. 2019. Supervised methods to extract clinical events from cardiology reports in Italian. Journal of biomedical informatics, 95:103219.

Sameer Zaman, Camille Petri, Kavitha Vimalasvaran, James Howard, Anil Bharath, Darrel Francis, Nicholas S Peters, Graham D Cole, and Nick Linton. 2021. Automatic diagnosis labeling of cardiovascular MRI by using semisupervised natural language processing of text reports. Radiology: Artificial Intelligence, 4(1):e210085.

## **8. Language Resource References**

Bernardo Magnini et al. 2020. E3C - European Clinical Case Corpus.

# Towards AI-supported Health Communication in Plain Language: Evaluating Intralingual Machine Translation of Medical Texts

Silvana Deilen<sup>1</sup>, Ekaterina Lapshinova-Koltunski<sup>1</sup>, Sergio Hernández Garrido<sup>1</sup>,  
Christiane Maaß<sup>1</sup>, Julian Hörner<sup>2</sup>, Vanessa Theel<sup>3</sup>, Sophie Ziemer<sup>3</sup>

<sup>1</sup>University of Hildesheim, <sup>2</sup>Text und Bild Verlag, <sup>3</sup> SUMM AI  
deilen,lapshinovakoltun,hernandezs,maass@uni-hildesheim.de,  
j.hoerner@wubv.de, sophie,vanessa@summ-ai.com

## Abstract

In this paper, we describe results of a study on evaluation of intralingual machine translation. The study focuses on machine translations of medical texts into Plain German. The automatically simplified texts were compared with manually simplified texts (i.e., simplified by human experts) as well as with the underlying, unsimplified source texts. We analyse the quality of the translations based on different criteria, such as correctness, readability, and syntactic complexity. The study revealed that the machine translations were easier to read than the source texts, but contained a higher number of complex syntactic relations than the human translations. Furthermore, we identified various types of mistakes. These included not only grammatical mistakes but also content-related mistakes that resulted, for example, from mistranslations of grammatical structures, ambiguous words or numbers, omissions of relevant prefixes or negation, and incorrect explanations of technical terms.

**Keywords:** plain language, medical discourse, accessible health communication, health literacy, machine translation

## 1. Introduction

In interlingual translation, CAT (computer-aided translation) tools and machine translation systems such as DeepL or Google Translate have significantly changed the translation industry and have become an indispensable component in the translation process, as can be seen in the case of the European Commission: While until a few years ago, all legal texts and official documents were translated by human translators alone, today, the European Commission makes no secret of the fact that their translators now tend to revise and post-edit the texts rather than translate them themselves.

However, it seems that this "transition to a new era" (Canfora and Ottmann, 2020) has not yet reached the field of intralingual translation. Following Jakobson, intralingual translation is defined as "an interpretation of verbal signs by means of other signs of the same language" (Jakobson, 1959, p. 233). In this context, it refers to translating a text from standard language into a complexity-reduced language variety of the same language as described in Maaß (2020, p. 171ff) and in Maaß (2024, p. 265ff). Plain Language translation is also related to text simplification, which is an automatic procedure of changing complex structures into simple ones. However, from the perspective of translation studies and translation practice, this is a type of translation that involves more than reducing surface complexity. Unlike in interlingual translation (i.e. translating a text from one lan-

guage to another), in intralingual translation the use of CAT tools and machine translation systems is still not established (Maaß et al. 2014, Deilen et al. 2023).

However, especially in health communication, there is a high need for technological assistance, which is especially due to the population's alarmingly low health literacy, as Schaeffer et al. (2017) point out. Their findings lead to the National Action Plan of the German Federal Government to promote health literacy (*Nationaler Aktionsplan Gesundheitskompetenz*, Schaeffer et al., 2018a) that lists Plain Language among the instruments to secure better access to information as the basis for better health literacy (see Section 2.2 below).

In our study, we analyse machine translations of medical texts into Plain German. The texts were taken from the website of the German health magazine *Apotheken Umschau*, which publishes healthcare articles and health information both in standard German and in Plain German. We evaluate the machine-translated output comparing it with human translations from the magazine's website, as well as with the underlying sources. We present the results of the qualitative and quantitative analysis.

## 2. Related Work

### 2.1. Plain German

Both Easy Language and Plain Language are complexity-reduced language varieties which aim to improve readability and comprehensibility of

texts (Bredel and Maaß, 2016; Maaß, 2020). They are used in different communication scenarios, e.g. in legal communication (Maaß and Rink, 2021) or health communication (see the contributions in Ahrens et al., 2022), and have different target groups (Maaß and Schwengber, 2022). While Easy Language is characterized by a maximally reduced complexity on all language levels and is mainly intended for people with communication impairments and disabilities, the grammatical and lexical features of Plain Language are only slightly less complex than in standard language and are mainly a means to open expert contexts for lay people (Maaß, 2020). Therefore, the main target audience of Plain Language is lay people with average or slightly below average language or reading skills (Maaß, 2020). In Germany, Easy Language has become a subject of scientific research since 2014 with rapidly growing output of publications in the following years. The studies point in two basic directions: studies on text qualities and possible barriers in various forms of communication on the one side (see, for example, Rink 2019) and studies on comprehensibility and recall by different target groups on the other (see, for example, Gutermuth 2020, Deilen 2021).

Unlike Easy Language, Plain Language is a dynamic variety. Plain Language does not have a fixed set of rules, but the linguistic complexity of Plain Language texts is adapted to the needs of the intended audience in a specific target situation (Bredel and Maaß 2016, Maaß 2020). Therefore, Plain Language is a flexible concept that varies depending on the presumed reading skills of its target group (for a more detailed distinction between the two varieties see Maaß 2020). In comparison to Easy Language, Plain Language has the advantage of not stigmatising the target audience (Maaß, 2020), which is one of the reasons why it is also more acceptable than Easy Language. However, due to the higher degree of linguistic complexity, Plain Language texts are far less comprehensible than Easy Language texts and therefore not necessarily accessible for people with very low literacy skills (Maaß, 2020). Maaß (2020) therefore models the variety Easy Language Plus, which is situated between Easy Language and Plain Language and strikes a balance between comprehensibility and acceptability.

In Germany, Plain Language is used in different fields and different settings, such as by *Deutschlandfunk*, a public-broadcasting radio station that publishes weekly news in Plain Language for a broad audience with reading difficulties or reduced language skills. However, one of the most prominent application areas of German Plain Language is health communication (Ahrens et al., 2022).

## 2.2. Accessibility in Medical Domain in Germany

In 2016, findings from the Health Literacy Survey (HLS-GER) revealed that over half of the German population (54,3%) experiences significant difficulties in locating, comprehending, evaluating and effectively using health-related information (Schaeffer et al., 2017). These results, which were "significantly worse than expected" (Schaeffer et al., 2020, p. 2), led to an increased awareness of the need for accessible health information and resulted in the development of the National Action Plan Health Literacy, which was published in 2018 (Schaeffer et al., 2018b). According to the National Action Plan, one strategy to promote health literacy in Germany is the use of Plain Language, which "aims to adapt complex texts to the literacy skills of large population groups" (Schaeffer et al., 2018b, p. 43); the National Action Plan cites the model put forward in Bredel and Maaß (2016) for reference. Considering new data from the second Health Literacy Survey (HLS-GER 2) in 2021, Plain Language in German health communication becomes even more relevant, for even more persons (58,8 %) experience difficulties navigating the health system (Schaeffer et al., 2021). One of the most prominent examples of implementing this strategy is the *Apotheken Umschau*<sup>1</sup>. The *Apotheken Umschau*, which is Germany's leading health publisher and the largest consumer medium in the German-speaking area with a traffic of 6.94 m. visits and 64.42 m. page impressions per month<sup>2</sup>, has so far published more than 220 texts in Plain Language on their website in a co-operation with the Research Centre for Easy Language (University of Hildesheim)<sup>3</sup>. By publishing information in both standard German and Plain German, they aim to "make reliable and helpful information on diseases, medications and preventive health care accessible to everyone with as few barriers as possible" (Hörner, 2022, p. 77). The project is based on the linguistic model for Plain Language by Bredel and Maaß (2016) and Easy Language Plus by Maaß (2020).

## 2.3. NLP for Plain Languages

Although the potentials of using computer-aided translation (CAT) tools for Plain Language translation were discussed almost a decade ago (Maaß et al., 2014), the role of automation and CAT tools in this area is still a major research desideratum. These potentials were re-explored and extended by Hansen-Schirra et al. (2020). In gen-

<sup>1</sup><https://www.apotheken-umschau.de>

<sup>2</sup><https://ausweisung-digital.ivw.de>, retrieved 10.10.2023

<sup>3</sup><https://www.uni-hildesheim.de/leichtesprache>

eral, intralingual translation poses a number of challenges for CAT tools: terminology management and sentence alignment (see e.g. [Kopp et al. 2023](#)) differ from those common in interlingual translation and, therefore, pose additional workload for translators instead of decreasing it. The theoretical set-up for a CAT tool for intralingual translation was suggested by [Welch and Sauberer \(2019\)](#). However, to our knowledge, such tools, as well as their analysis, are still missing.

While there are plenty of studies on automatic text simplification methods that aim to automatically convert a text into another text that is easier to understand but ideally conveys the same message as the source text which contributes to textual accessibility ([Sheang and Saggion, 2021](#); [Maddela et al., 2021](#); [Martin et al., 2020](#); [Saggion, 2017](#)), most of them do not consider the needs of the target audience. [Scarton and Specia \(2018\)](#) showed that using target audience oriented data helps to build better models for automatic text simplification using the Newsela corpus<sup>4</sup>. However, this corpus contains news texts only, whereas we are looking into the medical discourse, where texts in Plain Language enable accessibility to health literacy. Biomedical lay summarization is also related to automatic translation into Plain Language. [Gold-sack et al. \(2023\)](#) present results of a shared task on lay summarization of biomedical research articles (BioLaySumm 2023). In this case, medical information in expert language (expert-to-expert communication) is summarized for non-experts (expert-lay-communication). However, it is important to state that Plain Language translation, even if translators select and add information as described in [Bredel and Maaß \(2016, p. 202 ff.\)](#), is not the same textual practice as text summarization.

Specific problems of automatic systems of intralingual translation, e.g. copying source segments into the output, were addressed by [Säuberli et al. \(2020\)](#) and [Spring et al. \(2023\)](#) who showed that pretrained and fine-tuned NMT models have promising results in automatic text simplification. However, as stated by [Anschütz et al. \(2023\)](#), even though there are improvements in the systems of automated intralingual translation, the outputs might not be used by the target groups directly. Nevertheless, they may serve as a draft for professional intralingual translators to reduce their workload.

[Deilen et al. \(2023\)](#) drew similar conclusions for the outputs produced with ChatGPT. The authors investigated the feasibility of using this tool for intralingual translation. They analysed the quality of the generated texts according to such criteria as correctness, readability, and syntactic complex-

ity. Their results indicated that the generated texts were easier than the standard texts, but the content was not always rendered correctly. Besides that, the automated intralingual output did not fully meet the standards which human translators follow.

In the present study, we follow a similar approach. However, while the authors analysed intralingual translation into German Easy Language, a simplified, controlled language variety adapted to the needs of people with reading impairments, we focus on translation into Plain German. Besides that, we focus on medical texts, whereas the authors translated citizen-oriented administrative texts. Moreover, we investigate the feasibility of a tool which was specifically trained for intralingual translation into Easy and Plain Language instead of using a chatbot designed for various tasks.

### 3. Research Design

#### 3.1. Data Collection

We selected thirty texts from the website of the German health magazine *Apotheken Umschau*. The texts cover a broad range of topics such as insect bite, vaccination, cystitis, lumbago, food poisoning, heel spur and others. For all texts in the sample, a translation in Plain Language was already available, which was done by human translators. Both the source texts and the human translations were reviewed by medical or pharmaceutical professionals from the editorial team of *Apotheken Umschau* and comply with the guidelines of evidence-based medicine. Content accuracy is therefore guaranteed for the sample. This sample of thirty texts was translated using the machine translation system SUMM AI<sup>5</sup>.

Then, we analysed machine-translated texts comparing them with human translations, as well as with the source texts following [Deilen et al. \(2023\)](#). For this, we used three different criteria, namely the correctness of the content (see 3.2.1), the readability of the texts (see 3.2.2), and their syntactic complexity (see 3.2.3). The first criterion was applied to the machine translations only, the second and the third criteria were applied on all the three subcorpora (source texts, human translations, and machine translations)<sup>6</sup>.

<sup>5</sup>SUMM AI (<https://summ-ai.com/en/>) is a tool for translating texts into Easy German and Plain German. The company SUMM AI offers different licenses for freelancers, authorities and companies.

<sup>6</sup>The whole dataset we analysed is published on GitHub, i.e., the selected texts (sources, human and machine translations), including the raw data, the parsed data (conllu) and the Textlab analyses per text, and can be accessed under <https://github.com/katjakaterina/MT4plainDE>.

<sup>4</sup><https://newsela.com/data>

## 3.2. Data Analysis

### 3.2.1. Correctness

The content of the machine-generated texts was first analysed for correctness. This content evaluation was done manually, whereby each text was assessed independently by two researchers, who checked whether the medical information in the target text is still valid despite reduction of complexity and shortening of information. In cases where an accurate assessment required specialized knowledge, a healthcare professional from the *Apotheken Umschau* team was consulted. No quantitative error analysis was performed. Consequently, a translation was already considered incorrect if it contained one content-related error. This is because the study seeks insights into who artificial intelligence (AI) powered translation tools are suitable for: For translators, content providers, or end users? In order for machine translation into Easy or Plain Language to be safely usable by end users, the target texts must not contain errors. The presence of errors in the target texts therefore indicates usability for users other than the end users.

### 3.2.2. Readability

We also compared the comprehensibility of the human and machine translations, as well as of the source texts. For this, we use the Hohenheim Comprehensibility Index (HIX). The HIX is a meta index that calculates the readability of a text taking into account the four major readability formulas common in Easy Language Research (Bredel and Maaß, 2016, p. 61ff). They include the Amstad index, the simple measure of gobbledygook (G-SMOG) index, the Vienna non-fictional text formula (W-STX) and the readability index (LIX), with a HIX of 0 indicating extremely low comprehensibility and a HIX of 20 extremely high comprehensibility (for further details see: <https://klartext.uni-hohenheim.de/hix>). The benchmark for a text to be classified as a text in Easy German, which is the least complex variety of German, is set at 18 points (Rink 2019, p. 77). As Plain German is more complex than Easy German, we suggest setting the benchmark for Plain German at 16 points.

### 3.2.3. Syntactic Complexity

We operationalised syntactic complexity as a distribution of specific syntactic relations, i.e. specific clauses. We automatically identified syntactic relations using dependency parsing that we obtained with the Stanford NLP Python Library Stanza (v1.2.1)<sup>7</sup> with all the models pre-trained on the Universal Dependencies v2.5 datasets. Our

<sup>7</sup><https://stanfordnlp.github.io/stanza/index.html>

list of selected structural categories include adnominal clauses or clausal modifiers of noun (acl), adverbial clause modifiers (advcl), clausal complement (ccomp), clausal subjects (csubj), open clausal elements (xcomp) and parataxis relation (parataxis). These selected categories are all listed under the clause dependents<sup>8</sup> in the Universal Dependency. More details on dependency relations and their definitions across languages can be found in (De Marneffe et al., 2021). We collected and compared the distribution frequencies of these categories in the three subcorpora under analysis (source texts, human translations, and machine translations). We interpreted the results based on the assumption that the higher the number of these dependency relations in the corpus, the more complex the texts contained in these subcorpora are.

### 3.2.4. Automatic Evaluation Measures

We also used other indices that are commonly used in the field of automatic text simplification. Specifically, we applied SARI (Xu et al., 2016), which is a quantitative measure to evaluate automatic text simplification systems. SARI is suitable for evaluation of automatic text simplification models and could so be also suitable for the task of evaluating intralingual machine translation. In order to be able to compute these metrics, we aligned the source texts, machine translations and human translations on a paragraph level and scored them with respect to their grade of alignment. Out of the 935 analysed paragraphs more than 70%, namely 676 paragraphs, had no alignment between source text and human translations. This means that for 547 paragraphs in the source text no matching simplification could be allocated in the human translation and for 123 paragraphs in the human translation no matching source paragraph could be identified.

## 3.3. Results

### 3.3.1. Correctness

The analysis of the correctness of the machine translations showed that only one of the 30 texts was correctly translated. The other 29 texts showed problems with regard to their correctness in different aspects. Overall, the results are disparate and inconsistent. The texts do not follow a uniform structure and are not action-oriented. In practice, they would have to be completely post-edited. In some cases, the source texts (ST) are more stringent and comprehensible than the target texts (TT). We encounter grammatical errors and misspellings, omissions of relevant prefixes or negation, incorrect explanations of technical terms, incomplete listings, contradictory state-

<sup>8</sup><https://universaldependencies.org/u/dep/>

ments etc. It should be emphasized once again that no quantitative evaluation was performed because the mere presence of the errors themselves was considered a risk for the primary users. Furthermore, so far we have not classified or ranked the error types based on severity levels, but we plan to do so in our future work (see 4).

Some examples of the errors we found are given in the following.

Grammatical errors and misspellings are for example:

- homophonic but not homographic words are not correctly selected: "dass" (the connective "that") vs. "das" (the article "the" or the relative pronoun "that"). They are used in German to differentiate a function and they are not interchangeable (1):

1. "das" instead of "dass": "Die Zahl 1 bedeutet, **das** der Tumor weniger als 1 Millimeter dick ist." ["The number 1 means **the** the tumor is less than 1 millimeter thick" instead of "**that** the tumor"]

Another example is given in (2):

2. "isst" ("eat") vs. "ist" ("is"): "Wenn man nüchtern **isst**, geht es sehr schnell" ["When you **eat** sober, it happens very quickly" instead of "When you **are** sober"]

- wrong prepositions: "durch" ("through") vs. "von" ("of") (3):

3. "Erkrankung **von** Lebensmitteln" ["illness **of** food" instead of "**through** food"]

- wrong genus of nouns (4):

4. "**Das** Rücken wird dann immer schwerer und schwerer" ["The back then becomes heavier and heavier"], "Es soll **kein** Rückfall geben" ["There should be no relapse"]

Other errors contain verb numerus, genus of nouns, the syntax of clauses or sentences, especially in the passive voice, and other.

Semantic errors or inaccuracies are for example:

- Wrong explanations (5):

5. "am Tage mehrfach **wegdösen**" [ST, "**dozing off** several times during the day"] vs. "man **fällt** am Tag mehrmals **weg** und muss dann wieder aufwachen" [TT, "**falling away** several times during the day and then having to wake up again"]

- Terminology inaccuracies, e.g. (6):

6. "Ein erhöhtes Schlafbedürfnis am Tag, eine sogenannte Hypersomnie, ist eine oft kennzeichnende Folge solcher nächtlichen unbewusst oder bewusst erlebten Unterbrechungen" [ST, "Hypersomnia is characterized by increased need for sleep during the day caused by sleep interruptions at night."]. In the translation, this connection is no longer clear due to the information being abridged: "Manchmal kann man nachts aufwachen und dann nicht mehr einschlafen. Das nennt man dann Hypersomnie." [TT, "Sometimes you can wake up at night and then not go back to sleep. This is then called hypersomnia"]

- Polysemous words errors: In German, "Satz" means "sentence", but also "leap" (7):

7. "Ansonsten vermutlich der übliche **Satz** morgens aus dem Bett" [ST, "Otherwise probably the usual **leap** out of bed in the morning"] vs. "Sonst ist es wahrscheinlich der übliche **Satz**, den Sie morgens sagen" [TT, "Otherwise it is probably the usual **sentence** you say in the morning."].

- Sentences with conditional meaning have a particularly high error rate, like in the following example (8):

8. "Bei fortgeschrittenen Tumorstadien [...] ist eine umfassendere Behandlung notwendig" [ST, "In advanced tumor stages [...] more comprehensive treatment is necessary"] vs. "Wenn der Hautkrebs schon weiter fortgeschritten ist, gibt es mehr Möglichkeiten zur Behandlung" [TT, "If the skin cancer is already more advanced, there are more options for treatment"]

Correctness is not yet present for the system under study to the extent that texts would be usable without post-editing. The human translation corpus does not have such errors, but has a high degree of correctness.

### 3.3.2. Readability

Comparing the comprehensibility of the human and machine translations, as well as of the source texts, revealed that the machine translations had the highest comprehensibility, with a mean HIX value of 19.15 (SD: 0.49). In comparison, the human translations yielded a mean HIX value of 17.74 (SD: 1.67). Based on the HIX, the source texts were the least comprehensible (mean: 10.46, SD: 2.76). Given the low variance

in the machine translations (see Figure 1), all of the 30 texts could be classified as Plain Language texts.

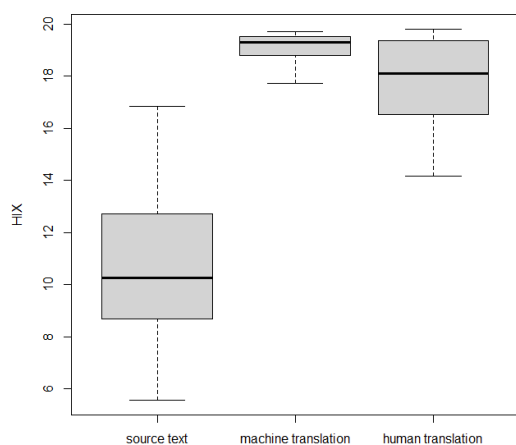


Figure 1: HIX values of the source text, the machine translation, and the human translation.

From the human translations, however, only 83% of the texts reached the predefined Plain German benchmark. As seen from the boxplot, human translations reveal a much greater variation in the HIX values than the machine-translated texts.

It is important to highlight that HIX values only consider overt complexity. Therefore, these values represent a starting point for evaluating comprehensibility, but have to be complemented with further qualitative analysis.

### 3.3.3. Syntactic Complexity

In the next step, we analysed the distribution of the dependency relations in human and machine translation, as well as in the source texts. We summarise the results (frequencies normalised per 10000) in Figure 2.

The distribution numbers reveal that both the source texts and machine translations seem to contain a higher number of complex syntactic relations than the human translation. For the latter, we observed higher number for parataxis relations only. This goes in hand with the results by Deilen et al. (2023) who reported similar tendencies for machine-generated texts - they contained more complex constructions than those generated by humans. However, the authors did not compare machine-generated texts with human translations. In our study, we had human translations at our disposal and observed that they are the most simplified amongst the subcorpora under analysis. Interestingly, sources contain more clausal subjects (csubj) as well as clauses modifying nouns (acl) than machine-translated texts do. It is also interesting to note that human translations do not contain any clausal subjects at all. At the same

time, clausal complements of verbs and adjectives (ccomp) along with clauses modifying verbs and adjectives (advcl) predominate in machine translations. So do subjectless clausal complements (xcomp), whose number is significantly higher in machine translated-texts if compared to the other two subcorpora.

A sentence with a clausal subject that is frequent in source texts is illustrated in example 9a (*Wer... [Who...]*). Its corresponding machine translation in 9b contains a clause modifying a verb (*Wenn man merkt/If you realise*) complement of a verb and a clausal verb complement (starting with *dass.../that*). The only subclause contained in the human translation (in 9c) is parataxis. The other parts are simple sentences.

9a. *Wer bei sich Probleme im Umgang mit Alkohol feststellt, sollte daher unbedingt das Gespräch mit dem Arzt suchen. [...]* (source)

9b. *Wenn man merkt, dass man mit Alkohol Probleme hat, sollte man unbedingt zum Arzt gehen. [...]* (machine translation)

9c. *Sie glauben: Ich bin vielleicht alkoholsüchtig? Dann sprechen Sie mit Ihrem Arzt.* (human translation)

### 3.3.4. Automatic Evaluation Measures

In the last step, we analysed the SARI score of machine translated texts which is a quantitative measure of text simplification. The boxplot visualising the SARI score computed on all aligned segments is displayed in Figure 3. As already mentioned in Section 3.2.4 above, SARI compares machine translated output with the human translations and the sources measuring added, deleted or kept words. Higher SARI values indicate better machine translated outputs.

The system used in the analysis achieves an average SARI-Score of 40.67 (SD: 6.79), which is in line with state-of-the-art text simplification models reported by Sheang and Saggion (2021). We also see from the box plot that our maximum values achieved by the system are around 55. Moreover, the data contains many outliers, i.e. segments with the score of over 55.

## 4. Discussion and Future Work

The present paper evaluates the use of a machine translation system for translating medical texts into Plain German. Our results showed that in terms of readability, the machine translations are much easier than the source texts and even easier than the human translations. Analysing the syntactic complexity, however, revealed that machine translations contain a significantly higher number of



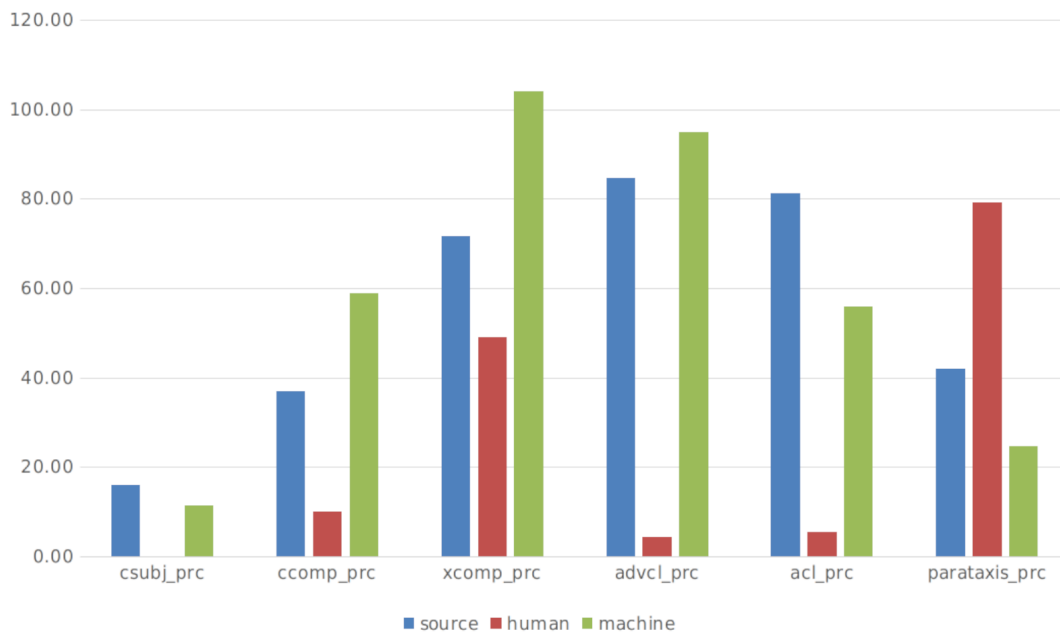


Figure 2: Distribution of syntactically complex dependency relations in the source texts, human and machine translations (normalised frequencies per 10000).

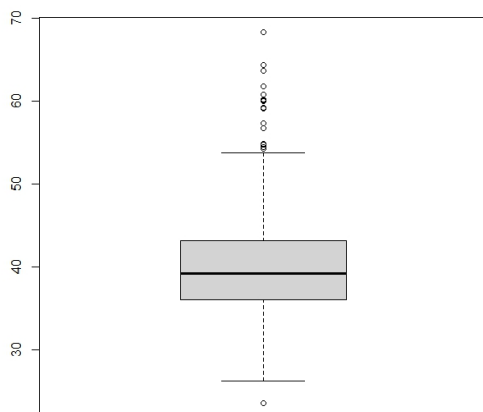


Figure 3: SARI score of the aligned source text, machine translation and human translation.

complex syntactic relations than human translations. Particularly interesting and against our expectations was the result that in most cases, the machine translations are even more complex than the source texts. Furthermore, our analysis revealed that the machine-translated texts contained various types of mistakes.

In our further research, we will proceed to classify the different types of machine translation errors and also look into different cases of partial correctness, where only some pieces of information were incorrect or missing. Furthermore, the present study only focused on the text perspective. However, to draw reliable conclusions about the

functionality of a translation, not only the text but also the user perspective needs to be considered. Therefore, in our future work, we plan to conduct empirical studies, consisting of eye-tracking and reading experiments, to gain insights into the cognitive processing costs of the target groups when reading machine translated texts. In addition, we plan to use questionnaires to investigate whether the end users accept the generated texts.

All in all, we conclude that the analysed tool is a promising text simplification tool, however, in terms of correctness and syntactic complexity, it still does not reach the human parity. The machine translation system showed its limitations in the field of selecting and prioritizing information, including adequate examples and images, and adapting the content to the prior knowledge of the target groups, i.e. adding for example explanations of difficult words and concepts. Human translators are therefore still indispensable. It becomes very clear that machine translated Plain Language texts cannot do without post-editing, but need intensive revision. The translation tools at hand are therefore not yet suitable for end users, but are rather to be used as CAT tools for professional translators or experts in the relevant domain.

Another aspect we want to point out is the aspect of liability: When pondering the use of AI in intralingual translation, the translator or company also has to keep in mind that the human translator still assumes full liability for the translation (since machines are not liable). This is especially important in "safety-critical domains", which [Canfora and](#)

Ottmann (2020) define as "those domains where translation errors can lead to injuries. Examples of safety-critical domains in translations are health-care, mechanical engineering, the chemical industry and power generation" (Canfora and Ottmann, 2020, p. 61). Thus, in medical discourse there is a high risk of safety-critical errors, which can result in serious damage. One of the reasons why these mistakes are especially dangerous is that post-editors seem to have difficulties to detect them in the raw machine translation output (Canfora and Ottmann, 2020). This underlines the importance of professional post-editing competences. Translators must be trained to detect and correct different types of errors, especially those that are critical for user safety.

Still, by constant training, SUMM AI is currently working towards improving their machine translation system using in-domain data. To investigate whether the trained and improved version of their machine translation system yields better results than the current one, we also plan to conduct a second, comparative study. A machine translation system that has evolved through several iterations and has achieved a satisfactory level of liability, coupled with the professional post-editing skills of a translator or a suitably trained editor would represent a breakthrough for the editorial process:

Editors would be able to publish a much larger volume of texts in Plain Language with greater frequency. Scientific review, however, would still have to be done with the same meticulousness as with human translations. But the translation process would be much faster. This could be a real milestone in the field of accessible health communication. As in the future, all essential questions – even current ones – on diseases, medications and preventive health care should also appear in the accessible and at the same time acceptable form of Plain Language. According to the National Action Plan for Health Literacy (Schaeffer et al., 2018b) this could contribute significantly to promoting health literacy in the population.

## 5. Acknowledgements

This research was funded by the Ministry of Science and Culture of Lower Saxony and its program "zukunft.niedersachsen".

## 6. Bibliographical References

Sarah Ahrens, Rebecca Schulz, Janina Kröger, Sergio Hernández Garrido, Loraine Keller, and Isabel Rink. 2022. Accessible communication and health literacy. *Accessibility–Health Literacy–Health Information: Interdisciplinary*

*Approaches to an Emerging Field of Communication*, 13:9.

Miriam Anshütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language Models for German Text Simplification: Overcoming Parallel Data Scarcity through Style-specific Pre-training. *arXiv preprint arXiv:2305.12908*.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen, Orientierung für die Praxis*. Dudenverlag.

Carmen Canfora and Angelika Ottmann. 2020. Risks in neural machine translation. *Translation Spaces*, 9(1):58–77.

Silvana Deilen. 2021. *Optische Gliederung von Komposita in Leichter Sprache. Blickbewegungsstudien zum Einfluss visueller, morphologischer und semantischer Faktoren auf die Verarbeitung deutscher Substantivkomposita*. Frank & Timme.

Silvana Deilen, Sergio Hernández Garrido, Ekaterina Lapshinova-Koltunski, and Christiane Maaß. 2023. Using ChatGPT as a CAT tool in Easy Language translation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability (TSAR)*, RANLP, Varna, Bulgaria. ACL.

Tomas Goldsack, Zheheng Luo, Qianqian Xie, Carolina Scarton, Matthew Shardlow, Sophia Ananiadou, and Chenghua Lin. 2023. [Overview of the biolaysumm 2023 shared task on lay summarization of biomedical research articles](#). In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 468–477, Toronto, Canada. Association for Computational Linguistics.

Silke Guterath. 2020. *Leichte Sprache für alle?: eine zielgruppenorientierte Rezeptionsstudie zu Leichter und Einfacher Sprache*, volume 5. Frank & Timme GmbH.

Silvia Hansen-Schirra, Jean Nitzke, Silke Guterath, Christiane Maaß, and Isabel Rink. 2020. Technologies for translation of specialised texts into easy language. *Easy Language Research: Text and User Perspectives*. Berlin: Frank & Timme, pages 99–127.

Julian Hörner. 2022. Accessible and reliable health information on the internet. *Accessibility–Health Literacy–Health Information: Interdisciplinary Approaches to an Emerging Field of Communication*, 13:77.

- Roman Jakobson. 1959. On linguistic aspects of translation. In *On translation*, pages 232–239. Harvard University Press.
- Tobias Kopp, Amelie Rempel, Andres P. Schmidt, and Miriam Spieß. 2023. Towards machine translation into easy language in public administrations: Algorithmic alignment suggestions for building a translation memory. In Silvana Deilen, Silvia Hansen-Schirra, Sergio Hernández Garrido, Christiane Maaß, and Anke Tardel, editors, *Emerging Fields in Easy Language and Accessible Communication Research*, pages 371–406. Frank & Timme, Berlin.
- Christiane Maaß. 2020. *Easy Language–Plain Language–Easy Language Plus: Balancing Comprehensibility and Acceptability*. Frank & Timme.
- Christiane Maaß. 2024. Translation into Easy Language. *Handbook of Accessible Communication*, pages 261–289.
- Christiane Maaß and Isabel Rink. 2021. Translating legal texts into easy language. *J. Open Access L.*, 9:1.
- Christiane Maaß, Isabel Rink, and Christiane Zehrer. 2014. Leichte Sprache in der Sprach- und Übersetzungswissenschaft. *Sprache barrierefrei gestalten. Perspektiven aus der Angewandten Linguistik*. Berlin: Frank & Timme, pages 53–85.
- Christiane Maaß and Laura Marie Schwengber. 2022. Easy Language and Plain Language in Germany. *Rivista internazionale di tecnica della traduzione= International Journal of Translation*.
- Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. [Controllable text simplification with explicit paraphrasing](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.
- Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. [Controllable sentence simplification](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.
- Isabel Rink. 2019. *Rechtskommunikation und Barrierefreiheit: Zur Übersetzung juristischer Informations- und Interaktionstexte in Leichte Sprache*. Frank & Timme.
- Horacio Saggion. 2017. Applications of automatic text simplification. In *Automatic Text Simplification*, pages 71–77. Springer.
- Andreas Säuberli, Sarah Ebling, and Martin Volk. 2020. Benchmarking data-driven automatic text simplification for German. In *Proceedings of the 1st workshop on tools and resources to empower people with reading difficulties (READI)*, pages 41–48.
- Carolina Scarton and Lucia Specia. 2018. [Learning simplifications for specific target audiences](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.
- D Schaeffer, EM Berens, S Gille, L Griese, J Klinger, S de Sombre, D Vogt, and K Hurrelmann. 2021. *Gesundheitskompetenz der Bevölkerung in Deutschland vor und während der Corona Pandemie: Ergebnisse des HLS-GER 2*. Interdisziplinäres Zentrum für Gesundheitskompetenzforschung.
- Doris Schaeffer, Svea Gille, and Klaus Hurrelmann. 2020. Implementation of the national action plan health literacy in Germany – lessons learned. *International Journal of Environmental Research and Public Health*, 17(12):4403.
- Doris Schaeffer, Klaus Hurrelmann, Ullrich Bauer, and Kai Kolpatzik. 2018a. Nationaler Aktionsplan Gesundheitskompetenz. *Die Gesundheitskompetenz in Deutschland stärken*. Berlin: KomPart, 10:0418–1866.
- Doris Schaeffer, Klaus Hurrelmann, Ulrich Bauer, and Kai Kolpatzik. 2018b. [National Action Plan Health Literacy. Promoting Health Literacy in Germany](#). *Journal of Public Health*, 31:905–915.
- Doris Schaeffer, Dominique Vogt, Eva-Maria Berens, and Klaus Hurrelmann. 2017. *Gesundheitskompetenz der Bevölkerung in Deutschland: Ergebnisbericht*. Bielefeld: Universität Bielefeld, Fakultät für Gesundheitswissenschaften.
- Kim Cheng Sheang and Horacio Saggion. 2021. [Controllable sentence simplification with a unified text-to-text transfer transformer](#). In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Nicolas Spring, Marek Kostrzewa, David Fröhlich, Annette Rios, Dominik Pfütze, Alessia Battisti, and Sarah Ebling. 2023. Analyzing sentence alignment for automatic simplification of german texts. In *Emerging Fields in Easy Language and Accessible Communication Research*, pages 339–369. Springer.

Birgit Welch and Gabriele Sauberer. 2019. Easy-to-read language and terminology: New needs, new rules, new software? In *Systems, Software and Services Process Improvement: 26th European Conference, EuroSPI 2019, Edinburgh, UK, September 18–20, 2019, Proceedings 26*, pages 647–658. Springer.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. [Optimizing statistical machine translation for text simplification](#). *Transactions of the Association for Computational Linguistics*, 4:401–415.

## 7. Language Resource References

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

# Large Language Models as Drug Information Providers for Patients

Luca Giordano, Maria Pia di Buono

UNIOR NLP Research Group

University of Naples "L'Orientale"

giordanoluca.uni@gmail.com, mpdibuono@unior.it

## Abstract

Recently, a significant interest has arisen about the application of Large Language Models (LLMs) in medical settings to enhance various aspects of healthcare. Particularly, the application of such models to improve knowledge access for both clinicians and patients seems very promising but still far from perfect. In this paper, we present a preliminary evaluation of LLMs as drug information providers to support patients in drug administration. We focus on posology, namely dosage quantity and prescription, contraindications and adverse drug reactions and run an experiment on the Italian language to assess both the trustworthiness of the outputs and their readability. The results show that different types of errors affect the LLM answers. In some cases, the model does not recognize the drug name, due to the presence of synonymous words, or it provides untrustworthy information, caused by intrinsic hallucinations. Overall, the complexity of the language is lower and this could contribute to make medical information more accessible to lay people.

**Keywords:** Large Language Models, Drug Package Leaflets, Italian

## 1. Introduction

Patients' knowledge about medications is crucial as it allows them to administer drugs safely. This knowledge frequently comes from written prescriptions, drug package leaflets, or from reading drug Web pages. Nevertheless, this information has been described as often inconsistent, incomplete, and difficult for patients to read and understand (Shrank and Avorn, 2007). Despite the fact that in 2009 the European Commission issued guidelines<sup>1</sup> to recommend the publication of drug package leaflets with accessible and understandable information for patients, several scholars (Rodríguez et al., 2009; Piñero-López et al., 2016; Segura-Bedmar and Martínez, 2017) account for the absence of improvement in the readability of such documents. Thus, educating patients about their medications seems to be a challenging task due to the linguistic nature of drug written information, which includes a high presence of specialized terms used to describe adverse drug reactions, diseases and other medical concepts that are not easy to understand.

Recently, a significant interest has arisen about the application of Large Language Models (LLMs) in medical settings to enhance various aspects of healthcare, ranging from medical education to clinical decision support (Yuan et al., 2023). A lot of specialized medical LLMs, resulting from entirely new pre-training processes or refinements of existing models, have been made available (Li et al., 2023). Furthermore, several evaluation campaigns

have been conducted to assess general-purpose LLMs in supporting knowledge access from both clinicians and patients (Sun et al., 2023; Xiong et al., 2023; Wang et al., 2023b).

To contribute to the topic, in this paper we investigate the possibility of using LLMs as patient assistants in drug administration. In our opinion, the capabilities of LLMs of providing information about drugs should be evaluated according to two main dimensions, that are the trustworthiness of the provided information and its readability. The former refers to LLM knowledge of drugs, while the latter pertains to the use of a simplified language to support the information access by patients. We conduct our experiment for the Italian language. Our contributions rely on the multidimensional evaluation of LLMs as drug information providers for patients and the release of a domain-specific corpus for the Italian language, namely D-LeafIT (Section 3.1), as result of a Ground Truth (GT) creation for the assessment of such models.

The paper is organized as follows. Section 2 delves into existing research on the topic, providing context for our approach. Section 3 outlines the specific methodology we applied. Section 4 details our experiment settings. Following this, Section 5 presents a thorough analysis of the obtained results. Finally, Section 6 discusses our conclusions based on the findings and outlines potential directions for future research.

## 2. Related Work

Three pivotal research avenues relevant to the present study include LLMs applied to medical and healthcare question answering in general, consumer question answering related to drugs and

---

<sup>1</sup>GUIDELINE ON THE READABILITY OF THE LABELLING AND PACKAGE LEAFLET OF MEDICINAL PRODUCTS FOR HUMAN USE - European Commission, 2009

medications and employing LLMs for medical text simplification.

In the realm of Natural Language Processing (NLP) applied to medical question answering and healthcare information extraction, several recent studies have delved into the capabilities and challenges of leveraging LLMs for these tasks. [Singhal et al. \(2023B\)](#) and [Korgul et al. \(2023\)](#) have explored the potential of LLMs in expert-level medical question answering, shedding light on the nuances and complexities involved in this domain. [Singhal et al. \(2023A\)](#) emphasized how LLMs can encode clinical knowledge, highlighting their significance in medical applications, and released a comprehensive evaluation benchmark called MultiMedQA, which combines six existing medical question answering datasets spanning professional medicine, research and consumer queries with HealthSearchQA, a new dataset of medical questions frequently searched online. [Reddy \(2023\)](#) proposed a framework for evaluating LLMs in healthcare based on translational value assessment, offering a structured approach to assess the practical implications and benefits of these models.

[Abacha et al. \(2017\)](#) and [Abacha et al. \(2019\)](#) have contributed significantly to bridging the gap between consumers' medication questions and trusted answers, and released two datasets, i.e., LiveQA and MedicationQA. [Nguyen et al. \(2023\)](#) introduced MedRedQA, a dataset for medical consumer question answering, contributing to the development of resources tailored to healthcare information retrieval. Additionally, [Gu et al. \(2023\)](#) conducted a case study on distilling LLMs for biomedical knowledge extraction, specifically focusing on adverse drug events, showcasing the practical implications of such models in healthcare.

On the topic of simplifying medical information, [Segura-Bedmar et al. \(2016\)](#) and [Segura-Bedmar and Martínez \(2017\)](#) explored text simplification techniques for drug package leaflets in Spanish, demonstrating efforts to enhance readability and accessibility of crucial healthcare information. Furthermore, [Sakakini et al. \(2020\)](#) addressed context-aware automatic text simplification of health materials in low-resource domains, underscoring the importance of adapting NLP techniques to diverse linguistic contexts for effective communication in healthcare settings. [Simões and Gamallo \(2021\)](#) developed LeMe-PT, a medical package leaflet corpus for Portuguese, contributing to language-specific resources in this domain. Moreover, [Li et al. \(2022\)](#) proposed PharmMT, a neural machine translation approach aimed at simplifying prescription directions, aligning with the broader goal of improving patient understanding and adherence to medication instructions.

These studies collectively underscore the diverse

applications and implications of NLP techniques, particularly LLMs, in transforming healthcare information retrieval, text simplification for patient comprehension, and knowledge extraction from medical texts.

### 3. Methodology

As already mentioned, our methodology aims at assessing the application of LLMs as information providers to support patients' education in drug administration. The proposed LLM assessment is designed to evaluate two main aspects: (i) the trustworthiness of information and (ii) the improvement of accessibility and understandability of such information. The former is evaluated through a two-fold approach: on the one hand, we estimate the semantic similarity between the LLM answers and a GT, formed by the drug package leaflets, and on the other hand we compare the overlapping between medical named entities in the LLM outputs and in the GT. Whereas, the latter aspect is estimated through a readability index of the LLM outputs.

Since posology, namely dosage quantity and prescription, contraindications and adverse drug reactions seem to be the sections most difficult to understand ([Rodríguez et al., 2009](#)), in our experiment we focus on these four types of information. We define different prompts and ask the model to answer medical questions about drug administration.

In this section, we first define the data collection, data cleaning and information extraction procedures that led to the creation of our corpus. Then, prompt design is discussed and two sets of four prompts are presented and explained, which will be used in Section 4 for the experiment. Follows a description of our evaluation methodology, i.e., a cosine similarity-based semantic evaluation. Later, we delve deeper in the evaluation, exploiting Named Entity Recognition (NER) tagging to provide a more fine-grained assessment of model performance by looking at exact-match entity overlapping between the ground truth and the model output. Finally, we compute a readability index for both GT and model outputs and compare them to assess patient's accessibility and comprehensibility.

#### 3.1. Data

AIFA<sup>2</sup>, the Italian Medicines Agency, adheres to Open Data principles and distributes its data and resources under CC-BY 4.0 license, making it freely accessible to anyone. An example of AIFA's Open Data are the pharmaceutical lists and the so-called "transparency lists"<sup>3</sup>, i.e., comprehensive, electron-

---

<sup>2</sup>AIFA - Homepage

<sup>3</sup>AIFA - Pharmaceutical and Transparency Lists

ically accessible lists of various types of pharmaceutical products, drugs and medications together with respective metadata, such as active ingredient, packaging, MA code (Marketing Authorisation) and more. AIFA also maintains a publicly accessible drug database<sup>4</sup> that contains all summaries of product characteristics (SmPCs) and package leaflets (PLs) approved by AIFA or EMA (European Medicines Agency).

**Data Collection** To create our corpus, we first gain access to the pharmaceutical lists. We select two, namely the generic drugs list and the Class A drugs list. A generic drug is defined by EMA as a medicine that is developed to be the same as a medicine that has already been authorised<sup>5</sup>, while Class A drugs in Italy are defined by AIFA as essential medicines and medicines for chronic diseases.<sup>6</sup> We select these two types of drugs because they are widely used and are relevant to our aims.

Since the AIFA drug database does not allow to download all package leaflets in bulk, we write a Python script tailored to the database website to automate data scraping, using the MA code of each pharmaceutical product in the lists as a query. We exploit the Python libraries *pandas*<sup>7</sup>, *requests*<sup>8</sup> and *selenium*.<sup>9</sup> We scrape 2968 generic drug package leaflets and 1299 Class A drug package leaflets for a total amount of 4267.

**Data Cleaning** Since the documents are in PDF format, in order to make them easier to process electronically, we convert them in TSV format using the Python library *PyPDF2*.<sup>10</sup> Furthermore, since many package leaflets refer to the same medication, only in different formats and packages (e.g., 50mg vs 100mg tablets), we consider these as duplicates and remove them, going down to 2037 unique drugs. Finally, we also remove samples with segmentation errors due to file format conversion, ending up with 1819 package leaflets (1439 generic + 380 Class A), among which are 338 unique active ingredients.

**Information Extraction** Drug package leaflets contain a wide range of useful information regarding the specific medication they refer to, such as drug definition and intended use, contraindications, dosage, administration, adverse drug reactions,

Drug type	# of PLs	# of tokens
Generic	1439	6,154,007
Class A	380	1,650,879
<b>TOTAL</b>	<b>1819</b>	<b>7,804,886</b>

Table 1: Corpus Description

storage indications and further information such as content of the package, marketing authorisation, manufacturer and date of approval.

For the scope of the present study, we consider relevant only information about drug definition, intended use, contraindications, dosage, administration and adverse drug reactions. Therefore, some information extraction techniques are needed. Since the structure of this text genre is highly standardized thanks to shared institutional guidelines<sup>11</sup>, we can easily identify the four leaflet sections containing the information of our interest and extract them using regular expressions and heuristics specifically tailored to the typical structure of these texts. It is worth remarking that not all the leaflets present the same structure and section titles. For instance, the first paragraph, which contains the drug definition and its intended use, is usually introduced by an explicit title such as *Che cos'è [DRUG\_NAME] e a cosa serve* (What is [DRUG\_NAME] and what is it used for), while for some drugs different paragraph titles are found, such as *Indicazioni terapeutiche* (Therapeutic indications). These cases required an adjustment to extract the information and align them with the paragraphs of the leaflets that comply with the shared institutional guidelines.

**D-LeafIT Corpus** Our corpus D-LeafIT is made up of 1819 Italian drug package leaflets, among which 1439 refer to generic drugs and 380 to class A drugs. The generic drug leaflets amount to 6,154,007 tokens while the class A to 1,650,879 tokens, for a total amount of 7,804,886 tokens (Table 1).

Each entry in the corpus contains a unique numerical identifier, the drug name, the MA code, the drug class (generic or class A), the text of the whole leaflet, the four relevant paragraphs extracted, the active ingredient and specifications concerning the packaging and format (e.g. number of tablets and unit of measurement). Furthermore, we also provide PoS tagging and relevant NER tagging annotations (Section 3.3). The corpus is publicly available<sup>12</sup> for future use and further exploration by the research community.

<sup>4</sup>AIFA - Banca Dati Farmaci

<sup>5</sup>EMA - Glossary:Generic medicine

<sup>6</sup>Ministry of Health - Drug classes

<sup>7</sup>pandas

<sup>8</sup>requests

<sup>9</sup>selenium

<sup>10</sup>PyPDF2

<sup>11</sup>See the [CMD\(h\) ANNOTATED QRD TEMPLATE FOR MR/DC PROCEDURES](#)

<sup>12</sup>D-LeafIT corpus on GitHub

### 3.2. Prompt Design

We define two different types of zero-shot prompts (ZSP), the first type (ZSP1) contains specific questions, stating clearly the knowledge needs and specifying the context (e.g., that the question deals with a drug); the second type (ZSP2) holds simpler questions without any further specification about the domain. The main reason for designing ZSP1 and ZSP2 is motivated by the goal of simulating different patients' approaches to test the model resilience to variation in user-defined prompts, which is a known challenge for LLMs (Huyen, 2023).

This issue highlights the importance of carefully crafting and monitoring prompts when interacting with LLMs, employing strategies such as setting specific constraints within prompts, ensuring consistent output formats, controlling parameters like output temperature or adopting prompt engineering techniques such as few-shot prompting (Brown et al., 2020) or Chain-of-Thought (CoT) prompting (Wei et al., 2022). However, it has been shown that non-experts and laymen (in our case, any potential patient) tend to struggle with prompt engineering due to incorrect assumptions about LLM capabilities, difficulties in generalizing prompt designs across contexts, and challenges in understanding the behavior of LLMs (Zamfirescu-Pereira et al., 2023). Therefore, we create the second set of prompts to more closely mirror non-expert, potential patient prompting.

Both prompt types are formed by four questions (Q), each one related to the four main sections of the drug package leaflet we are taking into account, i.e., introduction (Q1), precautions (Q2), dosage/administration (Q3) and adverse drug reactions (Q4), as it follows:

#### • ZSP1

- **Q1:** *Cos'è il farmaco [DRUG\_NAME] e a cosa serve?* (What is the drug [DRUG\_NAME] and what is it used for?)
- **Q2:** *Ci sono particolari avvertenze, precauzioni o potenziali interazioni con altri farmaci per il farmaco [DRUG\_NAME]?* (Are there any special warnings, precautions, or potential interactions with other drugs for the drug [DRUG\_NAME]?)
- **Q3:** *Quando, in quali dosi e in che modo devo prendere il farmaco [DRUG\_NAME]?* (When, in what doses, and how should I take the drug [DRUG\_NAME]?)
- **Q4:** *Quali sono i possibili effetti indesiderati del farmaco [DRUG\_NAME]?* (What are the potential side effects of the drug [DRUG\_NAME]?)

#### • ZSP2

- **Q1:** *A cosa serve [DRUG\_NAME]?* (What is [DRUG\_NAME] used for?)
- **Q2:** *Ci sono controindicazioni per [DRUG\_NAME]?* (Are there any precautions for [DRUG\_NAME]?)
- **Q3:** *Come devo assumere [DRUG\_NAME]?* (How should I take [DRUG\_NAME]?)
- **Q4:** *Ci sono effetti collaterali per [DRUG\_NAME]?* (Are there any side effects for [DRUG\_NAME]?)

### 3.3. Quality Evaluation

To evaluate the model answers to our prompts, we consider D-LeafIT corpus the GT and compare the model results against that.

**Trustworthiness assessment** Traditional metrics based on n-gram overlap such as Exact Match, BLEU, ROUGE or METEOR have been shown to be inadequate to evaluate modern generative AI systems in open-ended, free-form question and answering settings given the stochastic nature of such models and the variability of their outputs (Chen et al., 2019). Most importantly, these metrics generally fail to capture semantic nuances such as paraphrasing or synonymy, which, instead, could be of utmost relevance, for example in simplification tasks.

BERTScore (Zhang et al., 2019), on the other hand, is an automatic evaluation metric, used mainly for machine translation, that computes token similarity using contextual embeddings, rather than exact matches as it was for other metrics.

Inspired by BERTScore, we compare the ground truth and the model outputs by computing the cosine similarity of their respective contextual embeddings. We exploit a pre-trained, multilingual, freely accessible Sentence-BERT embedding model.<sup>13</sup>

This model maps sentences and paragraphs to a 768 dimensional dense vector space. However, since the average length of both the ground truth package leaflet sections and the model answers are longer than the embedding model's max sequence length (128 tokens), we compute the text embedding by mean-pooling (also referred to as "chunking" in OpenAI Cookbook guide by de Avila Belbute Peres, 2023), i.e., we generate contextual embeddings for each sentence separately and then derive the whole-text embedding by averaging those of the sentences contained in the text. Finally,

<sup>13</sup>Specifically, we use the model [sentence-transformers/paraphrase-multilingual-mpnet-base-v2](#), available on HuggingFace.



we look at the distribution of similarity scores for answers to both sets of prompts.

To further evaluate the model performance, we also extract named entities from our GT and all the model answers to measure the overlapping score between them. Specifically, we look for unique, exact-matching NER-tagged tokens, which we consider to be a sign of similarity, trustworthiness and accuracy of model performance compared to our GT. Considering the textual genre we are dealing with, we focus on a select few relevant tags. We exploit a fine-tuned, Italian-specific, freely accessible, BERT-based model called `Italian_NER_XXL`.<sup>14</sup> This model is allegedly capable of identifying 52 categories with a 79% accuracy. However, to the best of our knowledge, any further information on fine-tuning data, test data and metrics used is missing at the time of writing. Nonetheless, despite the scarce transparency, we tested this model and manually checked for performance. We choose to use this model because of its accuracy and the relevance of its tags to our domain.

We select 8 tags highly relevant to our aims: (i) ETA (person's age), (ii) MALATTIA (disease), (iii) MEDICINA (drug), (iv) STRENGTH, (v) FREQUENZA (administration frequency), (vi) DURAZIONE, (vii) DOSAGGIO (dosage), (viii) FORMA (e.g., tablet, capsule, injection). After extracting all unique entities from the texts, we compute precision, recall and F1 score to quantify the exact-match overlap of unique entities between the ground truth and the model answers.

Finally, we also compute Spearman's  $\rho$  and Kendall's  $\tau$  correlation coefficients to investigate the relationship between cosine similarity and named entity's overlap F1 scores (see Section 5 for the result discussion).

**Readability assessment** To assess the degree of readability of the model outputs, we compute the Gulepease readability index, which formula is tailored to the Italian language (Lucisano et al., 1988). It is a function of two linguistic variables, i.e., character-based word length and sentence length. Results range from 0 to 100, where 0 means extremely low readability and 100 extremely high readability.

## 4. Experiment

We configure the experiment as a zero-shot, open-ended, free-form, domain-specific QA. As explained in Section 3.2, we intentionally avoid exploiting advanced prompt engineering techniques to more closely simulate real-world user approaches to conversational systems.

<sup>14</sup>[DeepMount00/Italian\\_NER\\_XXL on HuggingFace](https://huggingface.co/DeepMount00/Italian_NER_XXL)

**Model Description** We select the SOTA at the time of this experiment (February 2024) in the Italian landscape of language-specific, open-source LLMs, i.e., Cerbero-7B<sup>15</sup> (Galatolo and Cimino, 2023). Specifically, we exploit `cerbero-7b-openchat-gguf`<sup>16</sup>. It is based on OpenChat 3.5 (Wang et al., 2023a), which was fine-tuned on a large, partly-synthetically generated chat corpus in Italian. It has 7 billion parameters and a context size of 4086 tokens. At the time of writing, evaluation on well-known benchmarks such as SQuAD-it and three tasks of the shared task EVALITA related to toxicity detection, irony detection, and sentiment analysis show that this Italian LLM outperforms all other Italian models, and the authors claim performance on par with or superior to ChatGPT 3.5. Due to limited computational resources, we use the 8-bit quantized version of the model available on HuggingFace, although aware of the decrease in precision and potentially lower performance.<sup>17</sup>

**Environmental Setup** All code is written and compiled in Python 3.10 on Linux Ubuntu 23.10. The model runs locally on an NVIDIA GeForce RTX™ 3060 Laptop GPU with CUDA v12.0.

**Implementation and Inference** The model is implemented using the `llama-cpp-python` framework<sup>18</sup> and all parameters are set to default except the output's maximum token length, which is set to unlimited, i.e., the model stops generating the sequence whenever it would generate the model's own stopword, in this case "[[Umano]]".

During inference, a simple and short system prompt is used every time the model is prompted. The system prompt is *Questa è una conversazione tra un umano ed un assistente AI. L'assistente AI risponde con parole semplici alle domande dei pazienti sui farmaci*<sup>19</sup>. The aim of this system prompt is to guide the model towards the patient's needs with a view to patient-oriented conversational AI systems.

The model is then prompted with the two sets of four prompts, where the placeholder [DRUG\_NAME] is iteratively replaced by one of the 1819 drug names.

## 5. Result Discussion

As mentioned before, we evaluate the result quality according to two criteria, that are trustworthiness, as the result of cosine similarity and named entities overlapping, and readability.

<sup>15</sup>[Cerbero-7B on GitHub](https://github.com/Cerbero-7B)

<sup>16</sup>[galatolo/cerbero-7b-openchat-gguf on HuggingFace](https://huggingface.co/galatolo/cerbero-7b-openchat-gguf)

<sup>17</sup>[HuggingFace documentation - Quantization](https://huggingface.co/docs/llama-cpp-python/quantization)

<sup>18</sup>[llama-cpp-python](https://github.com/jmorgan3/llama-cpp-python)

<sup>19</sup>"This is a conversation between a human and an AI assistant. The AI assistant answers patients' questions about medications in simple words".

**Trustworthiness** Table 2 shows the values for the mean cosine similarity and corrected sample standard deviation. We can notice that overall results span between .70 for Q4 in ZSP1 and Q1 in ZSP2 and .78 for Q3 in ZSP1. The best result is achieved in the ZSP1 setting on Q3 related to dosage/administration. Similarly, in ZSP2 setting the best result relates to the same question. Q1

Prompt	ZSP1	ZSP2
Q1	.71 (s .14)	.70 (s .16)
Q2	.72 (s .10)	.72 (s .11)
Q3	.78 (s .09)	.77 (s .09)
Q4	.70 (s .11)	.73 (s .10)

Table 2: Mean cosine similarity and corrected sample standard deviation  $s$  (between brackets) of answers to ZSP1 and ZSP2 questions

value in ZSP2 is caused mainly by the presence of ambiguous words, that prevent the model to identify [DRUG\_NAME] as a drug. The results of ZSP1 for Q1 are quite similar (i.e., .71), therefore we can assume that specifying that we are asking information about a drug does not always help the model in providing the correct information. Indeed, we notice that in such cases there exists a data interference, which affects the results, despite the fact that there is a well-defined information request. As far as the results are concerned, we notice the presence of four cases, that are: (i) the model is able to recognize the drug and provides the correct answer; (ii) the model recognizes the drug but fails in the generated answer; (iii) the model does not recognize the drug thus it does not provide the correct information; (iv) the model does not recognize the drug and refuses to answer.

The first type of results is considered correct, while

Type	Drug	Info
Case I	Yes	Yes
Case II	Yes	No
Case III	No	Yes
Case IV	No	No

Table 3: Result types

the other three are classified as errors. Case II, namely when the model recognizes the drug but the answer is not correct, is the most common type of error.

With reference to the cases where the model does not recognize the drug, we observe that this error is influenced by the presence of ambiguous drug names. In such cases of ambiguity, the model presents two different answers, i.e., Case III and Case IV, that are related to its ability of solving the phenomenon.

For instance, Case III happens when we prompt Q1

in ZSP1 setting and ask about ABBA (ID 1), specifying that we are seeking for information about a drug. The model produces an output as it follows:

- *ABBA non è un farmaco, ma è un gruppo musicale pop svedese attivo negli anni '70 e '80. I membri sono Benny Andersson, Björn Ulvaeus, Agnetha Fältskog e Anni-Frid Lyngstad. La loro musica è caratterizzata da melodie orecchiabili, armonie armoniche e testi divertenti, rendendo la musica di ABBA popolare in tutto il mondo.*<sup>20</sup>

In the provided answer, we can notice that the model does not take into account our specification, due to some kind of interference in the data. The answer is also affected by an extrinsic hallucination, caused by the presence of additional and not required text (one would expect just the first sentence, not the whole explanation about the music band).

In ZSP2, when we are not making explicit our information need about the drug, the model replies differently and assumes we are asking about an English word, referring to a family or a list of people. In such a case the cosine similarity value is .19.

The way in which the model solves this type of ambiguity seems to be consistent across prompts but not across Qs. This means that when it answers to some other questions, the model is able to recognize the drug and to generate the output accordingly. Thus, when we prompt Q2 in ZSP2, asking information about precautions, ABBA is recognized as a drug and the answer is adequate to the context (Case I).

- *Sì, ABBA è controindicato nelle persone allergiche all'acido ascorbico o ad uno qualsiasi dei suoi componenti. Inoltre, può interagire con altri farmaci, come antibiotici e antiacidi, quindi consultare sempre il medico prima di assumere questo prodotto.*<sup>21</sup>

In such case, the cosine similarity value is .66. Then, in the same ZSP setting, to Q3, the name ABBA is not recognized as drug once again and the model refuses to provide information (Case IV). Indeed the answer states that:

<sup>20</sup>ABBA is not a drug, but is a Swedish pop music group active in the 1970s and 1980s. The members are Benny Andersson, Björn Ulvaeus, Agnetha Fältskog, and Anni-Frid Lyngstad. Their music is characterized by catchy melodies, harmonic harmonies and funny lyrics, making ABBA's music popular all over the world.

<sup>21</sup>Yes, ABBA is not recommended for people who are allergic to ascorbic acid or any of its components. Also, it may interact with other medications, such as antibiotics and antiacids, so always consult with your doctor before taking this product.

- *Abba è il nome di una famosa band rock. Non c'è alcun farmaco chiamato "ABBA". Per chiarire, ABBA non è un farmaco, ma un nome di band rock, quindi non esiste alcun modo di assumere ABBA come farmaco. Se stavi chiedendo come si prendono i farmaci, dovresti chiedere di cosa hai bisogno di prendere e come dovrebbe essere preso.*<sup>22</sup>

The cosine similarity between this answer and the GT is .51. What is interesting in the generated text is the fact that it contains redundant information about ABBA not being a drug and the presence of a suggestion to address the question properly<sup>23</sup>. Finally, if we look at the answer to Q4, the result belongs to Case I, as the model provides information about the adverse drug reactions of ABBA and the cosine similarity with the GT is .78.

We also notice the presence of intrinsic hallucinations, mainly when the model does not recognize the drug name and it proposes an answer that is completely unverified and/or false. These answers are classified as belonging to Case IV. For instance, in the ZSP1 setting to Q4 the model answer that BODINET is not a drug but a city in France (that actually does not exist). To the same Q, in ZSP2, even though we do not specify our information needs, the model replies listing the adverse drug reactions, thus recognizing BODINET as a drug.

Since the cosine similarity is not informative enough, as we can see from the result to Q3 in ZSP2 for ABBA, we also evaluate the presence of named entities overlapping between the GT and the answers, considering this as an additional criterion of trustworthiness (Table 4).

Prompt	ZSP1	ZSP2
Q1	.38 (s .24)	.37 (s .24)
Q2	.33 (s .22)	.32 (s .22)
Q3	.43 (s .21)	.40 (s .21)
Q4	.30 (s .32)	.27 (s .31)

Table 4: Named Entity Overlap mean F1 score and corrected sample standard deviation  $s$  (between brackets) of answers to ZSP1 and ZSP2.

<sup>22</sup>Abba is the name of a famous rock band. There is no drug called "ABBA". To clarify, ABBA is not a drug, but a rock band name, so there is no way to take ABBA as a drug. If you were asking how to take medication, you should ask what you need to take and how it should be taken.

<sup>23</sup>We do not evaluate the text generation in terms of language aspects but it is worth noticing that the use of *band rock* in Italian is not correct. The proposed word sequence seems to be the result of an adjustment translation rule, which switches the word order from adjective+noun to noun+adjective. Nevertheless, in Italian the expression is a loanword and, as such, it preserves the source language word order.

The results for both types of prompts are quite low for all the Qs. Since these scores are calculated against a GT that was automatically created and processed, we consider these as preliminary results, and stress the need for further refinement of the data. Comparing the prompt settings, we observe that ZSP2 prompts present lower scores and Q4 has the worst overlapping (i.e., .27). These results could derive from the use of a simplified language which employs less domain-specific terms and affects the scores. For the answer to Q3 for ABBA, the overlapping is equal to 0, meaning that the cosine similarity score (i.e., .51) does not derive from the presence of informative words, such as named entities related to the category of interest, and might therefore be inflated by other non-informative content, such as *Consulta sempre il medico prima di assumere questo prodotto* (Always consult your doctor before taking this product). To account for such cases, we calculate two correlation coefficients (Table 5) for each of the Qs in both settings.

The selected correlation coefficients, i.e., Spearman's  $\rho$  and Kendall's  $\tau$ , here are used to test the correlation between the cosine similarity scores and the NER overlap F1 scores. In other words, a high correlation between the two variables means a higher trustworthiness of the evaluation metric chosen, i.e., cosine similarity, while a low correlation means a lower trustworthiness of the metric, thus a potential sign of cosine similarity being inflated by non-informative and non-relevant information. We observe that the highest correlation values are on Q1 for both ZSP1 and ZSP2 according to both indices. While the lowest scores are on Q2 for both ZSP1 and ZSP2 according to both indices. This confirms a low variation across ZSP types and a high one across Qs.

**Readability** In the quality evaluation, we also take into account the readability scores for each ZSP and Q and compare them to the scores observed in the GT (Table 6).

This evaluation pertains to the possibility of using LLMs to improve the medical information access for patients. In this evaluation, we do not assess the readability distinguishing between trustworthy and untrustworthy answers, this means that the results from LLM can be affected by the presence of answers not pertaining to the medical domain or the drugs.

In all cases, across ZSP settings and Qs, the readability of the model answers improves with respect to the GT. ZSP2 answers readability is always higher than the ZSP1 readability. For Q1 and Q2, ZSP2 answers are slightly better than ZSP1 ones, which outperform ZSP2 only on Q3. The latter case represents the highest readability score across all

	<b>ZSP1</b> Spearman's $\rho$	<b>ZSP2</b> Spearman's $\rho$	<b>ZSP1</b> Kendall's $\tau$	<b>ZSP2</b> Kendall's $\tau$
Q1	.56, $p < .05$	.59, $p < .05$	.39, $p < .05$	.42, $p < .05$
Q2	.02, $p > .05$	.06, $p < .05$	.02, $p > .05$	.04, $p < .05$
Q3	.30, $p < .05$	.22, $p < .05$	.21, $p < .05$	.15, $p < .05$
Q4	.15, $p < .05$	.14, $p < .05$	.10, $p < .05$	.10, $p < .05$

Table 5: Spearman's  $\rho$  and Kendall's  $\tau$  correlation coefficients between cosine similarities and NER F1 scores of answers to ZSP1 and ZSP2

<b>Prompt</b>	<b>GT</b>	<b>ZSP1</b>	<b>ZSP2</b>
Q1	43.30 ( <i>s</i> 8)	48.80 ( <i>s</i> 7)	49.40 ( <i>s</i> 7.90)
Q2	41.50 ( <i>s</i> 5.80)	48.40 ( <i>s</i> 4.40)	48.80 ( <i>s</i> 6.60)
Q3	52.20 ( <i>s</i> 6.20)	55.60 ( <i>s</i> 5.40)	55.10 ( <i>s</i> 4.80)
Q4	37 ( <i>s</i> 9.40)	43.20 ( <i>s</i> 10.30)	47.50 ( <i>s</i> 31.10)

Table 6: Mean readability score and corrected sample standard deviation *s* (between brackets) of Ground Truth (GT) and answers to ZSP1 and ZSP2 questions

Qs, but still very close to the GT readability score, such as in the answer to Q3 for ID 124 whose readability score is 53.70 in the GT text and 52.20 in the model answer, while the cosine similarity between these texts is .65.

For Q4 answers, the difference between ZSP1 and ZSP2 of both readability scores and standard deviation is quite high (respectively, 4.30 points and 20.80 points). On this question, the improvement of ZSP2 with respect to the GT is the highest across Qs and prompt settings.

The lowest score of readability is found in Q4 for all the three observed settings. This result confirms that the understandability of adverse drug reactions is very low, mainly in the drug package leaflets that form our GT.

For instance, the answers provided by ZSP2 to Q4 for ID 188 shows a readability index of 64.40, for the same ID the GT scores 38.30. Indeed, when comparing the two texts, in the model answer we notice the use of common words (e.g., nausea and vomit) rather than domain-specific terms together with short sentences, while the GT presents medical terms such as glaucoma and epistaxis and it contains a higher number of longer sentences. Nevertheless, in this specific example the model provides untrustworthy information, as it does not list the same adverse drug reactions presented in the GT. This is also confirmed by the cosine similarity score that is .41 and the NER overlap score that is 0.

## 6. Conclusion

In this paper, we discuss our experiment on the capability of general-purpose LLMs to provide trustworthy and simplified information about drug package leaflets in Italian in a zero-shot setting. The results show that different types of errors affect the

LLM answers. In some cases, the model does not recognize the drug name due to the presence of synonymous words or it provides untrustworthy information caused by intrinsic hallucinations. On the other hand, the number of correct results are quite promising, even though an evaluation of the completeness of such answers is required. Overall, the complexity of the language is lower and this could contribute to make medical information more accessible to lay people.

Future directions on this research topic include (i) a deeper evaluation of the available LLMs, including domain-specific ones, also through a comparative perspective, (ii) the collection of more information on errors and (iii) the possibility of fine-tuning a model to reach better results. The application of LLMs as patient assistants to support drug knowledge and ease their administration seems very attractive, however it needs to be evaluated carefully due to the presence of model hallucinations, potentially causing medical malpractice (Vaishya et al., 2023), as any concealed inaccuracies in diagnoses and health advice could lead to severe outcomes (Lee et al., 2023). For these reasons, in the evolving landscape of AI applications in medicine, considerations have been raised regarding the regulatory approval of LLMs as medical devices, highlighting the ethical and legal dimensions associated with deploying such technologies in healthcare settings (Gilbert et al., 2023).

## 7. Acknowledgements

Luca Giordano has been supported by Borsa di Studio GARR "Orio Carlini" 2023/24 - Consortium GARR, the National Research and Education Network.

Maria Pia di Buono has been supported by Fondo FSE/REACT-EU - Progetti DM 1062 del

10/08/2021 "Ricerca a Tempo Determinato di tipo A) (RTDA)". Azione IV.4 - Dottorati e contratti di ricerca su tematiche dell'innovazione/Azione IV.6 - Contratti di ricerca su tematiche Green.

## 8. Bibliographical References

- Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*, pages 1–12.
- Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis R Goodwin, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Bridging the gap between consumers' medication questions and trusted answers. In *MedInfo*, pages 25–29.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Evaluating question answering evaluation. In *Proceedings of the 2nd workshop on machine reading for question answering*, pages 119–124.
- Felipe de Avila Belbute Peres. 2023. Embedding texts that are longer than the model's maximum context length. [https://cookbook.openai.com/examples/embedding\\_long\\_inputs](https://cookbook.openai.com/examples/embedding_long_inputs). Accessed: 28/02/24.
- Federico A Galatolo and Mario GCA Cimino. 2023. Cerbero-7b: A leap forward in language-specific llms through enhanced chat corpus generation and evaluation. *arXiv preprint arXiv:2311.15698*.
- Stephen Gilbert, Hugh Harvey, Tom Melvin, Erik Vollebregt, and Paul Wicks. 2023. Large language model ai chatbots require approval as medical devices. *Nature Medicine*, 29(10):2396–2398.
- Yu Gu, Sheng Zhang, Naoto Usuyama, Yonas Woldeesenbet, Cliff Wong, Praneeth Sanapathi, Mu Wei, Naveen Valluri, Erika Strandberg, Tristan Naumann, et al. 2023. Distilling large language models for biomedical knowledge extraction: A case study on adverse drug events. *arXiv preprint arXiv:2307.06439*.
- Chip Huyen. 2023. Building llm applications for production. [https://huyenchip.com/2023/04/11/llm-engineering.html#prompt\\_versioning](https://huyenchip.com/2023/04/11/llm-engineering.html#prompt_versioning). Accessed: 28/02/24.
- Karolina Korgul, Andrew M Bean, Felix Krones, Robert McCraith, and Adam Mahdi. 2023. Exploring the landscape of large language models in medical question answering: Observations and open questions. *arXiv preprint arXiv:2310.07225*.
- Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.
- Jiazhao Li, Corey Lester, Xinyan Zhao, Yuting Ding, Yun Jiang, and VG Vydiswaran. 2022. Pharmmt: a neural machine translation approach to simplify prescription directions. *arXiv preprint arXiv:2204.03830*.
- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).
- Pietro Lucisano, Maria Emanuela Piemontese, et al. 1988. Gulpease: una formula per la predizione della leggibilità di testi in lingua italiana. *Scuola e città*, pages 110–124.
- Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. 2023. Medredqa for medical consumer question answering: Dataset, tasks, and neural baselines. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–648.
- María Ángeles Piñero-López, Pilar Modamio, Cecilia F Lastra, and Eduardo L Mariño. 2016. Readability analysis of the package leaflets for biological medicines available on the internet between 2007 and 2013: an analytical longitudinal study. *Journal of medical Internet research*, 18(5):e100.
- Sandeep Reddy. 2023. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked*, page 101304.
- Prieto Rodríguez, Ruiz Azarola, Simón Lorda, Barrio Cantalejo, A Danet, et al. 2009. Quality improvement of health information included in drug information leaflets. patient and health professional expectations. *Atencion primaria*, 42(1):22–27.
- Tarek Sakakini, Jong Yoon Lee, Aditya Duri, Renato FL Azevedo, Victor Sadauskas, Kuangxiao Gu, Suma Bhat, Dan Morrow, James Graumlich, Saqib Walayat, et al. 2020. Context-aware automatic text simplification of health materials in

- low-resource domains. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 115–126.
- Isabel Segura-Bedmar and Paloma Martínez. 2017. Simplifying drug package leaflets written in spanish by using word embedding. *Journal of biomedical semantics*, 8:1–9.
- Isabel Segura-Bedmar, Luis Núñez-Gómez, Paloma Martínez Fernández, and Maribel Quiroz. 2016. Simplifying drug package leaflets. In *SMBM*, pages 20–28.
- William H Shrank and Jerry Avorn. 2007. Educating patients about their medications: the potential and limitations of written drug information. *Health affairs*, 26(3):731–740.
- Alberto Simões and Pablo Gamallo. 2021. Leme-pt: A medical package leaflet corpus for portuguese. In *10th Symposium on Languages, Applications and Technologies (SLATE 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023A. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023B. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.
- Yi-Xin Sun, Zi-Ming Li, Jiu-Zuo Huang, Nan-ze Yu, and Xiao Long. 2023. Gpt-4: the future of cosmetic procedure consultation? *Aesthetic Surgery Journal*, 43(8):NP670–NP672.
- Raju Vaishya, Anoop Misra, and Abhishek Vaish. 2023. Chatgpt: Is this version good for healthcare and research? *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 17(4):102744.
- Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023a. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.
- Yuqing Wang, Yun Zhao, and Linda Petzold. 2023b. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *Machine Learning for Healthcare Conference*, pages 804–823. PMLR.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.
- Mingze Yuan, Peng Bao, Jiajia Yuan, Yunhao Shen, Zifan Chen, Yi Xie, Jie Zhao, Yang Chen, Li Zhang, Lin Shen, et al. 2023. Large language models illuminate a progressive pathway to artificial healthcare assistant: A review. *arXiv preprint arXiv:2311.01918*.
- JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–21.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

# Towards Generation of Personalised Health Intervention Messages

Clara Wan Ching Ho<sup>1,2</sup>, Volha Petukhova<sup>1</sup>

<sup>1</sup>Spoken Language Systems Group, Saarland University, Saarbrücken, Germany

<sup>2</sup>Goethe University Frankfurt, Frankfurt am Main, Germany

c.ho@ub.uni-frankfurt.de; v.petukhova@lsv.uni-saarland.de

## Abstract

Self-care is essential in managing chronic diseases when patients could not always be monitored by medical staff. It therefore fills in the gap to provide patients with advice in improving their conditions in day-to-day practices. However, effectiveness of intervention messages in encouraging healthy behaviour is limited, as they are often delivered in the same manner for patients regardless of their demographics, personality and individual preferences. In this paper, we propose strategies to generate personalized health intervention messages departing from assumptions made by theories of social cognition and learning, planned behaviour and information processing. The main task is then defined as a personalised argument generation task. Specifically, an existing well-performing Natural Language Generation (NLG) pipeline model is extended to modulate linguistic features by ranking messages generated based on individuals' predicted preferences for persuasive arguments. Results show that the model is capable of generating diverse intervention messages while preserving the original intended meaning. The modulated interventions were approved by human evaluators as being more understandable and maintaining the same level of convincingness as human-written texts. However, the generated personalised interventions did not show significant improvements in the power to change health-related attitudes and/or behaviour compared to their non-personalised counterpart. Data collected for the model's training was rather limited in size and variation though.

**Keywords:** personalised medicine, health messages generation, content adaptation

## 1. Introduction

In the context of the global aging population and persistent pressure on healthcare providers to lower their service costs, self-care eHealth services that provide health interventions<sup>1</sup> increasingly gaining popularity. Offered health interventions often have however limited effects on patient motivation, therapy compliance and behaviour or attitude change; a personalised approach is necessary (Adler et al., 2016; Kee et al., 2018). The need for personalisation comes from two primary sources that are not necessarily exclusive: gaps in medical and personal knowledge (Rojas, 2021). *Medical knowledge* of patients may be insufficient to understand health intervention texts. Walsh and Volsko (2008) showed that internet-based consumer-health information articles were written above the recommended reading levels for the average adult. *Personal knowledge* of doctors means that they may be not aware of individual preferences, emotional state, social status and lifestyle of their patients. Knowing certain patient characteristics and preferences associated with those characteristics doctors could personalise their messages that are optimal for their patients (Kee et al., 2018).

Modern Artificial Intelligence (AI) systems enable many application scenarios which incorporate automated online interactions. Recent generative

AI methods, in particular Large Language Models (LLM), offer the possibility of building text generation agents such as ChatGPT that can provide personalised content. However, the pre-trained large models are not suitable for specific applications without explicit prompting, instructions, re-training and/or adaption to a particular domain and task.

The study presented in this paper aims first at assessment of personalisation factors that may influence the interaction quality outcome, i.e. effectiveness of intervention messages for decision-making support and high treatment adherence. Our assumptions are based on the key predictions made by established models of planned behaviour, social cognition, learning and information processing. We test these assumptions in human-based study and collect initial data to design our prediction and generation models. A pipeline model is proposed which modulates medical evidence-based arguments extracted from PubMed abstracts with respect to medical and personal knowledge of the patient. Effects of linguistic modulations are evaluated in post-test experiments where human participants rate, rank and select messages as most convincing, understandable, competent and helpful. Interaction effects between participants' personal profiles and message manipulations are assessed.

The paper is structured as follows. Section 2 reviews models of individual and social aspects of decision making and information processing. We identify factors that impact the generation of convincing personalised health interventions. Section 3 introduces related NLG work in the field of person-

<sup>1</sup>According to the World Health Organization (WHO, 2022), self-care interventions are evidence-based tools used to promote and maintain health, prevent disease and cope with illness outside of formal health services.

alisation. In Section 4, our methodology, resources and architecture design are presented. Section 5 discusses pre-testing, implementation and evaluation experiments. Section 7 summarizes our findings, discusses limitations and outlines directions for future research and development.

## 2. Aspects of Decision Making and Behavioural Change

In order to gain patient adherence and motivate them to change their attitude and/or behaviour, it is important to identify what communication strategies are most appealing to them. Knowing patient characteristics and preferences associated with these characteristics help constructing optimal targeted intervention messages. It has been observed that patients prefer a psycho-social model of communication compared to a biomedical model, which is more commonly used by medical personnel (Kee et al., 2018). Thus, along with truthfulness and logical coherence of the arguments proposed in health intervention messages, their effectiveness relates to individual beliefs and intervention delivery aspects. *Planned Behaviour Theory* (Ajzen, 1991) and *Social Learning Theory* (Bandura and Walters, 1977) specify factors behind intentions and decisions to change attitude and behaviour comprising (1) individual attitudes towards behaviour and its outcomes: perceived importance and perceived level of readiness; (2) perceived social norms; and (3) the individual beliefs (confidence) about abilities to perform and control behaviour and its outcome. *Elaboration Likelihood Model* (Cacioppo and Petty, 1984) explains processing of persuasive messages and factors that facilitate potential attitude change associated with them. *Stereotype Content Model* (Cuddy et al., 2008) predicts the emotions associated with social groups on perceived warmth and competence of communicated messages. Theoretical predictions made by these models equip us with initial assumptions concerning the utility of intervention messages in inducing intended potential attitude and behaviour change. Figure 1 provides an overview of the basic assumptions tested in this study.

We assume that the quality of reached outcomes in terms of therapy compliance, motivation and attitude/behaviour change will depend on the content quality of interventions and patient personal characteristics. These two major factors, in their turn, depend on the perceived levels of agency/competence and warmth/communion - the big two of social cognition (Fiske, 2018).

The level of competence of the arguments presented in intervention messages are defined in terms of: (1) quality of the information provided, e.g. level, expert language use and expressed cer-

tainty level; and (2) framing effects, e.g. presenting options in positive terms (survival rates for a procedure) or in negative terms (mortality rates for a procedure). We assume that interventions based on valid medical evidence formulated in professional, concrete and confident language, and appropriate framing effects will be perceived as competent, see also (Guenoun and Zlatev, 2023).

Personal characteristics influencing the perceived levels of competence and warmth concerns general characteristics of the communicators (i.e. power/status, gender and age) and their personality trait profile (i.e. BIG 5; (McCrae, 1992)). Certain personality traits could be associated with higher levels of perceived competence and warmth in humans and agents. In the line with Dubois et al. (2016), we expect a fit effect between levels of competence and warmth of the generated interventions and patient's preferences on outcome quality: if the competence and warmth levels match, the quality would be higher than when there is a mismatch observed. This is compliant with *Elaboration Likelihood Model* (Cacioppo and Petty, 1984), which suggests that potential attitude change in persuasion could be seen as an act of information processing determined by the use of 1) *central route* which involves more cognitive processes and elaborated processing or 2) *peripheral route* which involves heuristics and cues pickups in processing information, based on an individual's motivation and abilities. The theory states that when a person is motivated and able to process a persuasive message that reinforces one's attitude, with a change in cognitive structure, then likely the central route would be taken, resulting in an attitude change. Otherwise, either peripheral route would be taken to process the message leading to a potential attitude shift temporarily, or there would not be an attitude change.

## 3. Related Work on Generation of Natural Language Interventions

Reiter and Dale (1997) proposed a classical NLG pipeline model that has been widely used and modified to suit a range of purposes, generating texts from an abstract goal to specific wordings. The model includes three components: a *Text Planner*, a *Sentence Planner* and a *Linguistic Realiser*.

A more recent modification proposed by Pauws et al. (2019) adapted the data-to-text architecture for medical domain application. Medical reports are generated automatically from patient's clinical outcomes. Other than the three components in the classical pipeline, another layer of data analysis before content determination is added, allowing output to contain different messages according to one's clinical outcomes. Here, medical knowledge



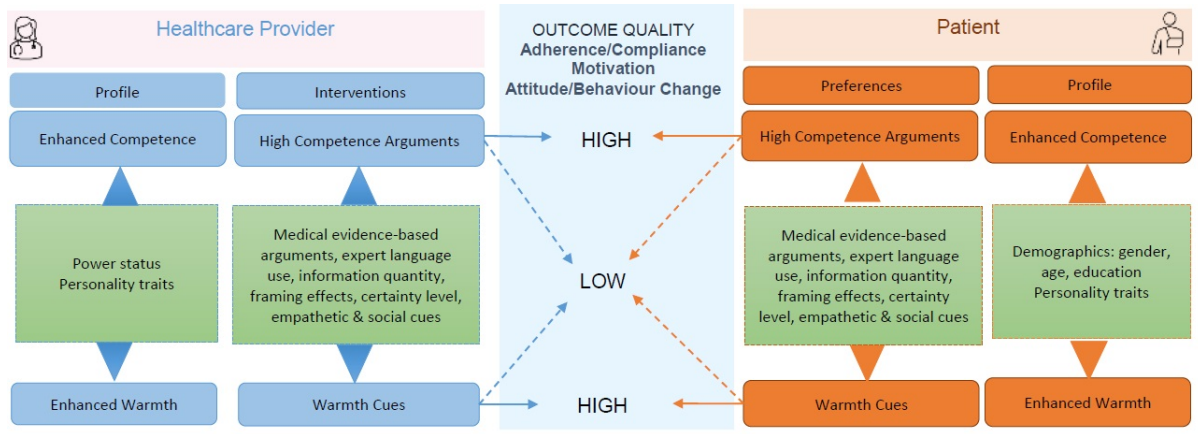


Figure 1: Overview of the key variables and their predicted interplay for health interventions.

was integrated. For example, when one's blood sugar is higher than a threshold as in the knowledge base, a warning message is generated and included in the report.

Mairesse and Walker (2007) implemented a model based on the classical pipeline model of Reiter and Dale (1997) to generate dialogues that mimic people with different levels of extroversion. Language cues for extroverted and introverted people, for example the frequently occurred negations, were considered. Parameters included, but were not limited to, self-references, lexicon frequency, hedge variation and concessions polarity. Two generation approaches were applied: (1) dialogues were generated based on the hypothesised parameters from previous studies, and (2) over-generating dialogues and selecting one that is the most similar to the target level of extroversion. Outputs were evaluated by human raters in terms of their perceived level of extroversion.

Guenoun and Zlatev (2023) compiled a list of linguistic cues as signals of competence and warmth. Their analysis showed that the use of infinitive verbs, nouns and determiners can be considered as accurate signals to quantify competence appeal, and the use of personal pronouns, verbs and wh-determiners as variables to quantify warmth appeal.

We follow the classical pipeline model of Reiter and Dale (1997), taking the approach of Pauws et al. (2019) integrating medical domain knowledge for the persuasive 'competent' content, and the over-generating and matching style applying regression approaches as by Mairesse and Walker (2007). Features studied by Guenoun and Zlatev (2023) are incorporated to quantify perceived competence and warmth appeals.

## 4. Methodology

The domain selected for our use case concerns the treatment of diabetes. To generate health interven-

tions, *data* were collected from PubMed abstracts and reports of the American Diabetes Association with reference to PubMed publications<sup>2</sup>. Data was manually segmented into an argument structure (Mayer et al., 2020), and used for further personalisation. For this, a *pre-test* was designed based on known persuasive strategies, personality traits and linguistic features.

In a *pre-test*, demographic and personality profiles of respondents were collected, along with their judgements of manually modified texts to assess our initial intuition on persuasiveness, understandability, perceived professionalism and perceived helpfulness. Correlations between respondent's personal profiles and linguistic preferences inferred from their judgements were analysed. Discovered effects were considered as parameters predicting one's preferred linguistic delivery of a persuasive intervention.

A pipeline generation model has been proposed to rely on the predicted linguistic parameters related to the perceived competence and warmth. Extracted evidence-based arguments were enriched with alternative medical terms and their definitions from the Unified Medical Language System (UMLS) term bank (Bodenreider, 2004).

Finally, the quality of modulated arguments incorporated into personalised interventions were automatically evaluated and assessed in a *post-test* by human evaluators.

### 4.1. Data and Pre-processing

Medical claims related to self-management actions were extracted, see Table 1 for an example. We assumed that (pre-)diabetic conditions, treatments and prevention measures are publicly relatively well known. According to the Centers for Disease Control and Prevention (2022) large portion of Western population suffers from diabetes, knows somebody

<sup>2</sup>[diabetesjournals.org](https://diabetesjournals.org)

Type	Content
Major claim	You should minimise alcohol intake.
Claim (support)	Minimal alcohol intake lowers health risk for people with diabetes.
Premise (support)	Alcohol intake may place people with diabetes at increased risk for delayed hypoglycemia.

Table 1: Example of a PubMed excerpt as an argument structure of Mayer et al. (2020).

in their family or close social group diagnosed with it or thinks to have sufficient knowledge about the disease. For instance, many studies report that the majority of respondents (up to 97.3% in Italy) had heard about diabetes (Pelullo et al., 2019). Thus, initial attitudes and respective potential changes can be tested rather reliably when assessing the effectiveness of the generated interventions.

From PubMed abstracts of Randomised Controlled Trials (RCT), 16 major claims concerning treatment or life quality improvement actions were selected. Excerpts were manually segmented at clause boundaries, to fit into the argument structure. Table 1 illustrates claims and evidence (premise) to persuade people to “Minimise alcohol intake”. For each major claim supporting and attacking claims and at least one premise were generated resulting in 32 claims and 35 premises in total.

## 5. Experimental Design

### 5.1. Pretest

A pre-test is conducted to collect preferences of respondents with various demographics for different linguistic delivery styles and to test initial assumptions that modulations of linguistic features: (1) are acknowledged by respondents; (2) have effects as predicted by theoretical models; and (3) lead to attitude change.

**Data:** Ten claims were selected for pre-testing: five are supporting the major claim, and other five are attacking those. Six variables, known from previous research, were selected for linguistic modulations and concern *Appeal* (competence/warmth), *Text length* (long/short), *Framing* (risk/benefit), *Lexical complexity* (complex/simple), *Concreteness* (numbers/textual delivery) and *Grammatical voice* (passive/active).

The tested claims were edited manually removing redundancy and generating the missing either attacking or supporting counterpart. This resulted in 12 intervention pairs, where in each pair only one linguistic variable is modulated.

**Questionnaire** has been designed comprising five parts to collect *participants profile*, to assess *pre-intervention attitudes*, to rate *individual inter-*

*ventions*, to compare *pairs of interventions*, and to identify *post-intervention attitudinal change* if any.

To design participant’s profile, information about one’s knowledge/experience with diabetes, gender, age and highest attained education level were collected. Further, participants were asked to complete an online Big Five Personality Test of Open-Source Psychometrics Project<sup>3</sup> Personality profiles corresponding to *extroversion*, *neuroticism*, *agreeableness*, *conscientiousness* and *openness* were computed.

To assess pre- and post-intervention attitudes, respondents were asked to rate on a 7-point Likert scale ten actions that have potentials in improving one’s diabetic conditions.

To assess intervention arguments, respondents were presented one major claim together with a relevant base claim (either attacking or supporting) and a premise, and asked to rate them on how much they agree that the arguments are understandable, helpful, professional and persuasive (7-point Likert scale). In pairwise comparison, respondents were given one major claim and a pair of modulated premises and asked to select one which fits the best the perceived level of the tested effects, e.g. perceived helpfulness.

**Results:** 32 respondents participated in experiments, all English non-native speakers; 58.1% of respondents were female and 38.7% were male; all respondents have at least heard of diabetes as a medical condition; about half of the respondents were between 16 and 30 years old, 32.3% of them between 46 and 60 years old, and 12.9% between 31-45 years old; 90% of the respondents have received at least one bachelor’s degree and over 30% had received at least one postgraduate degree.

The pre-test data has provided useful insights showing that the tested linguistic modulations were perceived by respondents as intended, and can be modelled as parameters in personalised intervention generation. However, it was concluded that a pairwise simple correlation between individual linguistic variable and profile factor is not sufficient to adequately quantify targeted modulation extents. Instead, the interplay between factors should be taken into account when implementing the personalisation model and therefore have contributed to the choice of incorporating random forest models in the pipeline model.

### 5.2. Pipeline Model

The pipeline has two streams, one dealing with the linguistic content (referred to as NLG Stream), the other dealing with the user’s personal profile

<sup>3</sup><https://openpsychometrics.org/tests/IPIP-BFFM/>

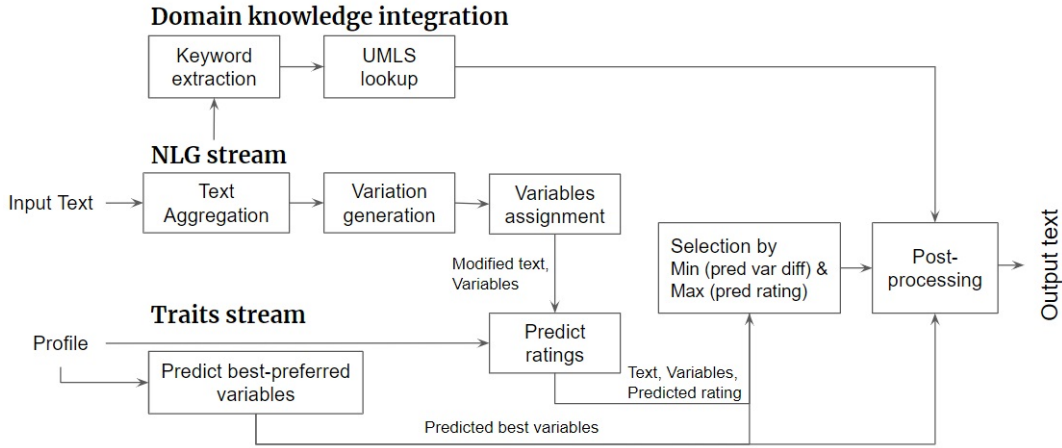


Figure 2: Architecture of the proposed pipeline model.

Parameter	Variables
Appeal to Competence	Average of: VBN tokens / total tokens NN tokens / total tokens DT tokens / total tokens
Appeal to Warmth	Average of: PRP tokens / total tokens VB tokens / total tokens WDT tokens / total tokens
Numeric delivery	local numeric token / max numeric token
Text length	local token count / max token count
Lexical complexity	local average token length / max average toke

Table 2: Parameters to modulate linguistic features. VBN stands for Verb, past participle; NN for Noun, singular or mass; DT for Determiner; PRP for Personal pronoun; VB for Verb, base form; WDT for Wh-determiner.

(referred to as Traits Stream). Additionally, medical domain knowledge (i.e. UMLS) is consulted to look up definitions of medical terms. Decisions to integrate the definition to augment an intervention argument is made at the post-processing step.

In the Traits Stream, information of the user’s profile including age, gender and personality traits scores (Goldberg, 1993) serves as input. Preferences for designated linguistic variables are predicted by two Random Forest regression models. Those are trained on the pre-test data where one model predicts a rating given an individual’s profile and linguistic features, and the other model predicts and ranks linguistic features given an individual’s profile. As a result, weights are assigned to respective linguistic features in an intervention argument and passed for comparison with weights of the generated options in the NLG Stream.

In the NLG Stream, an excerpt of the same major claim and premise serve as input. They are processed by the sentence aggregation component

based on the BART paraphrase model (Lewis et al., 2019) which generates interventions of different lengths with minimal lexical or syntactic changes, and redundant content removed. Repetitive interventions generated are filtered out by Levenshtein distance. The selected interventions are paraphrased with T5 paraphrase model PARROT (Damodaran, 2021). In this way, interventions with a great diversity in lexical, syntactic and potentially semantic changes are (over-)generated. The linguistic variables of the over-generated intervention arguments values are assigned, compared with the predicted values of the corresponding arguments of the Trait Stream and ranked. The best matching intervention, i.e. one with the highest predicted rating and minimal difference between the variables in linguistic features of the predicted preference and generated options, is selected for generation and returned to the user.

### 5.2.1. Linguistic Features for Personalisation

To quantify linguistic features of the generated interventions, five parameters were considered: *appeal to competence*, *appeal to warmth*, *relative level of numeric delivery*, *relative text length* and *relative lexical complexity*, see Table 2. With reference to the output of paraphrase generation, values for the variables were assigned in batches. A batch is the set of paraphrases generated from the same intervention claim or premise. For each batch, the maximal counts of numeric tokens, maximal token count and maximal average token length were computed. The model looped through all entries in the same batch and divided the local counts by the computed maximal counts to assign their relative level of numeric delivery, relative text length and relative lexical complexity, resulting in five values of the five parameters, each between 0 and 1.

To estimate appeal to competence and appeal to warmth, an average of three local variables

as listed in Table 2 were considered. For this, Part-of-Speech (POS) tagging with Python Natural Language Toolkit (nltk) library was performed, and count estimates were computed as explained above. Values of the five parameters, along with a participant's profile, were sent to a Random Forest model to predict a convincingness score for each of the intervention generated.

### 5.2.2. Domain Knowledge Integration

The domain knowledge integration component essentially identifies the key medical concepts in the text and looks up definitions for the respective term in knowledge base. Keywords and phrases were extracted as candidates using KeyBERT model (Grootendorst, 2020) from a pypi package. Subsequently, terms were queried in the Consumer Health Vocabulary (CHV) term bank with UMLS API. Given that the CHV is a medical term bank of common medical terms, if a term was found in CHV, it was considered unnecessary to provide readers with additional information about the term as it is already commonly known. The remaining terms were queried with the UMLS API in the available English medical term banks and the term entry with its respective definitions were retrieved. For simplicity, only the first matching entry was returned. The list of filtered terms, their first matching term entry and their respective definitions were passed on to the final post-processing component for rule-based term-replacement after suitable intervention arguments were generated.

## 6. Evaluation

**Intervention Quality as Texts** The quality of generated interventions was assessed automatically based on cosine similarity and well-formedness. While cosine similarity assesses the degree of semantic information retained in the modulated intervention message, well-formedness assesses its grammaticality.

An off-the-shelf similarity model from sentence transformers (Reimers and Gurevych, 2019) was used. Semantic similarity scores ranging from 0.8 to 0.97 were obtained for all generated intervention messages when compared to their unmodified counterparts. These values indicated that the modulated messages largely retained the information of the original interventions.

The well-formedness was automatically assessed with the off-the-shelf model of Kumar (2020). The unmodified interventions got a mean well-formedness score of 0.5, with a range of approximately from 0.3 to 0.65. The generated modulated interventions exhibited larger variations, where their well-formedness ranged between 0.1 and 0.9. Two-

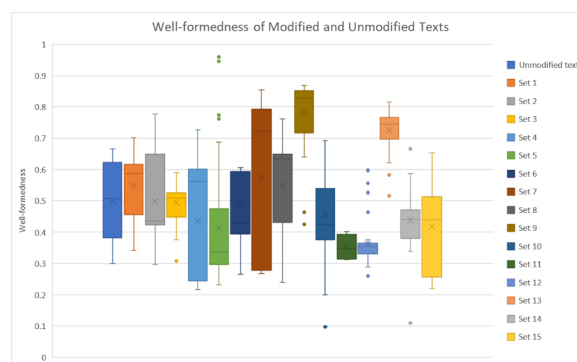


Figure 3: Mean well-formedness of unmodified and 15 sets of modified intervention messages.

thirds of modulated interventions had a lower average score than the unmodified ones. 60% of the modulated intervention messages were within the range of  $\pm 10\%$  of the mean of the unmodified ones. Figure 3 summarises the results.

### Association with Decision Making Aspects

The pipeline model did not specify personalisation strategies in linguistic features modulation, instead interventions were generated by selecting preferences expressed in linguistic cues relevant for decision making aspect. Based on the in-depth analysis and achieved effects such strategies can be defined. The following samples of interventions demonstrate how different syntactic structures, subjectivity, mood, information load and vocabulary use can be associated with perceptions triggered by generated interventions.

From pre-defined prompts of “You should [action]” for major claims, four patterns of subjectivity expressions were observed in modified intervention messages:

- (i) identical to the original input of “You should [action]”;
- (ii) “I recommend you [action]”. Both (i) and (ii) display a higher level of subjectivity and start with a personal pronoun, where (i) is stronger in tone than (ii);
- (iii) “It is recommended to [action]” is seen as objective and neutral; and
- (iv) “The best ... is [recommendation]”, phrases statements with a recommended action in replacement of a potentially harmful one.

Modified prompts which differ in subjectivity are presented in (1):

- (1) You should lose some weight.  
I recommend you lose weight.  
It is recommended to extend the time spent sleeping.  
The best replacement for sugar-sweetened drinks is water.

Apart from subjectivity, generated interventions differed in mood, including indicative, imperative, conditional and interrogative as exemplified in (2) .

- (2) It is important to do resistance training and aerobic exercises. (indicative)  
Take zinc supplements to slow the development of diabetes. (imperative)  
If you want your health to improve you should take supplements that contain b12. (conditional)  
Do you have to do a balance exercise? Short-term structured strength and balance training did not affect HRQoL; there were no significant differences between groups on the primary outcomes of PCS score and EQ-5D-5L index score. (interrogative)

Modified intervention messages A and B below in (3) show how the diversity in language cues may encourage intended attitude hence behavioural change.

- (3) **Unmodified Intervention Message**  
You should minimize alcohol intake. Minimal alcohol intake lowers health risk for people with diabetes. Alcohol intake may place people with diabetes at increased risk for delayed hypoglycemia. This is particularly relevant for those using insulin or insulin secretagogues who can experience delayed nocturnal or fasting hypoglycemia after evening alcohol consumption.

#### **Modified Intervention Message A**

Reduce the quantity of alcohol. Recommended for those using insulin or insulin secretagogues who experience delayed nocturnal or fasting Hypoglycemia (Abnormally low level of glucose in the blood) after evening alcohol consumption.

#### **Modified Intervention Message B**

It is important that you cut down on your alcohol consumption. This is particularly relevant for those using insulin or insulin secretagogues that may experience delayed nocturnal or fasting Hypoglycemia (Abnormally low level of glucose in the blood) after evening alcohol consumption.

Imperative and indicative moods are observed respectively at sentence beginnings of the two messages. The imperative mood in A conveys a sense of certainty and urgency, relevant to a higher level of perceived readiness according to the Planned Behaviour Theory (Ajzen, 1991). The expression “it is important that you” in B is related to the increase of perceived importance in the aspect of behavioural intention. Intervention Message B signals closeness with addressees when using personal pronouns, encouraging an in-group association of the warmth appeal in the Stereotype Content Model (Cuddy et al., 2008).

Diversity in vocabulary use is observed, such as the replacement of “minimize” to “cut down on” and “reduce”. They can be seen as presentations

that are more or less colloquial, establishing different levels of closeness, which is relevant for the competence/warmth appeal.

If parameters are set correctly, the model can personalise texts with a great variation in linguistic features to closely match individual linguistic preferences or targeted perception effects.

## **6.1. Post-test**

To assess the intended effects of personalisation, understandability, likeability and convincingness and the quality of the generated interventions, a post-test has been conducted. We also assessed naturalness, perceived redundancy and coherence of the generated messages.

**Data** From the 16 major claims presented earlier, 15 were selected for the post-test: eight expressing the supporting stance, and the other seven the attacking stance. The base claims were used as unmodified interventions and proposed for personalisation.

Three types of interventions were tested: (1) unmodified arguments from medical excerpts; (2) the best matching intervention generated by the pipeline model and matching the individual preferences; and (3) the worst matching intervention generated by the pipeline model and selecting the least matching individual profile. Note that the best matching and worst matching interventions vary for each participant as they were generated based on their individual profiles.

Five parameters (appeal to competence, appeal to warmth, relative level of numeric delivery, relative text length and relative lexical complexity) were modulated.

**Questionnaire Design** The post-test was distributed as a questionnaire with two parts: (1) collection of participants’ profiles in terms of their demographics and personality traits and is identical to that of the pre-test; and (2) a total of 15 sets of personalised and at least one non-personalised interventions were *ranked* and *rated* on a 7-point Likert scale. Additionally, the level of *information retention* was assessed. Three randomly selected unmodified, best and worst matching intervention messages were evaluated on criteria of well-formedness (coherent and natural), understandability, redundancy and likeability (convincing). The later criteria were meant to test some of the study’s hypotheses in the perception of personalised linguistic delivery.

**Results** 21 respondents participated in the study. All respondents were required to have at least heard of diabetes as a medical condition and have not

Evaluation Criterion	Preference Matching Setting		
	Best	Worst	Unmodified
<i>Text Quality Evaluation</i>			
Coherence	4.73* ( $\pm 1.5$ )	4.70* ( $\pm 1.6$ )	5.67 ( $\pm 1.5$ )
Naturalness	4.36* ( $\pm 0.2$ )	3.93* ( $\pm 0.1$ )	5.83 ( $\pm 0.3$ )
Redundancy	2.20* ( $\pm 1.6$ )	2.53* ( $\pm 1.5$ )	3.17 ( $\pm 1.7$ )
<i>Perception Evaluation</i>			
Likeability	4.37 ( $\pm 1.7$ )	4.31 ( $\pm 1.4$ )	4.53 ( $\pm 1.6$ )
Understandability	4.86* ( $\pm 1.9$ )	5.27 ( $\pm 1.7$ )	4.33 ( $\pm 1.7$ )
Convincingness	4.73 ( $\pm 1.5$ )	4.70 ( $\pm 1.5$ )	5.67 ( $\pm 1.8$ )

Table 3: Overview of the average ratings in text quality and perception evaluation experiments on the 7-point Likert scale. \* marks statistically significant differences when compared to an unmodified intervention argument. Standard deviation is provided in brackets.

participated in the pre-test. 15 respondents successfully completed both parts of the questionnaire, all of them were between 16 and 30 years old, with 53.3% male and 46.7% female; 93.3% of the respondents have received at least one bachelor’s degree and 13.3% have received a master’s degree.

The results showed that the generated interventions were rated as more understandable than unmodified ones, see Table 3. This is most probably due to the simplification and added definitions of medical terms. Results show statistically significant differences where the best matching rated approximately 30% higher than unmodified ones in understandability ( $p=0.049$ ).

Likeability fluctuates between test sets (Figure 4), which may be a result of the instability in paraphrasing quality. Nevertheless, results show that the likeability of the generated interventions is at least competitive with that of the unmodified arguments. The mean ratings of the five sets show that the three types of interventions were rated similarly in terms of likeability, where the unmodified ones receive the highest and the worst matching ones the lowest scores.

In both rating in ranking tasks, unmodified interventions are rated as the most redundant and best matching texts are the least redundant.

The rated naturalness and coherence of the modified interventions are noticeably lower than human-written unmodified texts. The results are understandable given the lack of grammatical and semantic check in selection of paraphrases. There are no statistically significant differences observed in convincingness of generated modulated and unmodified interventions, suggesting that the automatically generated messages are at least not less convincing after the performed modulations.

According to the post-test results, interventions generated by the model are in general less redundant, more understandable and as likeable and convincing as the unmodified arguments. However they are less natural and potentially less coherent.

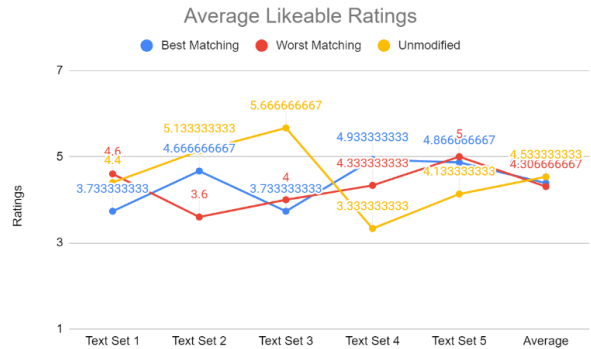


Figure 4: Average likeability ratings.

## 7. Discussion and Conclusion

This study evaluated the argument generation approach for medical domain application in personalising intervention messages. A pipeline model was implemented to process health interventions containing medical evidence based arguments and convert them into personalised health intervention messages. Medical domain knowledge is integrated to simplify and explain medical terms for higher understandability.

The implemented model was evaluated and produced good quality health interventions. Despite perceived as less natural, modulated interventions were rated by human evaluators as likeable and convincing as the unmodulated ones, while performing better on understandability and conciseness criteria.

Modulated interventions exhibited a high diversity in lexical and syntactic structures given different profiles, which potentially can be used to specify various persuasion strategies. Currently, no module in the model that defines or selects persuasive strategies is designed.

Further work is required to improve system’s personalisation capabilities. Personalisation factors are numerous and show complex interplay, these additional effects need to be analysed in a more detailed study with higher number of participants of various demographics, personalities and emotional states. Real patient data recorded in authentic doctor-patient communicative settings is ideal but hard to obtain. We, therefore, opt for better patient simulations and simulations of interactive situations which will allow better experimental control.

## 8. Acknowledgments

The authors are also very thankful to anonymous reviewers for their valuable comments.

## 9. Bibliographical References

- Rachel F Adler, Francisco Iacobelli, and Yehuda Gutstein. 2016. Are you convinced? a wizard of oz study to test emotional vs. rational persuasion strategies in dialogues. *Computers in Human Behavior*, 57:75–81.
- Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211.
- Albert Bandura and Richard H Walters. 1977. *Social learning theory*, volume 1. Englewood cliffs Prentice Hall.
- Olivier Bodenreider. 2004. [The unified medical language system \(umls\): integrating biomedical terminology](#). *Nucleic Acids Res.*, 32(Database-Issue):267–270.
- John T Cacioppo and Richard E Petty. 1984. The elaboration likelihood model of persuasion. *ACR North American Advances*.
- Centers for Disease Control and Prevention. 2022. [National diabetes statistics report](#). Technical report.
- Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149.
- Prithviraj Damodaran. 2021. Parrot: Paraphrase generation for nlu.
- David Dubois, Derek D Rucker, and Adam D Galinsky. 2016. Dynamics of communicator and audience power: The persuasiveness of competence versus warmth. *Journal of Consumer Research*, 43(1):68–85.
- Susan T Fiske. 2018. Stereotype content: Warmth and competence endure. *Current directions in psychological science*, 27(2):67–73.
- Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist*, 48(1):26.
- Maarten Grootendorst. 2020. [Keybert: Minimal keyword extraction with bert](#).
- Bushra S Guenoun and Julian J Zlatev. 2023. Sending signals: Strategic displays of warmth and competence. *Working Paper 23-051*.
- Janine WY Kee, Hwee Sing Khoo, Issac Lim, and Mervyn YH Koh. 2018. Communication skill in patient-doctor interactions: learning from patient complaints. *Health professions education*, 4(2):97–106.
- Ashish Kumar. 2020. [Query wellformedness scoring](#).
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. [Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#).
- François Mairesse and Marilyn Walker. 2007. Personage: Personality generation for dialogue. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 496–503.
- Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020*, pages 2108–2115. IOS Press.
- RR McCrae. 1992. Revised neo personality inventory (neo-pi-r) and neo five-factor inventory (neo-ffi) manual. *Psychological Assessment Resources*. Odessa, FL.
- Steffen Pauws, Albert Gatt, Emiel Krahmer, and Ehud Reiter. 2019. Making effective use of healthcare data using data-to-text technology. In *Data Science for Healthcare*, pages 119–145. Springer.
- Concetta P Pelullo, Riccardo Rossiello, Roberto Nappi, Francesco Napolitano, Gabriella Di Giuseppe, et al. 2019. Diabetes prevention: knowledge and perception of risk among italian population. *BioMed research international*, 2019.
- Nils Reimers and Iryna Gurevych. 2019. [Sentencebert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.
- Julian Andres Ramos Rojas. 2021. *Exploring AI-based personalization of a mobile health intervention and its effects on behavior change, motivation, and adherence*. Ph.D. thesis, Carnegie Mellon University Pittsburgh, PA.
- Tiffany M Walsh and Teresa A Volsko. 2008. Readability assessment of internet-based consumer health information. *Respiratory care*, 53(10):1310–1315.
- WHO. 2022. [Self-care interventions for health](#).

# Analysing Emotions in Cancer Narratives: A Corpus-Driven Approach

Daisy Monika Lal<sup>1</sup>, Paul Rayson<sup>1</sup>, Sheila Payne<sup>2</sup>, Yufeng Liu<sup>3</sup>

<sup>1</sup>UCREL Research Centre, Computing and Communications, Lancaster University, UK

<sup>2</sup>Health Research, Lancaster University, UK

<sup>3</sup>Linguistics and English Language, Lancaster University, UK

{d.m.lal, p.rayson, s.a.payne, yufengliu}@lancaster.ac.uk

## Abstract

Cancer not only affects a patient's physical health, but it can also elicit a wide spectrum of intense emotions in patients, friends, and family members. People with cancer and their carers (family member, partner, or friend) are increasingly turning to the web for information and support. Despite the expansion of sentiment analysis in the context of social media and healthcare, there is relatively less research on patient narratives, which are longer, more complex texts, and difficult to assess. In this exploratory work, we examine how patients and carers express their feelings about various aspects of cancer (treatments and stages). The objective of this paper is to illustrate with examples the nature of language in the clinical domain, as well as the complexities of language when performing automatic sentiment and emotion analysis. We perform a linguistic analysis of a corpus of cancer narratives collected from Reddit. We examine the performance of five state-of-the-art models (T5, DistilBERT, Roberta, RobertaGo, and NRCLex) to see how well they match with human comparisons separated by linguistic and medical background. The corpus yielded several surprising results that could be useful to sentiment analysis NLP experts. The linguistic issues encountered were classified into four categories: statements expressing a variety of emotions, ambiguous or conflicting statements with contradictory emotions, statements requiring additional context, and statements in which sentiment and emotions can be inferred but are not explicitly mentioned.

**Keywords:** Clinical narratives, medical language processing, social media, cancer, sentiment analysis, emotion analysis

## 1. Introduction

Cancer is one of the most prevalent diseases impacting the lives of millions of individuals and families worldwide. According to cancer statistics, 14.1 million people worldwide are affected by the disease (Torre et al., 2015). A cancer diagnosis can be upsetting and cause challenging psychological reactions in patients, including anxiety, despair, isolation, and feelings of shame and self-blame. Some individuals may experience heightened emotions contemplating the emotional impact of this news on their loved ones (Muzzin et al., 1994; Ahn et al., 2009; Singer et al., 2010; Cho et al., 2013; Al-Azri et al., 2014a,b). According to estimates, up to one-third of cancer patients receiving hospital treatment also suffer from a prevalent mental health issue (Singer et al., 2010). To provide effective therapy for cancer patients it is important to monitor their emotional state and we also aim to support cancer patients, their families, and healthcare providers to better understand their options. Analysing emotions and sentiments is one part of the evidence base to support patients' treatment and care choices, at each stage of disease and treatment. Experts feel that focusing on patients' emotions can improve their health, self-efficacy (patient engagement or involvement in improving the

quality of healthcare) (Lacy, 2016; Marzban et al., 2022), and well-being, hence assessing their mood is an important part of their treatment (Ryan et al., 2005; Harvey and Lawson, 2009).

Social media platforms have become more and more prevalent in providing a common place for patients and their loved ones to express their experiences with cancer (Bender et al., 2011, 2013; Kent et al., 2016; Domínguez and Sapiña, 2017). As a result, social media data can be used to examine the way patients and carers (family member, partner, or friend) talk about their journeys. To facilitate this large-scale analysis, we can use NLP approaches like sentiment analysis (SA), which have evolved over time from fundamental concepts to powerful deep learning (DL) algorithms that are becoming a valuable tool for a variety of NLP applications. Reddit has a huge collection of forums covering news, discussion, entertainment, and just about any topic. Through a network of discussion boards known as subreddits, hundreds of millions of active users regularly share their unfiltered opinions, experiences, ideas, and feelings on a wide range of topics.

In this exploratory study, we analyse the moods and attitudes among cancer patients and their loved ones by acquiring posts from various cancer-related forums on Reddit. We conducted both qualitative



and quantitative assessments to explore cancer-related attitudes and emotions for different cancer stages (I, II, III, IV) and cancer treatments (diagnosis, clinical trials, chemotherapy, radiation therapy, targeted therapy, and palliative care)<sup>1</sup>, as well as to see how closely the automated techniques match manual annotation. Moreover, we dive deeper into challenges associated with medical or clinical data processing.

## 2. Related Work

A wide range of modern techniques, including rule-based, conventional machine learning, and more advanced deep learning approaches, have been effectively applied to the task of SA in health and well-being (Zunic et al., 2020). Prior research, however, reveals that relatively few attempts have been made to use NLP to conduct a large scale examination of cancer patients' views during their journey, specifically the course of their treatment. The majority of research in this area focuses on patient experiences (positive, negative, and neutral) to enhance patient satisfaction (how satisfied patients are with a specific type of treatment as opposed to how they feel during that treatment).

Aspect-based Sentiment Analysis has previously been applied to the healthcare industry, most notably for SA of social media data on drug reviews (Gräßer et al., 2018; Sweidan et al., 2021), the COVID-19 disease and its vaccination (Aygün et al., 2021; Jang et al., 2021; Chaudhary et al., 2023), as well as psychological clinical records concerning suicide (George et al., 2021). Gräßer et al. (2018) performed several experiments related to drug reviews using data scraped from online drug review websites. The study involved looking at people's attitudes toward their overall experiences, side effects, and the usefulness of certain treatments. It also addressed the problem of the absence of annotated data and looked into the transferability of learned classification models across domains. (Sweidan et al., 2021) aimed to create a hybrid ontology-XLNet transfer learning strategy for identifying Adverse Drug Reactions (ADRs) from social data using sentence-level ABSA.

## 3. Methodology

This section provides an elaborate discussion of the strategy adopted for sentiment and emotion classification of the cancer-related Reddits.

<sup>1</sup>Data and Code are available at <https://github.com/4dpicture/Emotion-Analysis>.

Table 1: Cancer aspect-specific post-sets. Here, Min and Max represent the minimum and maximum word length of the posts.

Aspect Category	Total Count	Min (words)	Max (words)
Diagnosis	224	22	2396
Clinical Trials	232	30	3829
Surgery	230	11	2396
Chemotherapy	227	22	2396
Palliative Care	227	43	3554
Radiation Therapy	228	27	3829
Targeted Therapy	121	22	3829
Stage I	53	45	2396
Stage II	15	62	3829
Stage III	20	62	691
Stage IV	142	37	2396

### 3.1. Data Collection

We collected a sample of cancer-related English Reddit posts using PRAW<sup>2</sup>(Python Reddit API Wrapper) from several subreddits including 'r/cancer', 'r/cancersurvivors', 'r/cancerfamilysupport', and 'r/cancercaregivers'. We acquired around 1,500 public posts using cancer-specific (treatment and stage) aspects. For each aspect, we created a separate post-set comprised of all posts containing the aspect term as shown in Tables 1. As people discussed their journeys in these narratives, several posts featured overlapping aspects, such as diagnosis, which was nearly always reported, multiple stages and treatments, chemotherapy and surgery occurring concurrently in many posts, and so on. Because most posts contain several terms, selecting those covering a single aspect, for example, "stage 1" or "palliative care" makes it challenging to generate an independent set with adequate data samples. This yields significantly fewer data points for each set. To resolve this conflict, we just used the aspect-term to search the subreddits, and posts that had multiple aspects were added to all sets.

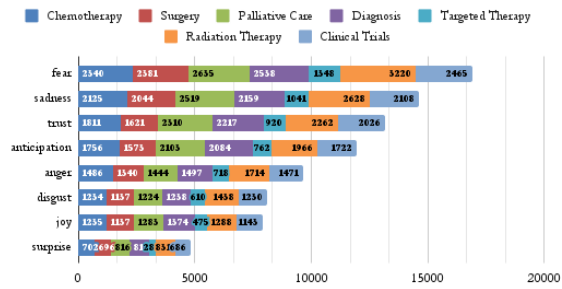
### 3.2. Sentiment and Emotion Classification Models

The idea is to analyse the sentiments or emotions keeping in mind the entire narrative (post). Since nearly all of the state-of-the-art algorithms have word count restrictions and because the entire narrative cannot be adequately tagged using a single sentiment or emotion, sentence-level classification was used to assess the attitudes and emotions in the post-sets. To examine the relative sentiments (positive, negative, or neutral) and emotions (sad-

<sup>2</sup>PRAW available at <https://praw.readthedocs.io/en/latest/>.

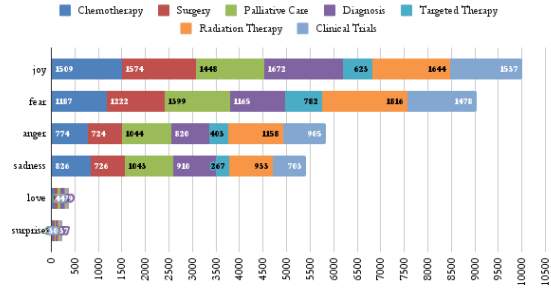
Figure 1: Cancer treatment-specific sentence-level emotion classification.

**NRCLex Emotions - Cancer Treatments**



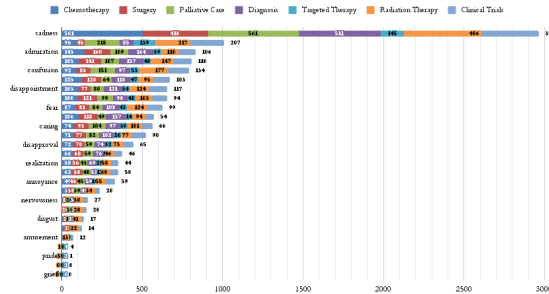
(a) Emotion classification using NRCLex. Emotion classes include fear, sadness, trust, anger, disgust, joy, surprise, and anticipation.

**T5 - Cancer Treatments**



(b) Emotion classification using T5. Emotion categories include joy, fear, anger, sadness, love, and surprise.

**RobertaGo Emotions - Cancer Treatments**



(c) Emotion classification using RobertaGo Emotions with 27 emotion classes.

ness, anger, joy, surprise, etc.) from the gathered post-sets, five different models (T5, DistilBERT, Roberta, RobertaGo, and NRCLex) were utilized. The models are characterized as follows: **T5**<sup>3</sup>: Google’s T5 (Raffel et al., 2020) is an emotion detection model fine-tuned using the emotion recognition dataset introduced in (Saravia et al., 2018). It provides six emotion classes: sadness, joy, love, rage, fear, or surprise. **DistilBERT**<sup>4</sup>: DistilBert (Sanh et al., 2019) was built using the Stanford Sentiment Treebank (SST) (Socher et al., 2013) composed of 11,855 single sentences collected from movie reviews. It classifies text into two categories: positive and negative emotions. **Roberta**<sup>5</sup>: Roberta was fine-tuned for the SA task using TweetEval (Rosenthal et al., 2017), which contains roughly 124M tweets from January 2018 to Decem-

ber 2021. This provides three sentiment classes: positive, negative, and neutral. **RobertaGo**<sup>6</sup>: RobertaGo (Liu et al., 2019) is a multi-label classification model tweaked on the largest manually annotated dataset, Go-Emotions (Demszky et al., 2020) consisting of 58k English Reddit comments, labeled for 27 emotion categories. **NRCLex**<sup>7</sup>: NRCLexicon is a PyPI project designed to gauge ten emotion categories. It is created using the NRC emotion lexicon (Mohammad and Turney, 2013) and the WordNet synonym sets from the NLTK library. It provides eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) and two sentiments (negative, positive).

**4. Key Findings**

This section reports the findings of our experiments. All percentages for a specific post-set were calcu-

<sup>3</sup>"t5-base-finetuned-emotion" available at <https://huggingface.co/mrm8488/t5-base-finetuned-emotion>.

<sup>4</sup>"distilbert-base-uncased-finetuned-sst-2-english" available at <https://huggingface.co/distilbert-base-uncased-finetuned-sst-2-english>.

<sup>5</sup>"twitter-roberta-base-sentiment-latest" available at <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest>.

<sup>6</sup>"roberta-base-go-emotions" available at [https://huggingface.co/SamLowe/roberta-base-go\\_emotions](https://huggingface.co/SamLowe/roberta-base-go_emotions).

<sup>7</sup>"Lexicon source 2016 National Research Council Canada (NRC)" available at <http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>.

Table 2: Examples of statements in the cancer narratives tagged with counter-intuitive sentiment (positive) and emotions (joy, happiness, etc.) by the models.

<p><b>"Going into town I will enjoy the delightful offerings from del taco and a Xanex. We dine like kings!"</b>          "T5": "&lt;pad&gt;joy"   "RobertaGo": "{label: 'joy', 'score': 0.8974}"   "Roberta": "positive = 0.9764"            "DistilBERT": "POSITIVE"   "NRCLex": "trust = 0.25, positive = 0.25, joy = 0.25, anticipation = 0.25"</p>
<p><b>"I made him breakfast, and dinner and tried to keep a positive mindset and just do the happy stuff we always loved doing, jamming, joking, eating waffles."</b>          "T5": "&lt;pad&gt;joy"   "RobertaGo": "{label: 'joy', 'score': 0.6846}"   "Roberta": "positive = 0.9417"            "DistilBERT": "POSITIVE"   "NRCLex": "positive = 0.5714"</p>
<p><b>"I told her I loved her and held her hand while they ended life support."</b>          "T5": "&lt;pad&gt;love"   "RobertaGo": "{label: 'love', 'score': 0.9029}"   "Roberta": "positive = 0.7132"            "DistilBERT": "POSITIVE"   "NRCLex": ""</p>
<p><b>"I'm still battling side effects but I'm so relieved to be done! I get to come home to my beautiful wife and loving dog every time."</b>          "T5": "&lt;pad&gt;joy"   "RobertaGo": "{label: 'joy', 'score': 0.7250}"   "Roberta": "positive = 0.9543"            "DistilBERT": "POSITIVE"   "NRCLex": "positive = 0.3333, joy = 0.3333"</p>
<p><b>"But hearing the phrase, 'you no longer have cancer' makes me feel invincible again."</b>          "T5": "&lt;pad&gt;joy"   "RobertaGo": "{label: 'joy', 'score': 0.3020}"   "Roberta": "positive = 0.7231"            "DistilBERT": "POSITIVE"   "NRCLex": "fear = 0.2857, negative = 0.2857"</p>

lated using the formula:

$$S_j^i/E_j^i = \left( \frac{N_{S_j^i/E_j^i}}{T_j} \right) \times 100 \quad (1)$$

where,  $i$  refers to the  $i^{th}$  sentiment or emotion  $\in$  {positive, negative, neutral, sadness, anger, joy, surprise, fear, etc.},  $j$  refers to the  $j^{th}$  aspect  $\in$  {stage 1, stage 2, ..., targeted therapy, surgery, and palliative care},  $N_{S_j^i/E_j^i}$  refers to the number of sentences tagged with sentiment or emotion  $i$  for aspect  $j$ , and  $T_j$  refers to the total number of sentences containing aspect  $j$ .

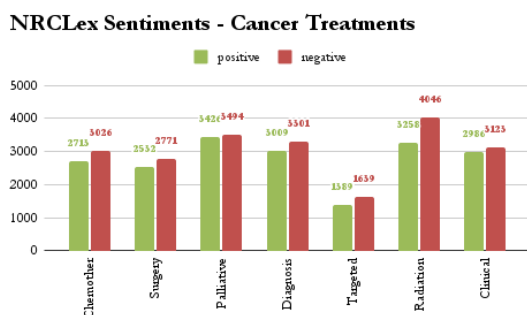
#### 4.1. Cancer Treatments

**Sentiment Analysis:** According to DistilBert (Figure 3b), negative sentiment is most typically relayed across all phases of treatment, in the order, diagnosis (64.6%), surgery (64.5%), chemotherapy (66.2%), palliative care (71.3%), clinical trials (72.4%), radiation therapy (72.4%), and targeted therapy (75.0%), indicating that patients have the greatest quantity of negative emotions during radiation therapy, followed by clinical trials and palliative care. Roberta (Figure 3c) detected more neutral feelings across multiple treatment aspects, clinical trials (45.0%), radiation therapy (44.2%), targeted therapy (51.8%), and palliative care (43.3%) while negative for the others. Across all phases, NRCLex (Figure 3a) detected more negative sentiments, diagnosis (52.3%), clinical trials (51.1%), chemotherapy (52.7%), surgery (52.3%), radiation therapy (55.4%), targeted therapy (54.1%), and palliative care (50.5%), with a narrow gap between the categories. Roberta,

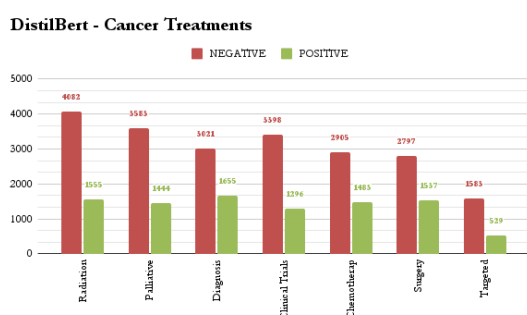
like DistilBert, has a wide disparity between the sentiment categories, with radiation therapy eliciting the most negative sentiment, followed by palliative care, diagnosis, clinical trials, and chemotherapy. In comparison to other aspects, Roberta indicated greater positive sentiments about diagnosis and surgery (see Table 2).

**Emotion Analysis:** According to the NRCLex model (Figure 2a), the two emotions that happen to surface most consistently throughout all treatment-specific aspects are *fear* (18 - 22%) and *sadness* (15 - 17%). Other emotions include *trust* (13 - 16%), *anticipation* (12 - 14%), and *anger* (10 - 11%). Besides pointing out how emotions like *curiosity* (2 - 6%), *admiration* (2 - 3%), *approval* (2 - 3%), and *confusion* (2 - 3%) are conveyed in the posts, RobertaGo (Figure 2c) ranks *sadness* (6 - 11%) as the most frequently relayed emotion throughout the phases. T5 (Figure 2b), in contrast to the two models, demonstrates that for all treatment elements, *joy* (28 - 36%) is the most prominent emotion, followed by *sadness* (24 - 37%), and *anger* (16 - 21%). Almost all emotion classifiers agree that *fear* and *sadness* are the most prevalent emotions. Every model also highlights *joy* and *happiness* as frequently seen emotions in the posts (see Table 2). Among the top five most expressed emotions, RobertaGo and T5 identify *joy* as one of the major feelings reported in the narratives. Apart from *joy*, RobertaGo mentions *gratitude*, *approval*, *caring*, and *admiration* as recurring emotions over the stages. T5 also reports *love* as an uncommon emotion encountered mostly during diagnosis and palliative care. NRCLex identifies *trust* and *anticipation* as significant, along with *fear*, *sadness*, and

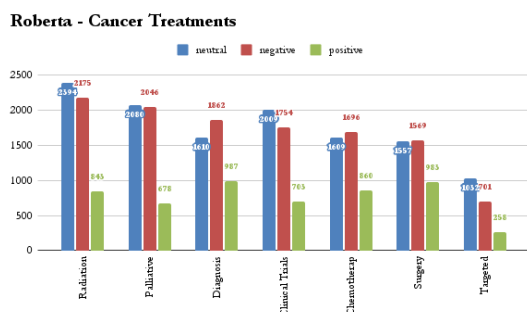
Figure 2: Cancer treatment-specific sentence-level sentiment classification.



(a) Sentiment classification using NRCLex. Sentiment categories include positive and negative.



(b) Sentiment classification using DistilBERT. Sentiment categories include positive and negative.



(c) Sentiment classification using Roberta. Sentiment categories include negative, positive, and neutral.

anger.

## 4.2. Cancer Stages

**Sentiment Analysis:** Negative sentiment is most commonly communicated across all cancer stages, particularly Stage IV cancer, as determined by all the classifiers. According to the DistilBERT (Figure 5b), negative sentiment is expressed prominently across all cancer stages, Stage I (62.7%), Stage II (60.9%), Stage III (64.7%), and Stage IV (64.6%) where there exists a substantial difference between the sentiment categories. NRCLex (Figure 5a), on the other hand, exhibits a small gap between

the sentiment categories throughout all stages and relays more positive attitudes for Stages I (52.2%) and III (53.3%) as compared to Stages II (46.3%) and IV (48.1%). At Stages III (37.7%) and IV (41.0%), Roberta (Figure 5c), expresses more negative sentiment, while at Stages I and II, it showcases a more neutral attitude.

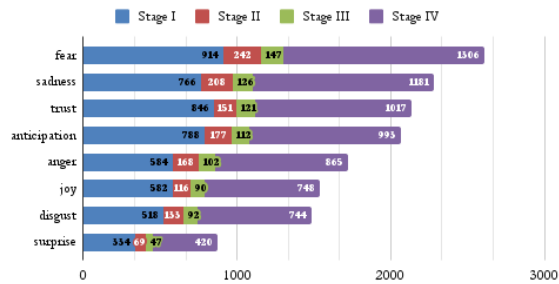
**Emotion Analysis:** The two emotions that appear most consistently throughout all cancer stages, according to the NRCLex model (Figure 4a) are *fear* (17 - 19%) and *sadness* (14 - 16%). RobertaGo (Figure 4c) highlights *sadness* (15 - 24 %) as the most frequently relayed emotion. In addition to that, it also emphasizes emotions like *admiration* (8 - 10%), *optimism* (5 - 9%), and *joy* (5 - 6%) being conveyed. According to NRCLex and RobertaGo, sadness is mostly expressed in Stage IV cancer posts as compared to other stages. T5 (Figure 4b) contrasts the two models and reveals that for all cancer stages, *joy* (33 - 39%) is the emotion that is most displayed, followed by *fear* (23 - 29%). The most prevalent emotions extracted by all classifiers are *fear* and *sadness*. Every model also points out that the feelings *joy* and *happiness* can frequently be seen in posts (see Table 2). *Joy* is one of the top five most frequently reported emotions by patients and carers, according to RobertaGo and T5. Apart from *joy*, RobertaGo mentions *gratitude*, *approval*, *caring*, and *admiration* as persistent emotions throughout the stages. NRCLex identifies *trust* and *anticipation* as significant, along with *fear*, *sadness*, and *anger*.

## 5. Human Evaluation

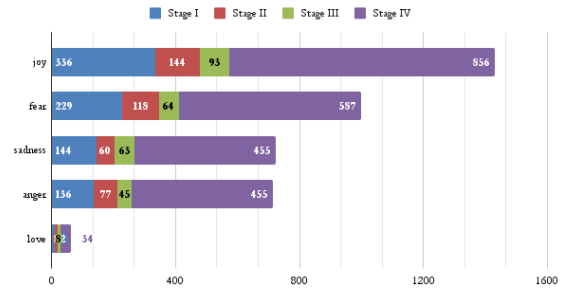
We reviewed a random subset of the collected posts having 50 sentences that were manually examined with the assistance of two distinct groups of annotators, three NLP/Linguistic researchers, and two medical domain specialists, to determine the impact of domain knowledge on the tasks for cancer-related data. The annotators evaluated the **Sentiment Polarity:** To determine whether the language conveys a positive, negative, or neutral attitude, and **Emotion Class:** To identify the emotion conveyed in the statement by using one of the following categories: sadness, anger, fear, joy, love, surprise, and neutral. Cohen's Kappa coefficient was used to determine the inter-rater agreement. Human evaluation was performed using three sentiment categories (positive, negative, and neutral) and seven emotion categories (sadness, anger, fear, joy, love, surprise, and neutral). The annotators were asked to choose one sentiment and emotion category they felt best suited in either scenario. Although some sentences contained

Figure 3: Cancer stage-specific sentence-level emotion classification.

NRCLex Emotions - Stages of Cancer



T5 - Stages of Cancer



(a) Emotion classification using NRCLex. Emotion classes include fear, sadness, trust, anger, disgust, joy, surprise, and anticipation.

(b) Emotion classification using T5. Emotion categories include joy, fear, anger, sadness, love, and surprise.

RobertaGo Emotions - Stages of Cancer

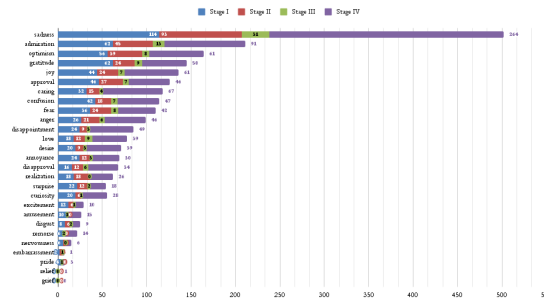
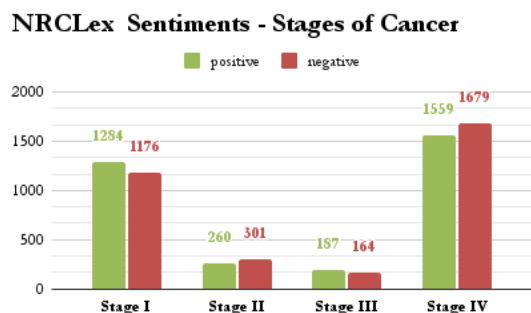
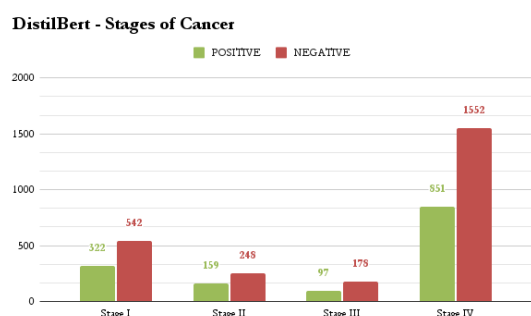


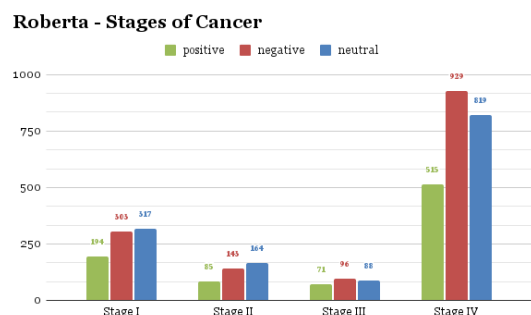
Figure 4: Cancer stage-specific sentence-level sentiment classification.



(a) Sentiment classification using NRCLex. Sentiment categories include positive and negative.



(b) Sentiment classification using DistilBert. Sentiment categories include positive and negative.



(c) Sentiment classification using Roberta. Sentiment categories include negative, positive, and neutral.

numerous emotion categories, the evaluation only permitted one choice for each category.

Corresponding to sentiment polarity (Figure 5), a substantial agreement (0.6383 - 0.6691) was discovered between the NLP/Linguistics researchers, and a substantial agreement (0.7534) exists among the domain experts. Moderate to substantial agreement (0.5064 - 0.6299) exists between the two groups indicating that scientific background and understanding somewhat influenced the sentiment polarity communicated in the sentences. Also, it was observed that the NLP/Linguistics researchers tended to classify

a slightly larger share of sentences as neutral compared to the domain experts. Corresponding to the emotion analysis (Figure 6b), a fair to moderate agreement (0.2946 - 0.5037) was found among the NLP/Linguistics researchers, while a substantial agreement (0.7258) was observed among the domain experts. Between the two groups, there is fair to moderate agreement (0.2273 - 0.5811), indicating that scientific background and understanding influence the emotions communicated in the sentences.

## 6. Comparative Analysis

To assess how well the SA models performed on data related to cancer, we compared the human and model-assigned sentiment tags using precision, recall, and accuracy (Table 4).

**Roberta:** The model achieves performance scores ranging from 0.6 to 0.8, indicating its ability to capture a significant portion of sentiment tags as assigned by annotators. With identical recall, precision, and accuracy scores of 0.7, it demonstrates balanced performance in identifying instances of the target emotion class while minimizing false identifications. Consistency in scores suggests that the model consistently captures the same instances as human-assigned labels, indicating comparable performance to human annotators. This implies effective training of the machine learning model, which remains accurate and reliable in assigning labels. A low trade-off between recall and precision indicates efficient identification of relevant instances while minimizing false positives, reflecting a well-tuned model. This reduces the need for manual verification or correction of labels, leading to significant time and cost savings in large-scale labeling tasks. Furthermore, the performance metrics indicate that Roberta surpasses both NRCLex and DistilBert on the SA task.

**NRCLex and DistilBert:** Both models achieve average recall and accuracy scores of 0.52, indicating a moderate performance. The precision scores are notably high, with NRCLex averaging 0.76 and DistilBert scoring 0.83. This suggests that the models tend to be conservative in their predictions, preferring to refrain from labeling instances as belonging to the emotion class unless they are highly confident in their prediction. However, the low recall indicates that a significant portion of instances belonging to the emotion class is being missed. This selective behavior, characterized by high precision and low recall, suggests that the models prioritize precision over recall, opting to make fewer predictions but ensuring their

Table 4: Precision, recall, and accuracy between human annotations and machine-generated sentiment tags for the sampled evaluation set. Here, **A**, **P**, **R** denote accuracy, precision, and recall, respectively. **a1**, **a2**, and **a3** denote experts in NLP/Linguistics, while a3 and a4 represent domain experts. **avg** denotes the average scores w.r.t all annotators.

	Roberta			NRCLex			DistilBert		
	A	P	R	A	P	R	A	P	R
<b>a1</b>	0.78	0.79	0.78	0.46	0.75	0.46	0.48	0.84	0.48
<b>a2</b>	0.74	0.73	0.74	0.62	0.83	0.62	0.58	0.85	0.58
<b>a3</b>	0.72	0.72	0.72	0.51	0.76	0.51	0.6	0.88	0.6
<b>a4</b>	0.70	0.71	0.70	0.54	0.78	0.54	0.5	0.83	0.5
<b>a5</b>	0.64	0.64	0.64	0.46	0.70	0.46	0.46	0.72	0.46
<b>avg</b>	<b>0.72</b>	<b>0.72</b>	<b>0.72</b>	<b>0.52</b>	<b>0.76</b>	<b>0.52</b>	<b>0.52</b>	<b>0.83</b>	<b>0.52</b>

correctness. Consequently, despite the moderate performance indicated by the accuracy score of 0.52, there remains substantial potential for enhancing the models' ability to capture more relevant instances.

All models exhibit relatively higher recall, precision, and accuracy when assessed by NLP/Linguistics researchers (a1, a2, a3) compared to domain experts (a4, a5). The low performance with domain expert annotators suggests that the model may not effectively capture the nuances or specific characteristics of the domain it was trained on. This could be due to limitations in the training data, inadequate representation of domain-specific features, or biases in the model architecture.

## 7. Linguistic and Semantic Challenges

Challenges that surfaced during the classification of cancer narratives are stated below. In all the examples, the color **red** represents a negative sentiment, **blue** represents a positive sentiment, and **green** represents a neutral or a sometimes ambiguous expression. The sentiment and emotion analysis is based on the tags provided by model predictions and human annotation.

### 1. Statements expressing a range of emotions:

When examining the posts, it appeared that people expressed a wide range of emotions in the same statement. For example, in the statement, "I have been told **I will never be cancer free**, I have **learnt to accept that**," the expression "never be cancer free" expresses a negative sentiment and a variety of emotions, sadness, fear, disappointment, etc. while "learnt to accept" displays a positive attitude and emotions like love, approval, optimism, admiration, etc. In another statement, "**I'm going to die**, and I'm going to **do it with as much dignity as possible** and have the **best last few months** I can possibly have," the patient expresses a negative

sentiment (though, given the circumstances, they may see it positively) in "going to die" while expressing a positive attitude in "do it with as much dignity as possible." The phrase "best last few months" conveys approval, adoration, optimism, sadness, and grief all at once. It is challenging to gauge the overall feeling relayed in such texts. Another example includes, "The hospital I live in right now had **given me less than 2 months**, and **I outlived it**." One of the major challenges for automatic sentiment or emotion classification is the ability to identify the overall attitude and pick the most likely emotion when the text is capable of multiple interpretations based on the context.

### 2. Statements with contradicting emotions:

When analysing the posts it was also observed that individuals often express their feelings using contradicting emotions (see Table 3). For example, in the statement, "This may **sound like hell**, but **it's actually pretty peaceful**," the patient expresses a negative attitude towards something comparable to an experience in hell when they use "may sound like hell," while contradicting the assumption and concluding that the experience is positive when they use "it's actually pretty peaceful". The statement exemplifies opposing feelings of disgust, fear, admiration, and optimism. This type of uncertainty is tough to capture and resolve not only for humans but also for machines. Furthermore, "I won't say that I hope my **long sleep comes soon**, but I **don't fear it**, it's **almost time for me to sleep forever**". In the preceding statement, the patient expresses a positive acceptance of a sad and undesirable circumstance. From the sentence "They **were able to cut out the tumor** but **weren't successful in getting clean margins** on the first pass," it is difficult to discern the sentiment because "able to cut out the tumor" indicates a positive sentiment or a sense of excitement, yet "weren't successful in getting clean margins" expresses a negative sentiment or sense of disappointment.

### 3. Statements that require additional context:

We discovered that sentiment and emotion analysis of cancer-related texts frequently necessitates prior knowledge and awareness of the healthcare domain, and that sentences cannot be successfully categorised without additional context. The sentence, "So I had my results today from my first Ct scan since treatment," is neutral in attitude and emotion because we do not know whether the results were favorable or not at this point. Some models selected joy as an acceptable emotion tag for this text, and two annotators suggested surprise. The second example, "I had an **aggressive cancer** in my left lung that **spread to my lymph nodes quickly**," lacks information regarding whether or not the situation was later improved. The NLP/Linguistic annotators identified a neutral sentiment and emotion for the sentence which might appear to be negative as identified by the domain experts who also chose fear and surprise as the appropriate emotion tag. Similarly the statement, "I'm having **really complicated feelings** about this" is ambiguous since we do not know what the individual is having "complicated feelings" about and, as a result, what kind of sentiment or emotion should be associated with it. It clearly portrays a sense of confusion, disapproval, disappointment, and nervousness, all at once.

**4. Statements where sentiment and emotions can be inferred but are not explicitly mentioned:** It has also been observed that, while emotions are not always explicitly mentioned in the sentence, they can be deduced using domain knowledge. For example, in the statement, "**Finished a year of treatment** and continued on maintenance chemotherapy for another year and **was cancer-free** for a while **until I relapsed at 22**," the expression "**until I relapsed at 22**" can be construed as conveying a negative sentiment, but lacks any explicit emotional words to indicate that the person is afraid, sad, surprised, disappointed, etc. In these cases, we might hypothesise that human annotators might annotate a sentence using an inferred emotion, and dictionary-based NLP approaches would be less able to capture such emotions, whereas DL-based methods might detect such subtle clues to annotate an implied or inferred emotion.

## 8. Conclusion

In this study, we conduct sentiment and emotion analysis of Reddit forum data on aspects (specifically stages and treatments) unique to cancer. We intend to analyse spontaneous clinical narratives to better understand the wide range of emotions that a patient or carer experiences throughout the

Figure 5: Human evaluation of 50 instances of cancer-related Reddit data. The table shows the kappa reliability scores between annotators. Here,  $a_i$  denotes the  $i^{th}$  annotator.

	a2	a3	a4	a5
a1	0.5468	0.6691	0.5166	0.6285
a2		0.5791	0.7534	0.6299
a3			0.5791	0.6383
a4				0.5064

(a) Sentiment polarity.

	a2	a3	a4	a5
a1	0.3181	0.2946	0.2273	0.3724
a2		0.5718	0.7258	0.537
a3			0.4878	0.5037
a4				0.5811

(b) Emotion classification.

various stages of cancer or treatments from diagnosis to palliative care. Through this study we discovered that: 1) Besides negative emotions (fear, anger, and sadness), there are many (potentially unexpected) examples of positive emotions (joy, happiness, admiration, approval, and optimism) in cancer-related posts. 2) Human evaluation results further indicates the dependency of both tasks on domain knowledge and comprehension. 3) The precision, recall, and accuracy scores suggest difficulties in accurately capturing the nuances of the target domain. Addressing these challenges may necessitate domain adaptation, careful examination of biases in the training data, and potentially utilizing transfer learning techniques to enhance model performance across all domains. 4) Various challenges encountered in annotation, both manual and automatic, include statements expressing diverse emotions, ambiguity or inconsistency in statements with conflicting emotions, and statements requiring additional context. We believe linguists can gain useful insights from this study when manually annotating such narratives. Additionally, we think that NLP researchers conducting comparable studies or developing new models would benefit from the analysis of the NLP models.



## 9. Ethics Statement

The large-scale analysis of sentiment and emotions expressed in open or closed online forums, particularly related to sensitive topics such as cancer requires ethical approval, and we have been granted approval for secondary data analysis of previously analysed datasets. The research presented in this paper is part of a larger multilingual multinational research project, and each partner will apply it in their organization or country to replicate our analysis. The overall aim of the research is to improve the cancer patient journey and ensure personal preferences are understood and respected during treatment discussions with medical professionals, thereby supporting treatment and care choices, at each stage of disease or treatment.

## 10. Acknowledgements

This publication presents research from the 4D PICTURE project,<sup>8</sup> which is a collaboration of research teams from Austria, Belgium, Denmark, Germany, the Netherlands, Slovenia, Spain, Sweden and the UK. The research leading to these results has received funding from EU research and innovation programme HORIZON Europe 2021 under grant agreement 101057332 and by the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number 10041120.

## 11. Bibliographical References

- Eunmi Ahn, Juhee Cho, Dong Wook Shin, Byeong Woo Park, Sei Hyun Ahn, Dong-Young Noh, Seok Jin Nam, Eun Sook Lee, and Young Ho Yun. 2009. Impact of breast cancer diagnosis and treatment on work-related life and factors affecting them. *Breast cancer research and treatment*, 116:609–616.
- Mohammed Al-Azri, Huda Al-Awisi, Samira Al-Rasbi, Kawther El-Shafie, Mustafa Al-Hinai, Hamdan Al-Habsi, and Mansour Al-Moundhri. 2014a. Psychosocial impact of breast cancer diagnosis among omani women. *Oman medical journal*, 29(6):437.
- Mohammed H Al-Azri, Huda Al-Awisi, Samira Al-Rasbi, and Mansour Al-Moundhri. 2014b. Coping with a diagnosis of breast cancer among omani women. *Journal of health psychology*, 19(7):836–846.
- Irfan Ayygün, Buket Kaya, and Mehmet Kaya. 2021. Aspect based twitter sentiment analysis on vaccination and vaccine types in covid-19 pandemic with deep learning. *IEEE Journal of Biomedical and Health Informatics*, 26(5):2360–2369.
- Jacqueline L Bender, M Carolina Jimenez-Marroquin, Lorraine E Ferris, Joel Katz, and Alejandro R Jadad. 2013. Online communities for breast cancer survivors: a review and analysis of their characteristics and levels of use. *Supportive Care in Cancer*, 21:1253–1263.
- Jacqueline L Bender, Maria-Carolina Jimenez-Marroquin, and Alejandro R Jadad. 2011. Seeking support on facebook: a content analysis of breast cancer groups. *Journal of medical Internet research*, 13(1):e1560.
- Meghna Chaudhary, Kristin Kosyluk, Sylvia Thomas, and Tempestt Neal. 2023. On the use of aspect-based sentiment analysis of twitter data to explore the experiences of african americans during covid-19. *Scientific Reports*, 13(1):10694.
- Juhee Cho, Katherine Smith, Eun-Kyung Choi, Im-Ryung Kim, Yoon-Jung Chang, Hyun-Young Park, Eliseo Guallar, and Young Mog Shim. 2013. Public attitudes toward cancer and cancer patients: a national survey in korea. *Psycho-Oncology*, 22(3):605–613.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Martí Domínguez and Lucía Sapiña. 2017. “others like me”. an approach to the use of the internet and social networks in adolescents and young adults diagnosed with cancer. *Journal of Cancer Education*, 32:885–891.
- Amy George, David Johnson, Giuseppe Carenini, Ali Eslami, Raymond Ng, and Elodie Portales-Casamar. 2021. Applications of aspect-based sentiment analysis on psychiatric clinical notes to study suicide in youth. *AMIA Summits on Translational Science Proceedings*, 2021:229.
- Felix Gräßer, Surya Kallumadi, Hagen Malberg, and Sebastian Zaunseder. 2018. Aspect-based sentiment analysis of drug reviews applying cross-domain and cross-data learning. In *Proceedings of the 2018 international conference on digital health*, pages 121–125.
- JN Harvey and VL Lawson. 2009. The importance of health belief models in determining self-care behaviour in diabetes. *Diabetic Medicine*, 26(1):5–13.

---

<sup>8</sup><https://4dpicture.eu/>

- Hyeju Jang, Emily Rempel, David Roth, Giuseppe Carenini, and Naveed Zafar Janjua. 2021. Tracking covid-19 discourse on twitter in north america: Infodemiology study using topic modeling and aspect-based sentiment analysis. *Journal of medical Internet research*, 23(2):e25431.
- Erin E Kent, Julia H Rowland, Laurel Northouse, Kristin Litzelman, Wen-Ying Sylvia Chou, Nonniekaye Shelburne, Catherine Timura, Ann O'Mara, and Karen Huss. 2016. Caring for caregivers and patients: research and clinical priorities for informal cancer caregiving. *Cancer*, 122(13):1987–1995.
- Amanda Lacy. 2016. Patient emotions and patient education technology. In *Emotions, Technology, and Health*, pages 23–39. Elsevier.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Sima Marzban, Marziye Najafi, Arjola Agolli, and Ensieh Ashrafi. 2022. Impact of patient engagement on healthcare quality: A scoping review. *Journal of Patient Experience*, 9:23743735221125439.
- Saif M Mohammad and Peter D Turney. 2013. Nrc emotion lexicon. *National Research Council, Canada*, 2:234.
- Linda J Muzzin, NJ Anderson, Alvaro T Figueredo, and Susan O Gudelis. 1994. The experience of cancer. *Social science & medicine*, 38(9):1201–1208.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. [SemEval-2017 task 4: Sentiment analysis in Twitter](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.
- H Ryan, Penelope Schofield, Jill Cockburn, Phyllis Butow, Martin Tattersall, Jane Turner, Afaf Girgis, D Bandaranayake, and Deborah Bowman. 2005. How to recognize and manage psychological distress in cancer patients. *European journal of cancer care*, 14(1):7–15.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. [CARER: Contextualized affect representations for emotion recognition](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.
- S Singer, J Das-Munshi, and E Brähler. 2010. Prevalence of mental health conditions in cancer patients in acute care—a meta-analysis. *Annals of oncology*, 21(5):925–930.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.
- Asmaa Hashem Sweidan, Nashwa El-Bendary, and Haytham Al-Feel. 2021. Sentence-level aspect-based sentiment analysis for classifying adverse drug reactions (adrs) using hybrid ontology-xlnet transfer learning. *IEEE Access*, 9:90828–90846.
- Lindsey A Torre, Freddie Bray, Rebecca L Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. 2015. Global cancer statistics, 2012. *CA: a cancer journal for clinicians*, 65(2):87–108.
- Anastazia Zunic, Pdraig Corcoran, and Irena Spasic. 2020. Sentiment analysis in health and well-being: systematic review. *JMIR medical informatics*, 8(1):e16023.

# Study of Medical Text Reading and Comprehension Through Eye-Tracking Fixations

Oksana Ivchenko, Natalia Grabar

Univ. Lille, CNRS, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France  
oksana.ivchenko.etu@univ-lille.fr, natalia.grabar@univ-lille.fr

## Abstract

Reading plays a crucial role in cognitive processes, acting as the primary way in which people access and assimilate information. However, the ability to effectively comprehend and understand text is significantly influenced by various factors related to people and text types. We propose to study the reading easiness and comprehension of texts through the eye-tracking technology, which tracks gaze and records eye movement during reading. We concentrate on the study of eye-tracking measures related to fixations (average duration of fixations and number of fixations). The experiments are performed on several types of texts (clinical cases, encyclopedia articles related to the medical area, general-language texts, and simplified clinical cases). Eye-tracking measures are analysed quantitatively and qualitatively to draw the reading patterns and analyse how the reading differs across the text types.

**Keywords:** Medical Texts, Types of Texts, Simplification, Reading, Eye-Tracking, Fixations

## 1. Introduction

Reading plays a crucial role in cognitive processes, acting as the primary way in which people access and assimilate information (Wylie et al., 2018; K. and Ismail, 2011). However, the ability to effectively comprehend and understand text is significantly influenced by various factors, including a person's level of education, proficiency in the language of the text, and general health (Aarsland et al., 2021; Javourey Drevet et al., 2022; Gala et al., 2018; Pandey et al., 2021). These elements can both facilitate and impede the reading process by impacting how information is processed and comprehended. Therefore, understanding the relationship between these factors and reading comprehension is essential for developing strategies to make information more comprehensible and accessible to a wider audiences.

In addition, text comprehension also varies considerably depending on the text type, and is influenced by key factors that interact with both reader capabilities and text properties (Pugh et al., 2023; Fulcher, 1997). Hence, in the medical area, several types of texts can be distinguished, such as scientific literature, drug inserts, patient leaflets, clinical documents, clinical trial protocols, and encyclopedia articles. In each case, the texts have specific structure, expected recipients, and content (Zweigenbaum et al., 2001; Biber and Finegan, 1994). The reading and comprehension of the different text types may vary due to these different reasons.

Eye-tracking technology provides possibility to study the reading easiness through specific measures. Researchers employ eye-tracking, a technique that tracks gaze patterns and records eye

movements while reading, in different contexts. Eye tracking provides objective measures of reading behaviour such as saccade duration, fixation size, pupil dilation and regression frequency (Rayner et al., 1989; Ekstrand et al., 2021; Clifton et al., 2007; Singh et al., 2016).

Longer fixation durations are often indicative of increased cognitive effort, as individuals spending more time at a specific point likely face challenges in interpreting the information presented (Just and Carpenter, 1980; Holmqvist et al., 2011; Ozeri-Rotstain et al., 2020). Thanks to this technique, the past decade has witnessed considerable advancements at the intersection of linguistics, cognitive science, and computer science, enhancing our understanding of cognitive processes, including text readability and comprehension. One of the known limitations of eye-tracking is that only short spans of texts can be processed (Duchowski, 2007). Hence, eye-tracking permits to collect precise reading measures but requires to split the texts into segments that fit the screen.

In our study, we aim to investigate the ease of reading and comprehension of different types of texts (medical and general-language). We employ eye-tracking methodologies, which enable the observation of gaze behaviour and the recording of eye movement data during the reading.

We hypothesise that eye-tracking measures can identify the complexity of texts, highlighting technical passages and terms, that are difficult to read and may require simplification. To investigate this hypothesis, we conducted an experiment where participants were exposed to four types of texts: original clinical texts (clinical cases), medical texts from Wikipedia, general-language texts, and simplified clinical cases. Our objective is to compare

the average fixation duration and the number of fixations across these text types to statistically discern their differences. As already indicated, this experiment provides objective measures about the complexity of texts and points out the content which should be simplified.

In what follows, we first introduce the eye-tracking technology and metrics in more detail (Section 2). We then describe the data used in the experiment (Section 3). Sections 4 and 5 are dedicated to the presentation of the principles for statistical analysis of eye-tracking measures related to fixation measures (the average fixation duration and the number of fixations), and the obtained results. Section 4 is focused on the analysis of each type of texts individually, while Section 5 provides an analysis across the types of texts making their comparison. A qualitative analysis of words that require most readers' attention is presented in Section 6. Finally, Section 7 is dedicated to the general conclusion and outlines some directions for future work.

## 2. Fixations in Eye-Tracking

Eye-tracking provides several objective measures collected during the reading process. Among them, the measures related to fixations occupy an important place.

Fixations, characterised by brief pauses during reading, are pivotal for information processing, serving as indicators of cognitive engagement and interaction with the text. Longer fixations often signal processing difficulty or heightened interest, while more frequent fixations may indicate the text's challenging nature or its ability to engage the reader. The average duration of an eye fixation on a word during reading varies depending on several factors, including the complexity of the text, the reader's familiarity with the content, and the purpose of reading (Hyönä and Kaakinen, 2019).

However, in general, research in the field of eye movement and reading suggests that the average eye fixation duration on a word is approximately 200-250 milliseconds (ms) for adults reading texts in their native language under normal conditions (Rayner and Reingold, 2015; Rayner et al., 2006). Hence, fixations tend to be longer for less common or more complex words, as the reader may require additional time to process the meaning of such words. Conversely, familiar or highly predictable words may receive shorter fixations, or even be skipped entirely, as the reader's brain can efficiently predict their meaning based on context.

Figure 1 illustrates a visual representation of eye-tracking data superimposed on a passage of a medical case text. Each circle represents a fixation point, where the number within the circle indicates the sequential order of fixations, and the size of the

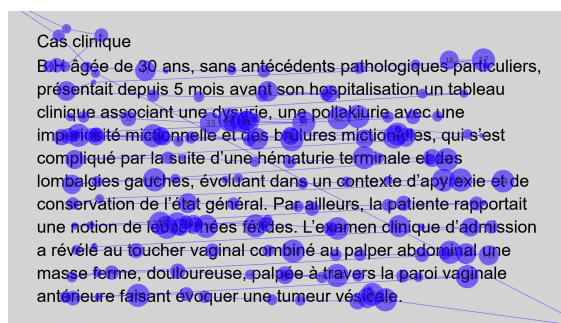


Figure 1: Example of fixations in a medical text. Translation of the excerpt from the clinical case: *Clinical case. B.H, a 30-year-old female with no particular pathological history, presented with symptoms for 5 months before her hospitalization, which included dysuria, frequent urination with urinary urgency, and burning sensations during urination. These symptoms were later complicated by terminal hematuria and left-sided lower back pain, progressing in a context of no fever and maintained general health. Additionally, the patient reported having foul-smelling leukorrhea. The admission clinical examination revealed through a combined vaginal touch and abdominal palpation a firm, painful mass palpated through the anterior vaginal wall, suggesting a bladder tumor.*

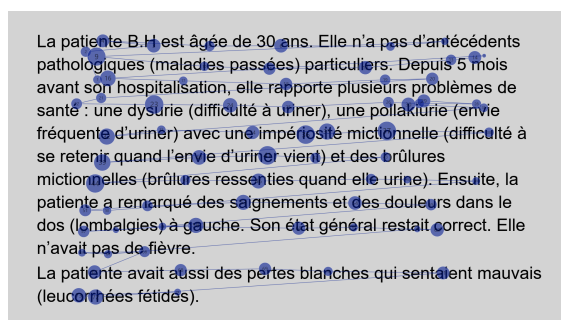


Figure 2: Example of fixations in a simplified clinical case. Translation of the excerpt from the clinical case: *Clinical Case. Patient B.H. is 30 years old. She has no particular past medical history. For five months before her hospitalization, she reported several health problems: dysuria (difficulty urinating), pollakiuria (frequent urge to urinate), urinary urgency (difficulty holding urine when the urge comes), and micturition burns (burning sensations when urinating). Subsequently, the patient noticed bleeding and pain in the left side of her back (left lumbar pain). Her overall health remained stable. She did not have a fever. The patient also experienced foul-smelling white discharge (fetid leukorrhea).*

circle corresponds to the duration of the fixation: the larger the circle, the longer the reader looked at

that point in the text. The lines connecting the circles show saccadic movements between fixations, demonstrating the trajectory of the reader's gaze as the reader progresses through the text. In contrast, Figure 2 displays the reading of a simplified version of the clinical case, showing a more uniform spread of colours with fewer fixations. This map of visual attention allows us to determine which parts of the text are subjected to deeper cognitive processing, as indicated by the number and size of fixations.

### 3. Data and Experiment

For our experiment, we utilise a medical corpus in French that includes the CLEAR corpus (Grabar and Cardon, 2018) and a corpus with Clinical Cases (Grabar et al., 2020). These corpora encompass a variety of materials, such as articles from Wikipedia, reviews, leaflets, and medical cases.

Clinical cases are detailed accounts of the symptoms, diagnosis, treatment and follow-up of an individual patient or a small group of patients. Their content is close to clinical documents, such as discharge summaries (Grabar et al., 2020). Hence, clinical cases are rich in medical terminology. Additionally, they may include a discussion of the rationale for treatment choices, making them complex and rich in specialised information. Such texts are invaluable for medical education and practice, providing insight into the practical application of theoretical knowledge in real-life scenarios. In the context of our study, clinical case texts serve as a key component of the corpus, providing a deep immersion into medical scenarios that require significant cognitive effort of patients and their families to process and comprehend medical information relevant to their health and care.

Overall, for the eye-tracking experiments, we compiled a corpus of 16 texts. The texts are in French. This collection was thoughtfully divided into two distinct sets to ensure a balanced representation of text types in each.

Specifically, Set 1 comprised one assortment of texts, while Set 2 featured a different assortment, with each set containing an equal mix of medical texts, clinical cases, and ordinary texts to maintain a uniform distribution of text complexity and subject matter across both sets. We manually simplified these texts, resulting in two variations for each text within the sets: Version A and Version B. In Version A, some texts were presented in their original form, while others were simplified. Version B reversed this configuration, providing a mirrored counterpart to Version A in terms of which texts were simplified. Participants in the study were assigned to read texts from either Set 1 or Set 2, but not both, to ensure focused exposure to a specific subset of texts.

For the purpose of the work presented here, we chose 8 texts with the aim to cover the variety of the available medical text types (Table 1 indicates the size of these texts):

- two clinical case texts, chosen for their complexity and their specificity to the clinical context. Indeed, the patients and their families often have to face such documents during their healthcare process;
- two Wikipedia articles related to medical topics: autopsy and erythema. These texts illustrate medical information freely accessible to the general public. As observed by researchers and associations, general population look for medical and health information online increasingly frequently (Fox, 2014);
- two general-language texts about popcorn and quince. They are provided from Wikipedia as well. They are selected to illustrate general-language topics and provide some contrast in terms of content and lexical density by comparison with medical texts;
- the simplified versions of clinical cases. Since the content of clinical cases is too technical for common people, we manually simplified these clinical cases to make their reading more friendly. The simplification was done at lexical (lexical substitutions with synonyms, hyperonyms, hyponyms, definitions...), syntactic (sentence structure modification) and semantic (addition of contextual and semantic information) levels.

Table 1: Text size

Text Category	Nb words
Clinical Cases	534
Medical Encyclopedia Texts	1,594
General Encyclopedia Texts	1,545
Simplified Clinical Cases	630

As indicated, we created two sets of texts and, in each set, there is a version A and a version B :

- version A contains text 1 in its original form, text 2 in its simplified form, text 3 in its original form, text 4 in its simplified form,
- version B contains text 1 in its simplified form, text 2 in its original form, text 3 in its simplified form, text 4 in its original form.

Hence, each person reads all the texts from a given set in either original or simplified forms. Yet, in this study, we analyze only the simplified versions of clinical cases. As indicated above, due to the

screen size limitations (Duchowski, 2007), the texts are divided into smaller segments.

The main purpose of this experiment is to collect eye-tracking indications on text reading. In addition, we also collect information on text understanding. For this, comprehension questions are asked after the reading of a given segment. The questions are related to the segment the participants just read. The possible answers to these questions are: *True*, *False*, or *I don't know*. In order to make the reading as natural as possible, the questions are asked only on some segments of text.

Participants read the texts using a Tobii Pro Spectrum eye-tracker, operating at 600 Hz.

For the experiment presented in this paper, we analyse the results from two groups of 5 participants each, totaling 10 participants. These individuals are French-speaking with French as first language, aged between 19 and 33 years, with no medical education, and coming from various social backgrounds, including students, PhD students, and full-time employees. Each group is tasked with reading texts of different types (clinical case, medical text, general text, and a simplified medical case), and sometimes answering the understanding questions.

In this preliminary analysis, we selectively focused on eye movement data from ten participants and specific text types to investigate characteristic patterns of eye movements during reading. The responses to comprehension questions from this subset of participants were not considered in the current analysis. This decision was made because the primary goal of this phase was to examine eye movement performance, and the limited sample size precludes a comprehensive analysis of text comprehension across the entire participant group based on their responses to questions. At this stage, we considered the comprehension questions mainly as active engagement with the text, thus ensuring that the recorded eye movements accurately reflect actual reading behaviour. It is important to note that this approach does not negate the value of comprehension data. As we move beyond this preliminary phase, we intend to conduct a more extensive analysis that includes eye movement characteristics in conjunction with comprehension responses for all texts and participants. This future analysis aims to offer deeper insights into how text comprehension correlates with specific eye movement patterns.

In the two following sections, we present the analysis principles and the results first for each type of texts individually (Section 4) and then across the types of texts making their comparison (Section 5). In addition, Section 6 concentrates on a qualitative analysis of words that require most readers' attention.

## 4. Statistical Analysis of Fixation Measures

To describe the fixation measures in different text types, and to infer the cognitive effort required to process the text content, we analyse the fixation metrics for four types of texts (clinical case, medical text, general text, and simplified clinical case). For each type of texts, the results are presented and discussed across three lines: general statistical analysis of fixation measures (Section 4); normality test (Section 4.2) to assess the normality of the data distribution and to define which further statistical tests can be applied; correlation of fixation measures (Section 4.3).

### 4.1. Collected Values of Fixation Measures

The collected average measures of the fixations are summarised in two tables: Table 2 details the average total duration of fixations, while Table 3 presents the average number of fixations for each text type. We indicate information on Mean values, the Standard deviation and Median values.

Table 2: Average Total Duration of Fixations Across Text Types in ms

Text Category	Mean	SD	Median
Clinical Case	395.89	328.25	307.8
Medical Text	359.29	239.37	321.8
General Text	323.78	239.95	271.9
S. Clinical Case	255.75	197.50	223.25

Table 3: Average Number of Fixations per Word

Text Category	Mean	SD	Median
Clinical Case	1.83	1.33	1.5
Medical Text	1.62	0.97	1.4
General Text	1.54	0.99	1.4
S. Clinical Case	1.27	0.87	1.2

The average total duration of fixations (Table 2) was the highest for clinical case texts, indicating these require more time to process, likely due to their complexity. In contrast, simplified clinical case texts showed the shortest average duration, suggesting that simplification effectively reduces cognitive load. Similarly, the average number of fixations (Table 3) followed a comparable pattern (clinical cases show the highest number of fixations indicating that this type of texts require more attention), further supporting the notion that text complexity influences reading behaviour.

The standard deviation for both fixation duration and number was notably higher in clinical case texts, underscoring a variability in complexity within

this category. Indeed, clinical cases typically contain technical terms (*dysuria, hematuria, leukorrhea...*) aside more common words with medical meaning (*pathological, symptoms, hospitalization, urgency, burning...*) or not (*female, sensation, frequent, combined...*). This variability was less pronounced in the simplified texts, indicating a more consistent level of difficulty.

The two other types of text (medical and general language encyclopedia articles) show intermediate values. Yet, it is worth to notice that general language texts show high SD values.

## 4.2. Normality Test

We apply the normality Shapiro-Wilk test (Shapiro and Wilk, 1965) to assess the normality of the data distribution. Understanding whether our data conform to a normal distribution is necessary to determine the most appropriate further statistical tests.

After conducting the Shapiro-Wilk test, the data for both number and duration of fixations across all text categories yielded p-values very close to 0, decisively indicating their non-normal distributions. This result is consistent across the diverse text types examined (clinical case, encyclopedia medical text, general text, and simplified clinical case) highlighting significant variability in fixation metrics that could be attributed to differences in syntactic and lexical density. The anticipation of such variability, given the distinct characteristics of each text type, underscores the complexity of the cognitive processes involved in text reading and comprehension. We included the mean measure in our analysis to provide a central tendency of eye-tracking measures, revealing how text complexity affects reader engagement. Despite the non-normal distribution of the data, the mean values illustrate the general reading behaviour across different text types (Tables 2 and 3), highlighting longer and more frequent fixations on complex texts, such as clinical cases. The median offers a more accurate reflection of central tendency than the mean, as it is less influenced by extreme values. Therefore, focusing on the median provides a clearer understanding of the typical reader engagement and comprehension levels across different text types.

## 4.3. Correlation of Fixation Measures

After establishing that the fixation data, both duration and number, do not follow a normal distribution across the various text types, we next explore the relationship between these two metrics. Understanding the correlation between average total duration of fixations and average number of fixations can offer deeper insights into how text complexity influences reading behaviour. Given the non-normal distribution of our data, we employ Spearman's

rank correlation coefficient (Spearman, 1904; Zar, 2005), a non-parametric measure that assesses the strength and direction of association between two ranked variables. This test is particularly suitable for our dataset given its observed non-normality.

The purpose is to verify whether a higher number of fixations correlates with longer total durations, suggesting more cognitive effort or processing time, in different types of texts. Such an analysis is crucial for understanding the nuances of reading patterns and how textual characteristics impact reader engagement and comprehension. Hence, employing this test, we analyse the correlation between average total duration of fixations and average number of fixations within each of the four text types.

Table 4: Spearman's Correlation Coefficients and P-values for Different Text Types

Text Type	Correl. Coeff.	P-value
Clinical Case	0.971	< 0.001
Medical Text	0.903	< 0.001
General Text	0.955	< 0.001
S. Clinical Case	0.968	< 0.001

Table 4 presents the Spearman's correlation coefficients and p-values for different text types, indicating a consistently strong positive correlation between the average total duration of fixations and the average number of fixations. These results suggest a robust relationship across all text types: increased fixation duration is associated with a higher number of fixations, reflecting varying levels of text complexity and cognitive engagement of readers.

## 5. Statistical Comparison between the Types of Texts

To determine whether significant differences exist across the text types in terms of fixation duration and frequency, we apply the Kruskal-Wallis test (Kruskal and Wallis, 1952). The analysis for average total duration of fixations yielded a test statistic of 85.137, with a corresponding p-value near 0.001. This indicates significant differences between the text types. Similarly, for average number of fixations, the test statistic was 55.191, with a p-value near 0.001, further confirming significant disparities between the text types.

Hence, we further apply the Dunn's post-hoc test (Dunn, 1961) to specify which text types differ significantly in fixation duration and frequency. This analysis facilitates pairwise comparisons between text types, elucidating the specific nature of the differences indicated by the Kruskal-Wallis test. We first present and discuss the findings for the average duration of fixations (Section 5.1) and then for the average number of fixations (Section 5.2).

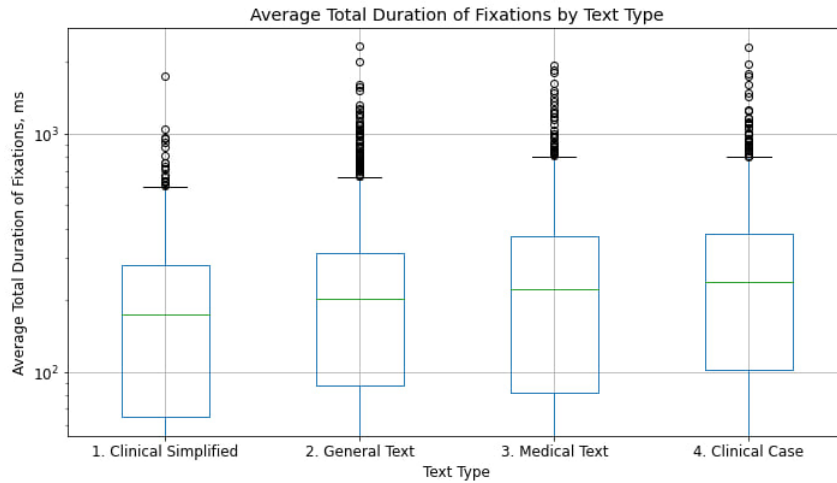


Figure 3: Boxplot of Average Total Duration of Fixations Across Text Types.

### 5.1. Average Total Duration of Fixations

Table 5: Post-hoc Dunn’s Test Results for Average Total Duration of Fixations. The following text types are indicated: CC (clinical case), SCC (simplified clinical case), MT (medical text), and GT (general text).

Text Pairs for Comparison	P-value
CC vs. SCC	$2.61 \times 10^{-11}$
CC vs. GT	0.009975
CC vs. MT	1.000
SCC vs. GT	$1.84 \times 10^{-6}$
SCC vs. MT	$2.49 \times 10^{-17}$
GT vs. MT	0.000035

As demonstrated in Table 5, the Post-hoc Dunn’s test reveals a statistically significant variation in the average total duration of fixations across different types of texts. Specifically, a Post-hoc Dunn’s test highlights a notable difference between clinical case texts and their simplified versions ( $p = 2.61 \times 10^{-11}$ ), indicating that text simplification leads to a measurable reduction in cognitive load. This is further supported by comparisons between simplified clinical case texts against general ( $p = 1.84 \times 10^{-6}$ ) and medical texts ( $p = 2.49 \times 10^{-17}$ ), suggesting that simplified texts are read and comprehended more efficiently by readers.

Interestingly, no significant difference is found between clinical case and medical texts ( $p = 1.0$ ), which suggests a comparable level of complexity from a cognitive load perspective. This observation is crucial for understanding the nuances of text engagement and comprehension, reinforcing the importance of text types for cognitive processing.

In Figure 3 the plot illustrates the distribution of average fixation durations across four categories of

texts. The central line in each box represents the median duration, while the top and bottom edges of the box delineate the third and first quartiles, respectively. We can observe that the texts present an increasing difficulty starting with simplified clinical case, going through general-language text, then up to medical text and clinical case. From this boxplot, it is evident that the simplified clinical text has the lowest range of fixation durations as measured by eye-tracking characteristics, followed by general text. Medical text exhibits higher fixation durations, and clinical case texts have the highest, indicating a trend of increasing fixation duration across the complexity of the text types.

### 5.2. Average Number of Fixations

Table 6: Post-hoc Dunn’s Test Results for Average Number of Fixations. The following text types are indicated: CC (clinical case), SCC (simplified clinical case), MT (medical text), and GT (general text).

Text Pairs for Comparison	P-value
CC vs. SCC	$3.99 \times 10^{-10}$
CC vs. GT	0.01323
CC vs. MT	1.000
SCC vs. GT	$1.38 \times 10^{-5}$
SCC vs. MT	$2.85 \times 10^{-10}$
GT vs. MT	0.112714

Similarly, the average number of fixations across text types indicates significant disparities, reinforcing the insights gained from the analysis of fixation durations. The Post-hoc Dunn’s test results (Table 6) demonstrate a significant reduction in fixations when comparing clinical case texts with their simplified counterparts ( $p = 3.99 \times 10^{-10}$ ), underlining



the effectiveness of simplification in enhancing text accessibility.

Furthermore, while general and medical texts exhibit no significant difference in the number of fixations ( $p = 0.112714$ ), the distinct contrast with simplified clinical case texts emphasises the impact of simplification on reader engagement and cognitive effort. Taken together, these results highlight the potential of targeted text simplification to improve their comprehension and accessibility, especially for non-expert readers who are exposed to complex medical content.

## 6. Analysis of Words that Require Most Readers' Attention

We propose an analysis of words with the longest fixation durations and the highest number of fixations across the four types of texts (Tables 7 to 10). For 8 to 10 top-words, we indicate the average duration of fixations and the average number of fixations.

Table 7: Top Words by Average Total Duration (ms) and Number of Fixations in Clinical Case Texts

Words	Avg. Dur.	Avg. Nb
urétéro-hydronéphrose	2322.4	9.6
immunohistochimique	1954.2	9.2
impériosité	1780.8	6.0
vésicoacétabulaire	1748.2	7.2
pollakiurie	1617.6	5.6
leucorrhées	1486.0	5.4
latéro-trégonale	1428.0	6.8
cystoscopie	1267.2	5.4
extrapéritonisation	1242.0	5.0
47/48/52 mm	1164.4	5.0

Table 8: Top Words by Average Total Duration (ms) and Number of Fixations in Medical Texts

Words	Avg. Dur.	Avg. Nb
télangiectasie	1936.8	7.6
n°2011-525	1852.4	6.8
d'anatomo-pathologie	1820.0	7.4
«médico-hospitalière»	1637.8	6.4
anatomo-pathologiques	1512.8	6.6
1335-11	1487.0	4.6
ataxie-télangiectasie,	1423.8	6.4
spécialistes	1376.4	6.0
polypathologies	1316.4	4.6
scarlatiniformes	1256.0	5.2

In Tables 7 and 8, we can observe that words with the longest average total duration of fixations in medical texts are predominantly medical terms (*immunohistochimique*, *vésicoacétabulaire*, *ataxie-télangiectasie*...). Moreover, there is a noticeable

overlap between words with the highest number of fixations and those with the longest fixation durations, indicating again a correlation between the complexity of medical terms and the cognitive effort required for their processing.

Table 9: Top Words by Average Total Duration in ms and Number of Fixations in General Texts

Words	Avg. Dur.	Avg. Nb
2-méthyle-2-butenoate	2337.6	9.8
proanthocyanidols	2017.8	8.6
anthocyanidols	1616.4	6.2
(sclérenchymateuses)	1576.0	5.8
Aromatnaya	1515.6	6.2
Balkans	1321.2	5.6
fabrique	1258.4	3.6
cuisson	1212.8	4.6

Table 9 demonstrates similar information for general-language text. Here, complex chemical names, such as *2-méthyle-2-butenoate* and *proanthocyanidols* (chemical substances contained in quince) attract the gaze of readers and show the highest fixation durations and frequencies when reading the general text. This suggests that these terms pose a cognitive challenge to readers, despite being part of general texts. They may also be the cause of the high standard deviation in this type of texts, where they neighbour common and frequent words (Section 4.1). Besides, the fixation data reveal that even non-medical, general texts may contain words that necessitate significant cognitive effort to be processed.

Table 10: Top Words by Average Total Duration (ms) and Number of Fixations in Simplified Clinical Case Texts

Words	Avg. Dur.	Avg. Nb
vésicoacétabulaire	1745.00	6.20
Staphylococcus	1052.40	4.80
(cystoscopie)	963.40	4.60
pollakiurie	956.25	3.5
(tomodensitométrie)	916.50	4.50
cystoscopie)	883.00	4.75
(intraveineuse)	810.00	3.60
d'orthopédie	759.40	3.80
(aminoside)	758.80	2.8
dysurie	727.75	3.50

Finally, Table 10 shows words from the simplified clinical cases, highlighting that medical terms, often enclosed in brackets and explained with simpler language, significantly reduce both the duration of fixations and the number of fixations per word when compared to the fixation values observed on original clinical case texts. This reduction indicates that simplification efforts effectively lower the cognitive load required to comprehend these medical

terms and that the comprehension is helped by the context.

These findings collectively underscore the influence of text types on reader engagement and cognitive processing. Indeed, the proposed qualitative and quantitative analysis of eye-tracking measures indicates the complexity elements and zones within the analysed texts.

## 7. Conclusion and Future Work

In this work, we propose an experiment using eye-tracking technology with the purpose of revealing text complexity and the inherent cognitive load it presents to readers. Four text types are considered: clinical cases, encyclopedia medical texts, encyclopedia general texts, and manually simplified clinical cases. These texts are read by 10 people. The collected eye-tracking measures related to fixations are analysed. We perform quantitative and qualitative analysis. The normality statistical test shows that the fixation values do not have a normal distribution, which means that the reading difficulty is uneven. This also suggests the complexity of cognitive processed involved in text reading. Further quantitative analysis with different statistical tests indicates that the text types significantly impact the reading easiness, as testified by the fixation measures (duration of fixations and their number): in our experiment, (1) clinical cases are the most difficult to read, (2) they are followed by medical and general language encyclopedia articles, (3) while the simplified version of clinical cases eases a lot the reading process. The statistical analysis also indicates that duration of fixations and their number are correlated: complex words usually require longer fixations and their number is higher. Such words need a stringer cognitive effort. Interestingly, this quantitative analysis indicates that clinical cases present the highest difficulty, yet the simplification of clinical cases makes these texts much easier to read and comprehend. As for the qualitative analysis, we presented top-words which require the most attention from readers in each type of texts. These words usually correspond to technical medical terms.

These measures from eye-tracking records across various text types can be utilised to detect the complexity zones within these texts. Besides, such an eye-tracking annotation of texts can be used to train a language model, thus enabling the automatic prediction of reading patterns for texts of different types. This is one of our objectives for future. Other objectives are related to the collection of eye-tracking measures from more people and on more texts.

## 8. Ethical Considerations and Limitations

Participation in this study is voluntary, with informed consent obtained from all participants, ensuring compliance with the European General Data Protection Regulation (EU) 2016/679 and the modified French Data Protection Act of January 6, 1978. All personal data collected in the course of this research are anonymized to protect participant privacy and are accessible only by the designated project manager. This study has been registered in the University of Lille's registry under reference 2022-075, affirming our commitment to upholding the highest standards of data protection and participant rights.

## 9. Acknowledgements

This work was partly funded by the French National Agency for Research (ANR) as part of the *CLEAR* project (*Communication, Literacy, Education, Accessibility, Readability*), ANR-17-CE19-0016-01.

We thank the FR "Sciences et Cultures du visuel" for the facility to perform the eye-tracking experiments.

We also would like to thank the anonymous reviewers for their helpful comments.

## 10. Bibliographical References

- Dag Aarsland, Lucia Batzu, Glenda Halliday, Gert Geurtsen, Clive Ballard, Kallol Ray Chaudhuri, and Daniel Weintraub. 2021. [Parkinson disease-associated cognitive impairment](#). *Nature Reviews Disease Primers*, 7:47.
- Douglas Biber and E. Finegan. 1994. Intra-textual variation within medical research articles. *Corpus-based research into language*, 12:201–222.
- Charles Clifton, Adrian Staub, and Keith Rayner. 2007. [Eye movements in reading words and sentences](#). *Eye movements: A window on mind and brain*.
- Andrew Duchowski. 2007. *Eye Tracking Methodology. Theory and practice*. Springer, London, UK.
- Anna Ekstrand, Mattias Nilsson, and Gustaf Öqvist Seimyr. 2021. [Screening for reading difficulties: Comparing eye tracking outcomes to neuropsychological assessments](#). *Frontiers in Education*, 6.
- Susannah Fox. 2014. The social life of health information. Technical report, Pew Internet & American Life Project, Washington DC.

- Glenn Fulcher. 1997. [Text difficulty and accessibility: Reading formulae and expert judgement](#). *System*, 25:497–513.
- Núria Gala, Thomas François, Ludivine Javourey-Drevet, and Johannes C Ziegler. 2018. [La simplification de textes, une aide à l'apprentissage de la lecture](#). *Langue française*, 199(3):123–131.
- Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. 2011. [Eye tracking: A comprehensive guide to methods and measures](#).
- Jukka Hyönä and Johanna K. Kaakinen. 2019. [Eye Movements During Reading](#), pages 239–274. Springer International Publishing, Cham.
- Ludivine Javourey Drevet, Stephane Dufau, Thomas François, Núria Gala, Jacques Ginestié, and Johannes Ziegler. 2022. [Simplification of literary and scientific texts to improve reading fluency and comprehension in beginning readers of french](#). *Applied Psycholinguistics*, 43:1–28.
- Marcel Adam Just and Patricia A. Carpenter. 1980. [A theory of reading: from eye fixations to comprehension](#). *Psychological review*, 87 4:329–54.
- Ooi K. and Khaidzir Ismail. 2011. [The relationship between cognitive processing and reading](#). *Asian Social Science*, 7.
- William H. Kruskal and W. Allen Wallis. 1952. [Use of ranks in one-criterion variance analysis](#). *Journal of the American Statistical Association*, 47(260):583–621.
- Aya Ozeri-Rotstain, Ifaat Shachaf, Rola Farah, and Tzipi Horowitz-Kraus. 2020. [Relationship between eye-movement patterns, cognitive load, and reading ability in children with reading difficulties](#). *Journal of Psycholinguistic Research*, 49.
- Mamata Pandey, Geoffrey Maina, Jonathan Amoyaw, Yiyang Li, Rejina Kamrul, C. Michaels, and Razawa Maroof. 2021. [Impacts of english language proficiency on healthcare access, use, and outcomes among immigrants: a qualitative study](#). *BMC Health Services Research*, 21.
- Alia Pugh, Devin Kearns, and Elfrieda Hiebert. 2023. [Text types and their relation to efficacy in beginning reading interventions](#). *Reading Research Quarterly*, 58:710–732.
- Keith Rayner, Kathryn Chace, Timothy Slattery, and Jane Ashby. 2006. [Eye movements as reflections of comprehension processes in reading](#). *Scientific Studies of Reading - SCI STUD READ*, 10.
- Keith Rayner and Eyal Reingold. 2015. [Evidence for direct cognitive control of fixation durations during reading](#). *Current Opinion in Behavioral Sciences*, 1:107–112.
- Keith Rayner, Sara Sereno, Robin Morris, Anna Schmauder, and Charles Clifton. 1989. [Eye movements and on-line language comprehension processes](#). *Language and Cognitive Processes - LANG COGNITIVE PROCESS*, 4.
- Abhinav Singh, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan. 2016. [Quantifying sentence complexity based on eye-tracking measures](#).
- Judith Wylie, Jennifer Thomson, Paavo Leppanen, Rakefet Ackerman, Laura Kannianen, and Tanja Prieler. 2018. [Chapter 3. Cognitive processes and digital reading](#), pages 57–90.
- Pierre Zweigenbaum, Pierre Jacquemart, Natalia Grabar, and Benoit Habert. 2001. [Building a text corpus for representing the variety of medical language](#). In *MEDINFO*, pages 290–294.

## 11. Language Resource References

- Olive Jean Dunn. 1961. [Multiple comparisons among means](#). *Journal of the American Statistical Association*, 56(293):52–64.
- N. Grabar, Clément Dalloux, and V. Claveau. 2020. [Cas: corpus of clinical cases in french](#). *Journal of Biomedical Semantics*, 11.
- Natalia Grabar and Rémi Cardon. 2018. [CLEAR – simple corpus for medical French](#). In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.
- S. S. Shapiro and M. B. Wilk. 1965. [An analysis of variance test for normality \(complete samples\)](#). *Biometrika*, 52(3-4):591–611.
- C Spearman. 1904. [The proof and measurement of association between two things](#). *The American Journal of Psychology*, 15(1):72–101.
- Jerrold H Zar. 2005. [Spearman rank correlation](#). *Encyclopedia of Biostatistics*, 7.

# A Neuro-Symbolic Approach to Monitoring Salt Content in Food

Anuja Tayal<sup>1</sup>, Barbara Di Eugenio<sup>1</sup>, Devika Salunke<sup>2</sup>, Andrew D. Boyd<sup>2</sup>,  
Carolyn A Dickens<sup>3</sup>, Eulalia P Abril<sup>4</sup>, Olga Garcia-Bedoya<sup>3</sup>, Paula G Allen-Meares<sup>3</sup>

<sup>1</sup> Department of Computer Science <sup>2</sup> Department of Biomedical and Health Information Sciences

<sup>3</sup> Department of Medicine <sup>4</sup> Department of Communications

University of Illinois Chicago, IL, USA

{atayal4, bdieugen, dsalun2, boyda, cdickens, eulalia, ogarciab, pameares}@uic.edu

## Abstract

We propose a dialogue system that enables heart failure patients to inquire about salt content in foods and help them monitor and reduce salt intake. Addressing the lack of specific datasets for food-based salt content inquiries, we develop a template-based conversational dataset. The dataset is structured to ask clarification questions to identify food items and their salt content. Our findings indicate that while fine-tuning transformer-based models on the dataset yields limited performance, the integration of Neuro-Symbolic Rules significantly enhances the system's performance. Our experiments show that by integrating neuro-symbolic rules, our system achieves an improvement in joint goal accuracy of over 20% across different data sizes compared to naively fine-tuning transformer-based models.

**Keywords:** Dialogue Systems, Neuro-Symbolic AI, Heart Failure

## 1. Introduction

The excessive consumption of salt poses significant public health risks, contributing to diseases such as high blood pressure and heart failure (He et al., 2020). Reducing salt intake has been shown to mitigate these health issues. In 2017, excessive sodium intake was associated with around three million deaths and a significant loss of healthy life years (Roth et al., 2018). Research, including clinical trials and population studies, supports the reduction of salt intake as a means to manage and prevent these conditions. Despite the clear benefits of sodium reduction, public understanding and action are lacking; only 58% of individuals can accurately read sodium content on nutrition labels, and merely 44% can classify food products as high or low in sodium based on standard labeling (Dickson and Riegel, 2009). This gap in knowledge and practice underscores the challenge of addressing dietary sodium intake, with only a handful of countries implementing effective public health interventions.

Therefore, we aim to develop a dialogue system that enables patients to inquire about the salt content in various foods. This system especially aims to support heart failure patients, who must meticulously monitor and reduce their salt intake. More specifically, African American individuals who are more prone to heart failure (Nayak et al., 2020), have a higher sensitivity to salt and face challenges like food deserts and higher consumption of junk foods. This necessitates a specialized dietary management approach to help them monitor and reduce their salt intake effectively. Furthermore, in (Gupta et al., 2020), the authors show that African American patients with heart failure often focus on discussions related to salt and food during heart

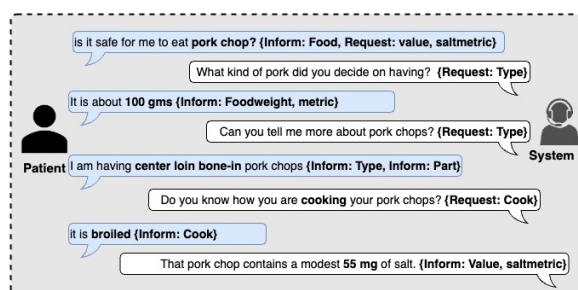


Figure 1: Sample Template Conversation which begins with the user asking about the salt content in food. The system asks clarification questions to determine the precise food item and its salt value.

failure educational sessions, indicating a significant interest and need for information in this area. By providing a tool that facilitates easy access to information about salt content in foods, we aim to empower patients to make healthier dietary choices, thereby addressing a critical aspect of managing heart failure. Having said that, the dialog system can be used by anyone who wants to inquire about the salt content in food.

Creating a dialog agent specialized in food-related conversations and nutrient information is challenging. This is primarily due to the lack of a conversational dataset specifically designed for this domain. Moreover, annotating the dataset is very costly and resource-intensive. To address this issue, we create a template-based conversational dataset (an example is shown in Figure 1) to identify various food items and their salt content. Our approach involves utilizing the USFDC (U.S. Food Data Central) (USFDC, 2022) dataset, which provides detailed food descriptions along

with their nutrient values. To enhance the system's ability to recognize the different components detailed in the food descriptions, we developed a food ontology. This ontology is constructed using the FoodOn (Dooley et al., 2018) framework and GloVe (Pennington et al., 2014) embeddings, facilitating the identification of various attributes related to the food's cooking and preparation methods. Leveraging this ontology, we create a structured food conversational dataset (Figure 1). As it is a template-based dataset, annotating it is easy and effective. We model the dataset after the state-of-the-art task-oriented dialog dataset MultiWOZ (Budzianowski et al., 2018).

To train the dialog system, we use the end-to-end dialog system PPTOD (Plug-and-Play Task-Oriented Dialogue System) Model (Su et al., 2022). PPTOD extends the T5 framework, especially designed for task-oriented dialogue (TOD) applications. To accurately provide salt content information, our dialog system will address vague user queries, (such as "What is the salt content in pork chops?"), by asking clarification questions. This approach ensures the model understands the specific preparation and consumption context of the food item, allowing us to determine the precise salt content based on how the food is prepared (For ex - beef can be consumed raw, cooked, or as part of a dish like a burger). The dataset along with the trained model is publicly available.<sup>1</sup>

Despite the remarkable proficiency of large pre-trained language models (PLMs) like GPT-3 and T5 (Brown et al., 2020; Raffel et al., 2020) in complex arithmetic reasoning tasks, they occasionally make calculation errors, especially as the mathematical operations in equations become more complex (Wei et al., 2022). Our observations align with these findings, notably that even after fine-tuning, the PPTOD model struggled to compute the salt values for various food items. Moreover, the USFDC provides salt content for standard food measurements, and users may not frame their queries in these standard terms. To rectify this, we propose **NS-PPTOD**, where we integrate PPTOD model with neuro-symbolic rules. These rules are designed to harness the strengths of PLMs while compensating for their weaknesses, enabling the system to retrieve and accurately calculate the salt content from the database. This ensures the system's adaptability in responding to queries about salt content in both standard and non-standard food quantities, thereby offering accurate salt content information and enhancing the system's overall accessibility and effectiveness.

Our experiments show that just fine-tuning a transformer model to predict salt content isn't

enough. The integration of neuro-symbolic rules significantly enhances the system's performance, evidenced by a 20% improvement in joint goal accuracy across different dataset sizes. This proves that combining pre-trained language models with neuro-symbolic rules is essential for better accuracy.

In summary, our contributions are

- We propose to develop a food conversation dataset that includes clarifying questions to infer the correct food item and its salt content.
- We finetuned PPTOD on our food conversation dataset using a few-shot approach.
- We propose NS-PPTOD which integrates PPTOD with Neuro-Symbolic rules to infer correct salt values across different food weights.
- We show a 20% increase in joint accuracy compared to the finetuned PPTOD.

## 2. Related Work

- **HealthCare Dialog Systems** Task-oriented dialogue systems have seen a significant rise in the healthcare sector, where they play a vital role in enhancing various aspects of healthcare. These systems are developed for a wide array of diseases including heart failure (Moulik, 2019; Gupta et al., 2020), cancer (Belfin et al., 2019), mental disorders (Ali et al., 2020), public anxiety (Wang et al., 2020) etc. Their applications extend to several areas, including disease diagnosis (Wei et al., 2018), patient education (Cai et al., 2023; Gupta et al., 2020), and health coaching (Zhou et al., 2022) among others. A comprehensive survey of NLP literature conducted in (Valizadeh and Parde, 2022) provides an in-depth analysis of these diverse healthcare-oriented dialogue systems, examining them from a computational perspective and highlighting their varied end-users.

(Gupta et al., 2020; Salunke et al., 2023) discuss the development of a dialog agent for self-care needs of heart failure patients, drawing upon insights from educational sessions. The work in (Kearns et al., 2020) explores the Wizard of Oz (WOZ) technique to craft a persona-based health counseling dialog dataset. Additionally, recent advancements have seen the application of Large Language Models (LLMs) in responding to patient inquiries (Chowdhury et al., 2023), though the importance of safety is emphasized. Addressing the limitations in the medical knowledge of LLMs, the study in (Li et al., 2023) undertakes the task of enhancing and fine-tuning the LLaMa model with

---

<sup>1</sup><https://github.com/anujatayal/NS-Monitoring-Salt-Content-in-Food>

a dataset of approximately 100,000 patient-doctor dialogues.

- **Pretrained Language Models (PLMs)** With the advancement in pre-trained language models (PLMs), different systems based on PLMs have been proposed including dialog systems (Lei et al., 2018; Peng et al., 2021). PLMs excel in various tasks, approaching human-like performance. Yet, they struggle in mathematical reasoning, as noted in (Wei et al., 2022). (Qian et al., 2023) shows the limitations of LLMs with complex or lengthy numerical operations. For instance, GPT-3 (Brown et al., 2020) performs well in simple two-digit additions but falters with longer numbers. Similarly, even a fine-tuned T5 model struggles with the accurate addition or subtraction of lengthy numbers (Nogueira et al., 2021), and the challenge escalates with numbers not covered in their training data.

- **NeuroSymbolic AI** Integrating neuro-symbolic approaches offers a solution by combining the inference capabilities of symbolic systems with the robustness of neural networks, creating a composite AI framework adept at reasoning, learning, and cognitive modeling (Garcez and Lamb, 2023). This blend addresses the inherent weaknesses of each system, promising enhanced performance and robustness.

To address the generalization issues in neural networks, particularly in task-oriented dialogue systems, various neuro-symbolic methodologies have been investigated. (Mehri and Eskenazi, 2021) proposes schema graphs to generalize across various unseen domains and tasks. In (Romero et al., 2021), the authors fine-tuned GPT-2 to generate the text and symbolic representations. DILOG (Zhou et al., 2020) employed inductive reasoning to formulate logical rules, enabling dialog policy training with minimal data to facilitate zero-shot domain transfer. (Arabshahi et al., 2021) used a neuro-symbolic approach to extract multi-hop reasoning and integrate commonsense in a dialog system. These strategies underscore the potential of neuro-symbolic integration to significantly improve the adaptability and efficacy of language models in complex and dynamic tasks.

- **Representing Food in Dialog Systems** Addressing the intricacies of food representation, FoodKG (Hausmann et al., 2019; Chen et al., 2021) explored knowledge graphs to represent food. FoodKG (Hausmann et al., 2019) integrates information from diverse recipe collections and the US Food Data Central (USFDC,

2022), primarily focusing on template-based queries related to ingredients and recipes. However, this framework has limitations, notably in identifying only the primary item in food descriptions and missing key details like the type of food, cooking methods, or quantity, which are included in our methodology. (Fu et al., 2022) explored the role of recommending food to improve mental health. RecipeQA (Yagcioglu et al., 2018) explored multimodal question answering within the context of recipes, while CookDial (Jiang et al., 2022) provides a platform for users to navigate and query cooking recipes more effectively.

### 3. Dataset Creation

Given the absence of a specialized dataset for conversational inquiries about salt content and the challenges in dataset collection and annotation, in this section, we show in detail how we created the dataset. Dataset creation involves developing a template-based conversational framework to accurately identify food items and their salt content. First, we used the USFDC dataset (USFDC, 2022), and created an ontology using FoodOn (Dooley et al., 2018) and Glove (Pennington et al., 2014) to describe the different components in the food description. By doing so, we were able to distinguish between different slot values. Using the ontology, we created a template-based conversational dataset that mimics human conversation, alternating between the user and the system turns. We define an average of 7 slots while creating the dataset namely- food, cook, type, animal, part, foodweight, metric.

**Data Source** To construct the dataset, we leverage the extensive food descriptions and nutritional data from the USFDC database (USFDC, 2022). It is renowned for its broad representation of diverse food items and is publicly available. The dataset was created with careful consideration of cultural differences, sourcing its data from the U.S. Department of Agriculture (USDA).

Each food description in the USFDC database consist of unstructured, comma-separated text detailing ingredients, cooking methods, and cutting styles. This format lacks clarity on the significance of each component as demonstrated in Table 6 of Appendix B. To address this, we concatenated these segments using underscores, transforming each into a distinct entity to enhance data clarity and interpretation.

Furthermore, we faced the difficulty of distinguishing whether a food item is a primary ingredient or as part of a larger dish. Items like lettuce and cheese, for example, can be both independent food items

key	Questions
food	What is the {nutrient} content in {food} ?
	How much {nutrient} in {food}?
	What is the {nutrient} content in {cook} {food}?
	How much {nutrient} in {foodWeight} {metric} of {food}?
	Can my partner with heart issues eat {food}?
	Is {food} okay for heart patients?

Table 1: Sample Template Questions that user asks to begin the conversation

and components in recipes like pizzas or burgers. To overcome this challenge, we developed a food ontology. This ontology aids in categorizing each food item more accurately, thus improving the overall understanding of the dataset.

**Ontology Construction** To develop the food ontology, an initial framework is established using FoodON (Dooley et al., 2018), focusing on key relations of *food*, *cook*, *animal* and *part*. These relations were chosen based on their significant impact on altering the salt content in various foods. Moving forward, these specific relations will be utilized to aid in creating and annotating the conversational dataset and pose clarifying questions to users to infer the salt amount in food. To address the limitations in the comprehensiveness of this initial ontology, pre-trained GloVe vectors (Pennington et al., 2014) are utilized to identify words similar to those in the ontology, thus expanding its scope. However, this method inadvertently introduces some items unrelated to food, necessitating manual preprocessing to eliminate irrelevant elements and maintain a focus on food context.

Further refinement of the ontology was needed to incorporate the items that do not associate with existing relations. A new relation, *type*, was created to integrate these components (For example type of cuisine, meat, other food ingredients etc). Drawing inspiration from FoodKG (Hausmann et al., 2019), the first item in each food description is categorized under the *food* relation. This enhanced ontology becomes a valuable tool for mapping each component of the comma-separated food descriptions to the relevant keys. In instances where multiple segments pertain to the *type* relation, their values are concatenated to ensure consistency and clarity.

### 3.1. Conversational Dataset Creation

Using the ontology and the food descriptions, we aim to develop a template-based conversational dataset that mimics human conversation. The conversation initiates with a user query about the salt content in a specific food item and alternates between the user and the system. The system poses clarification questions, drawing from ontology relations such as the type of food, cooking method,

and portion size, which are crucial determinants of salt content.

Leveraging the task-oriented dialogue framework, each turn  $t$  is annotated with a belief state  $B_t$ , encompassing a list of slot-value pairs and action states *inform* and *request*. Figure 1 illustrates a sample conversation highlighting the belief state and action state for each turn  $t$ . The figure demonstrates that within a single turn, it is possible to fill multiple slot values (the 3rd turn of the user). The dataset’s format and annotations draw inspiration from the advanced task-oriented dialog dataset, MultiWOZ (Budzianowski et al., 2018). Employing MultiWOZ as a benchmark not only validates the dataset but also enhances its replicability for crafting conversational datasets for other nutrients. Using template-based approach streamlines the annotation process, ensuring uniformity and efficiency, and reducing the cost and time needed for external annotators.

	Dialogue Statistics
# Dialogues	87,425
# Total turns	525,392
Avg turns per dialogue	6
Avg slots	7

Table 2: Dialogue Statistics of the template based conversation data

We consider 3 types of turns in a conversation. We have tried to encapsulate the range of dynamics that can occur in dialogues, ensuring the system is robust enough to handle the fluidity of human conversation.

- **Matching Answers**- This type involves turns where the user’s reply is directly pertinent to the system’s question regarding a particular slot value. An example is when the system queries about the cooking method, and the user responds specifically about the food’s cooking method (the 4th turn of the user in Figure 1).
- **Random Answers**- Occasionally, a user’s reply may not correspond to the query posed by the system. For instance, if the system asks about the *food type* and the user responds with information about the *weight* of the item, as demonstrated in the user’s second turn in Figure 1, the system needs to adapt. In such situations, the system should recognize and not repeat a question about the weight, since that information has already been provided. Instead, it should proceed to ask another question, possibly continuing to seek details about the *food type*. In .45% of conversations, a turn consists of a random answer.

- **Changing Answers-** People are very indecisive and often tend to change their responses. This category captures the scenarios where users might revise their previous responses. For example, if a user initially mentions that the cooking method is "pan-fried" but later changes it to "boiled," the system needs to update its understanding to reflect this new information, shifting its belief state from "pan-fried" to "boiled." In .45% of conversations, one user turn involves changing the answer.

The dataset is created by generating random conversations, in which, at each turn, the system’s questions and the user’s responses are randomly selected from the templates. The conversation starts with the user asking about the salt content in food. This initial question is informed by a user study of HFChat (Salunke et al., 2023), where participants frequently asked 3 categories of questions: 1) how much salt in {food} 2) Can I eat {food}? and 3) what kind of {food} can I eat?. These question types, along with similar ones, constitute the initial question, as exemplified in Table 1. The system’s objective is to engage in the dialogue by asking clarification questions to ascertain the values of different slots (cook, type, weight etc). To keep the dialogue dynamic and realistic, questions related to these slots are presented randomly. The number of questions is limited; for instance, if the food in question is eaten raw, queries about cooking methods are omitted. Users might not be aware of all system-initiated questions, in which case default values for each slot are assumed. This led to the creation of approximately 87k template-based conversations, each comprising 3-4 exchanges between the user and the system. The statistics of this extensive dataset are detailed in Table 2. As the dataset size is very large, around 87k conversations, a few-shot method is used to train the model.

## 4. Methodology

Once the conversational dataset was created, we built the NS-PPTOD model by fine-tuning PPTOD on the dataset using few-shot and integrating neuro-symbolic rules.

### 4.1. Plug and Play Task Oriented Dialog System (PPTOD)

Leveraging the T5’s model success, we adopted the PPTOD model for developing the task-oriented dialogue system. PPTOD extends the T5 framework, especially designed for task-oriented dialogue (TOD) applications, and pre-trained on a diverse range of dialog datasets spanning eleven different domains. PPTOD has integrated different TOD modules — Dialogue State Tracking (DST),

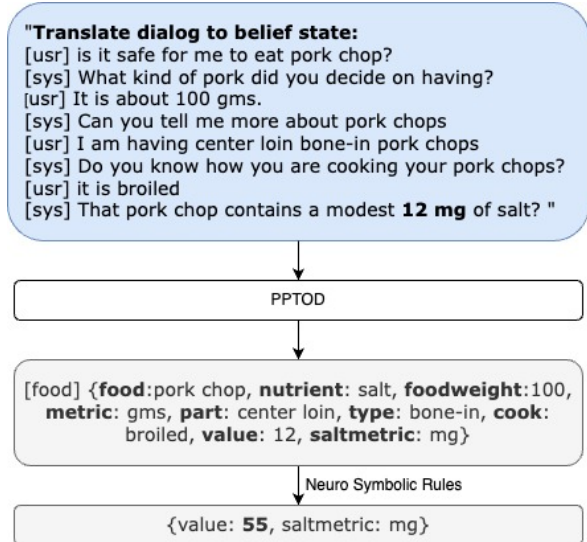


Figure 2: NS-PPTOD Model with Example: This example illustrates that PPTOD alone is not able to identify correct salt value for the food item

Natural Language Understanding (NLU), Dialogue Policy (POL) and Natural Language Generation (NLG) —into a single end-to-end architecture using a pipeline approach.

The PPTOD model is adept at in-context learning, employing customized prompts for each dialogue component, thus enhancing the relevance of model inputs to specific dialogue scenarios. Each training sample in PPTOD is represented as  $(x_t, y, z)$ , where  $t \in \text{NLU, DST, POL, NLG}$  specifies the type of Task-Oriented Dialogue (TOD) task to which the sample belongs. The term  $x_t$  refers to the task-specific prompt, formatted as *translate dialogue to A:*, where A represents different aspects such as *user intent*, *belief state*, *dialogue act*, and *system response* corresponding to the NLU, DST, POL, and NLG tasks respectively. The input dialogue context, a concatenation of all preceding dialogue utterances, encompassing both the system’s and the user’s contributions is represented by  $y$  whereas  $z$  indicates the target output text. This is exemplified in Figure 2. PPTOD is trained with a maximum likelihood objective and the loss function as shown in Equation 1 where  $\theta$  is the model parameters.

$$L_{\theta} = - \sum_{i=1}^{|z|} \log P_{\theta}(z_i | z_{<i}; x_t, y) \quad (1)$$

PPTOD also has an associated tokenizer, which supports a few-shot learning framework, enabling the system to identify new food-related terms not seen in training. We have fine-tuned it on only 1000 food-related dialogs using the same learning objective as PPTOD.



	Train Size	Epochs	Inform	Success	BLEU
PPTOD	100	8	71.43	0	24.99
NS-PPTOD	100	-	88.90	77.80	22.50
PPTOD	300	7	75.00	5.00	34.30
NS-PPTOD	300	-	81.50	63.00	26.90
PPTOD	500	9	82.86	2.86	29.81
NS-PPTOD	500	-	74.50	58.10	28.90
PPTOD	1000	7	93.50	2.70	29.00
NS-PPTOD	1000	-	85.90	71.70	30.00

Table 3: Increase in performance when using NS-PPTOD compared to PPTOD.

We chose PPTOD model for its adaptability, and its ability to support modular decomposition which in turn facilitates the incorporation of specific rules. To further enhance the model, we incorporated neuro-symbolic techniques into the DST framework, aiming to boost the system’s interpretative capabilities and its adaptability to intricate dialogue contexts. This integration seeks to fortify the dialogue system, ensuring it remains robust and flexible in managing diverse conversations.

In summary, we chose PPTOD because:

- PPTOD is a state-of-the-art (SOTA) model designed specifically for task-oriented dialogue and is based on the T5 model.
- PPTOD consists of a tokenizer making it possible to use few-shot approach to fine-tune it.
- PPTOD can be modularised to implement neuro-symbolic AI.

## 4.2. NS-PPTOD

NS-PPTOD is an integration of PPTOD with neuro-symbolic rules. The conversational dataset we created as described in section 3.1 is used to finetune PPTOD. Given the extensive size of the dataset of 87k template-based conversations, a few-shot learning approach is used. Instead of utilizing the entire dataset, limited subsets of samples are selected randomly to finetune PPTOD. As we use a few-shot approach, T5-small is used as the base model to train PPTOD. PPTOD model is finetuned for 10 epochs, employing a batch size of 16 and varying the total dataset size across 100, 300, 500, and 1000 samples. Within the dataset, 10% of the data was allocated as a development set, and another 10% served as the test set.

During the experiments - to be discussed in Section 5, we observed that the model correctly identified slot values but struggled to determine the correct salt values. Additionally, the dataset primarily comprised salt values for standard food weights, like 100 grams, 3 ounces, 1 packet, etc., and lacked data for non-standard food weights that users might inquire about. PPTOD model is fine-tuned so that it accurately learns other slot values, even if it also

Train Size	Epochs	Joint Accuracy	
		PPTOD	NS-PPTOD
100	6	55.56	73.08
300	4	51.92	72.8
500	6	58.75	83.2
1000	6	58.53	85.2

Table 4: Increase in Joint Accuracy when using NS-PPTOD compared to PPTOD across different training sizes

learns some incorrect salt values. Subsequently, we employ a neuro-symbolic approach that involves two key methodologies to correct the salt values:

- **Retrieval of the accurate salt value from the database:** This step is crucial for standard food weights where exact values are available and can be directly obtained. (as shown in Figure 2)
- **Mathematical calculation of the correct salt value for varying food weights:** This method is particularly beneficial for non-standard food weights, enabling the model to compute salt values based on weight. (In Figure 2, if the user requests the salt value for a different food weight, instead of the standard 100 grams of pork chops.)

Specifically, upon determining the slot values, the system queries the database for the salt content. If the database contains the salt value, it is then retrieved. In cases where the salt value is not available in the database, it is calculated mathematically, based on the weight of the food. We demonstrate this in Figure 2. When the dialog context and prompt, labeled *Translate dialog to belief state*, are processed through PPTOD’s DST model to determine belief states, the model successfully infers all slot values except for the salt value (12). The correct salt value (81) is then retrieved from the database by applying Neuro-Symbolic rules.

## 5. Evaluation

The evaluation of NS-PPTOD encompasses two task-oriented tasks. The first task involves end-to-end dialog modeling, assessed using metrics such as inform rate, success rate, and BLEU score (Papineni et al., 2002). The second task, the Dialog State Tracking (DST) module of PPTOD is evaluated through joint-accuracy. This evaluation was conducted over sample sizes of 100, 300, 500, and 1000.

**End-to-End Dialog Modeling** After seven epochs of training, the PPTOD model demonstrated a high inform accuracy rate of 93.5% across

1000 samples, as detailed in Table 3. Inform rate reflects the model’s adeptness in identifying slot values and the target goal slot. However, its success rate in accurately determining correct salt values was notably low, standing at just 2.7%, a point further highlighted in Table 3. This pattern of low success rate was consistent across other training sizes of 100, 300, and 500 samples. The limited 2% success rate is attributed to the PPTOD model’s tendency to predict values at random.

Implementing the NS-PPTOD model resulted in a substantial enhancement, achieving success rate of 71.7%. This improvement was not just limited to the training size of 1000 samples but was also observed consistently across the smaller training sizes. The integration of neuro-symbolic rules with PPTOD evidently plays a crucial role in enhancing the model’s capability to accurately predict and determine the correct salt values.

**DST Module** The evaluation also included the Dialog State Tracking (DST) module of PPTOD, trained for recognizing different belief states such as food, cook, type, weight, and value. Joint accuracy of 58.53% was achieved when PPTOD was used. This performance notably increased to 85.2% for 1000 samples with the addition of neuro-symbolic rules. Similar improvements in joint accuracy were observed for other training sizes. Table 4 displays the enhanced joint-accuracy achieved by using NS-PPTOD compared to PPTOD.

**Analysis** The improvement in both success rate and joint accuracy can be attributed to a key difference in approach. PPTOD, on its own, tends to memorize values instead of effectively retrieving them from the database, a critical process for accurately determining salt content. However, the application of neuro-symbolic rules in conjunction with PPTOD enhances its capability, enabling it to effectively retrieve values from the database.

### 5.1. Comparison with ChatGPT

With the advent of ChatGPT, questions have arisen about the necessity of systems such as our NS-PPTOD. To address this, we conducted a comparative analysis between NS-PPTOD and ChatGPT, to highlight the distinct capabilities and applications of NS-PPTOD that are not achievable by ChatGPT.

In this comparison, we specifically focused on their responses to queries about the salt content in foods. Appendix A illustrates ChatGPT’s response to the prompt *What is the salt amount in a pork chop?*. ChatGPT’s responses are generally comprehensive, explaining variations in salt quantity due to different cooking methods and weights, and

often include an average value. NS-PPTOD, conversely, poses targeted clarification questions to precisely identify both the food item and its salt content.

There is also a marked difference in the readability of responses from these two systems. Readability assessments, SMOG (Mc Laughlin, 1969), Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), and Flesch-Kincaid Reading Ease (FKRE) (Kincaid et al., 1975), which evaluates text comprehensibility in educational and informational contexts, were used for this comparison. The SMOG and FKGL estimates the education level needed to understand a text, while the FKRE measures the ease of comprehension.

As indicated in Table 5, understanding ChatGPT’s content requires an 11th-grade reading level, whereas NS-PPTOD’s content is accessible to those at a 2nd-grade reading level. NS-PPTOD’s responses align with recommendations for content to be understandable at a 5th-grade reading level or lower (Badarudeen and Sabharwal, 2010). This approach is particularly suitable for the intended audience of our system, the patients served by the hospital at our institution who are mostly of lower socio-economic status (and mostly minorities); this ensures the information is accessible and understandable to a wider audience.

	SMOG	FKGL	FKRE
ChatGPT	11.8	11.1	51.28
NS-PPTOD	0	2.5	95.17

Table 5: Readability Comparison of NS-PPTOD with ChatGPT: A higher FKRE score denotes better readability. Lower scores in SMOG and FKGL indicate better text readability.

## 6. Conclusion and Future Work

This work introduces NS-PPTOD, a conversational dialogue system specifically designed to help patients, particularly African-American individuals suffering from heart failure, in accurately identifying the salt content in various foods. Managing salt intake is crucial for these patients, and NS-PPTOD offers a specialized solution to aid in their dietary management. The study reveals that simply applying fine-tuning techniques to a transformer model for predicting salt content is inadequate for achieving high accuracy. However, the integration of neuro-symbolic rules into the system resulted in a significant performance boost, enhancing accuracy by 20% across different dataset sizes. This enhancement underscores the necessity of combining pre-trained language models with neuro-symbolic rules to achieve superior accuracy.

Looking ahead, there are plans to conduct an in-depth study comparing the performance of NS-PPTOD with a system based on ChatGPT. This study will involve African American patients with heart failure interacting with both systems in a within-subject setting. Utilizing pre and post-interaction surveys, the aim is to conduct both qualitative and quantitative analyses. These analyses will evaluate and contrast the effectiveness of each system offering insights into their respective impact in real-world scenarios.

## 7. Acknowledgement

For financial support, we thank the DPI Cycle 1 Seed Funding Program Award and NSF award IIS 2232307

## 8. Bibliographical References

- Mohammad Rafayet Ali, Seyedeh Zahra Razavi, Raina Langevin, Abdullah Al Mamun, Benjamin Kane, Reza Rawassizadeh, Lenhart K Schubert, and Ehsan Hoque. 2020. A virtual conversational agent for teens with autism spectrum disorder: Experimental results and design lessons. In *Proceedings of the 20th ACM international conference on intelligent virtual agents*, pages 1–8.
- Forough Arabshahi, Jennifer Lee, Mikayla Gawarecki, Kathryn Mazaitis, Amos Azaria, and Tom Mitchell. 2021. Conversational neuro-symbolic commonsense reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4902–4911.
- Sameer Badarudeen and Sanjeev Sabharwal. 2010. [Assessing readability of patient education materials: current role in orthopaedics](#). *Clinical orthopaedics and related research*, 468(10):2572–2580.
- RV Belfin, AJ Shobana, Megha Manilal, Ashly Ann Mathew, and Blessy Babu. 2019. A graph based chatbot for cancer patients. In *2019 5th international conference on advanced computing & communication systems (ICACCS)*, pages 717–721. IEEE.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multi-WOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.
- Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, Dan Berlowitz, and Hong Yu. 2023. [PaniniQA: Enhancing Patient Education Through Interactive Question Answering](#). *Transactions of the Association for Computational Linguistics*, 11:1518–1536.
- Yu Chen, Ananya Subburathinam, Ching-Hua Chen, and Mohammed J. Zaki. 2021. Personalized food recommendation as constrained question answering over a large-scale food knowledge graph. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM '21*, page 544–552, New York, NY, USA. Association for Computing Machinery.
- Mohita Chowdhury, Ernest Lim, Aisling Higham, Rory McKinnon, Nikoletta Ventoura, Yajie He, and Nick De Pennington. 2023. [Can large language models safely address patient questions following cataract surgery?](#) In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 131–137, Toronto, Canada. Association for Computational Linguistics.
- Victoria Vaughan Dickson and Barbara Riegel. 2009. Are we teaching what patients need to know? building skills in heart failure self-care. *Heart & Lung*, 38(3):253–261.
- D.M. Dooley, E.J. Griffiths, and G.S Gosal. 2018. Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2.
- Chengcheng Fu, Zhisheng Huang, Frank van Harmelen, Tingting He, and Xingpeng Jiang. 2022. Food recommendation for mental health by using knowledge graph approach. In *Health Information Science: 11th International Conference, HIS 2022, Virtual Event, October 28–30, 2022, Proceedings*, page 231–242, Berlin, Heidelberg. Springer-Verlag.

- Artur d'Avila Garcez and Luis C Lamb. 2023. Neurosymbolic ai: The 3 rd wave. *Artificial Intelligence Review*, pages 1–20.
- Itika Gupta, Barbara Di Eugenio, Devika Salunke, Andrew Boyd, Paula Allen-Meares, Carolyn Dickens, and Olga Garcia. 2020. Heart failure education of African American and Hispanic/Latino patients: Data collection and analysis. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 41–46, Online. Association for Computational Linguistics.
- Steven Haussmann, Oshani Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, Deborah L. McGuinness, and Mohammed J. Zaki. 2019. Foodkg: A semantics-driven knowledge graph for food recommendation. In *The Semantic Web – ISWC 2019*, pages 146–162, Cham. Springer International Publishing.
- Feng J. He, Monique Tan, Yuan Ma, and Graham A. MacGregor. 2020. [Salt reduction to prevent hypertension and cardiovascular disease](#). *Journal of the American College of Cardiology*, 75(6):632–647.
- Yiwei Jiang, Klim Zaporozets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2022. Cookdial: a dataset for task-oriented dialogs grounded in procedural documents. *Applied Intelligence*, pages 1–19.
- William R. Kearns, Neha Kaura, Myra Divina, Cuong Vo, Dong Si, Teresa Ward, and Weichao Yuwen. 2020. [A wizard-of-oz interface and persona-based methodology for collecting health counseling dialog](#). In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–9, New York, NY, USA. Association for Computing Machinery.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.
- Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. Association for Computational Linguistics.
- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. [Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai \(llama\) using medical domain knowledge](#).
- G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.
- Shikib Mehri and Maxine Eskenazi. 2021. [Schema-guided paradigm for zero-shot dialog](#). In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 499–508, Singapore and Online. Association for Computational Linguistics.
- Sanjoy Moulik. 2019. *DIL-A Conversational Agent for Heart Failure Patients*. Ph.D. thesis, The Claremont Graduate University.
- A Nayak, AJ Hicks, and AA Morris. 2020. Understanding the complexity of heart failure risk and treatment in black patients. *Circulation: Heart Failure*, 13(8):e007264.
- Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. [Investigating the limitations of transformers with simple arithmetic tasks](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Jing Qian, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. 2023. [Limitations of language models in arithmetic and symbolic induction](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9285–9298, Toronto, Canada. Association for Computational Linguistics.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena,

- Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Oscar J. Romero, Antian Wang, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. 2021. A task-oriented dialogue architecture via transformer neural language models and symbolic injection. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 438–444, Singapore and Online. Association for Computational Linguistics.
- Gregory A Roth, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. 2018. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1736–1788.
- Devika Salunke, Carolyn Dickens, Olga Garcia-Bedoya, Paula Allen-Meares, Barbara Di Eugenio, and Andrew Dallas Boyd. 2023. [Learnings from pilot testing hfchat, an interactive dialogue agent that can assist black/african american and hispanic/latino patients with heart failure](#). *Journal of Cardiac Failure*, 29(4):563.
- Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. [Multi-task pre-training for plug-and-play task-oriented dialogue system](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.
- USFDC. 2022. [Us food data central](#).
- Mina Valizadeh and Natalie Parde. 2022. [The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland. Association for Computational Linguistics.
- Jinping Wang, Hyun Yang, Ruosi Shao, Saeed Abdullah, and S. Shyam Sundar. 2020. [Alexa as coach: Leveraging smart speakers to build social agents that reduce public speaking anxiety](#). In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. [Task-oriented dialogue system for automatic diagnosis](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia. Association for Computational Linguistics.
- Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. [RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.
- Yue Zhou, Barbara Di Eugenio, Brian Ziebart, Lisa Sharp, Bing Liu, Ben Gerber, Nikolaos Agadakos, and Shweta Yadav. 2022. [Towards enhancing health coaching dialogue in low-resource settings](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 694–706, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Zhenpeng Zhou, Ahmad Beirami, Paul Crook, Pararth Shah, Rajen Subba, and Alborz Geramifard. 2020. Resource constrained dialog policy learning via differentiable inductive logic programming. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6775–6787, Barcelona, Spain (Online). International Committee on Computational Linguistics.

## A. Appendix: Sample ChatGPT Response

ChatGPT response to the prompt **How much salt in pork chops?**

*The salt content in pork can vary significantly based on how it is prepared. Fresh, unprocessed pork naturally contains a small amount of sodium, roughly 60 to 70 milligrams of sodium per 3-ounce serving. However, processed pork products like bacon, ham, sausages, and deli meats can have a much higher sodium content due to added salt during processing for flavor and preservation. For*

instance, a slice of bacon can have between 150 to 450 milligrams of sodium, and a serving of ham can contain over 1,000 milligrams.

For specific sodium content, it's best to check the nutrition labels on packaged pork products. For fresh pork, nutrition databases or guidelines from health organizations can provide general estimates. Remember, if you're watching your sodium intake, opting for fresh, unprocessed pork and limiting processed pork products can help manage your sodium consumption.

## B. Appendix: Sample Food Descriptions

Food Descriptions	Salt Value
Pork, fresh, loin, top loin (chops), boneless, separable lean and fat, raw	48
Pork, fresh, loin, center loin (chops), bone-in, separable lean and fat, cooked, broiled	55
Pork, fresh, blade, (chops), boneless, separable lean and fat, cooked, broiled	58
Pork, fresh, loin, sirloin (chops or roasts), boneless, separable lean only, raw	63
Pork, fresh, loin, blade (chops), bone-in, separable lean only, cooked, broiled	76

Table 6: Unstructured Food Descriptions along with their salt value for 100gms of food

# On Simplification of Discharge Summaries in Serbian: Facing the Challenges

**Anđelka Zečević, Milica Ćulafić, Stefan Stojković**

Mathematical Institute, Serbian Academy of Sciences and Arts  
Faculty of Pharmacy, University of Belgrade

Clinic for Gastroenterology and Hepatology, University Clinical Center of Serbia, Belgrade  
andjelkaz@mi.sanu.ac.rs, milica.culafic@pharmacy.bg.ac.rs, stefanstojkovic@gmail.com

## Abstract

The simplified information page (SIP) is a simplified discharge summary created to mitigate health risks caused by low medical comprehension. One of the most critical aspects of medical comprehension concerns interpreting medication instructions such as proper dosing, frequency, and duration. In our work, we examine the capacities of mainstream Large Language Models (LLMs) such as ChatGPT and Gemini to generate SIP-like medication-oriented pages based on the provided discharge summaries. We are sharing the initial qualitative assessments of our study based on a small collection of discharge summaries in Serbian, pointing to noticed inaccuracies, unfaithful content, and language quality. Hopefully, these findings might be helpful in addressing the multilingual perspective of patient-oriented language.

**Keywords:** patient safety, text simplification, discharge summary, LLMs

## 1. Introduction

Understanding discharge summaries is crucial for continuity of care and patient safety. However, low comprehension poses significant challenges in healthcare delivery. Inadequate comprehension of discharge summaries can lead to medication errors, treatment delays, and patient confusion. This is notably well-supported, especially regarding medication errors in post-discharge (Weetman et al., 2021; Alqenae et al., 2020).

Healthcare providers often struggle to communicate complex medical information effectively within discharge summaries, hindering patient understanding. On the other hand, the so-called patient literacy plays a key role in understanding discharge summaries. Regardless of education level, this literacy may be quite limited and patients, thus, face heightened challenges in understanding crucial health-related information. This issue lies at the core of health disparities (Murugesu et al., 2022).


Addressing the challenges associated with a low understanding of discharge summaries requires interdisciplinary efforts involving healthcare providers, policymakers, educators, and technology developers to enhance clarity, accessibility, and patient-centeredness in discharge communication (Geese et al., 2023; Bhati et al., 2023).

One line of research that improves patient understanding is the creation of simplified discharge instructions. The simplified information page (SIP) (DeSai et al., 2021) is a one-page patient discharge summary designed originally for emergency departments in accordance with the Centers for Medicare and Medicaid Services and


Joint Commission recommendations. It lists information related to diagnoses, recommended treatments (medications, diet, therapy, wound care, etc.), doctors or clinics needed to follow up, and symptoms or circumstances that should be monitored and urgently addressed (Figure 1). All information are presented in a simplified manner with the Flesch-Kincaid grade level 5. The SIP demonstrates that changing only the information structure and making it more accessible improves patients' comprehension. In the most critical segments that relate to medication dosage and duration, an improvement by over 22% is noticed across all demographics and education levels.

Patient Name: \_\_\_\_\_ Date: \_\_\_\_\_


Today, you were treated at the Clements Emergency Department for \_\_\_\_\_  
(Diagnosis)

 After you leave the Emergency Department, get these medicines from your pharmacy:

- Medicine Name:
  - Dose:
  - How long to take:
  - Purpose:
- Medicine Name:
  - Dose:
  - How long to take:
  - Purpose:
- Medicine Name:
  - Dose:
  - How long to take:
  - Purpose:

 You also need to visit your other doctors so they can check on you.

- Doctor Name \_\_\_\_\_, Specialty \_\_\_\_\_
  - Date, Time

 If you have these symptoms, visit the Emergency Department:

•

•

Figure 1: Simplified information page.

As the manual creation of discharge summaries in the SIP-like form requires additional personnel and time resources, we examine the potential of Large Language Models (LLMs) to recreate them. LLMs have already entered the world of biomedicine with models trained on medical publications (PubMedBERT, Gu et al., 2021), medical records (ClinicalBERT, Huang et al., 2019), or medical knowledge bases (UmlsBert, Kang et al. 2020). The performances on the relevant benchmarks, such as the BLURB - Biomedical Language Understanding and Reasoning Benchmark (Gu et al., 2021), spark the various capabilities of biomedical models. However, the complexity of medicine, the ever-growing medical knowledge, constant technology enhancements, and its safety-critical nature, always reveal the necessity for improvement.

Due to disparities in healthcare digitalization and regulatory policies across different regions, biomedical datasets in non-English languages are often scarce. The same holds for language models and appropriate language tools. Therefore, in our approach, we leverage publicly accessible chatbots such as ChatGPT<sup>1</sup> and Gemini<sup>2</sup>, to generate SIPs for the provided expert-written discharge summaries in Serbian. We mainly focus on the medication instruction part, including medication names, dosages, durations, frequencies, ways of administration, and their purpose. Although working with a small collection of discharge summaries, employing a qualitative approach, we were able to identify several pain points of language models that require additional attention and enhancement.

## 2. Related Work

Soon after progress had been made in language modeling (Vaswani et al., 2017), document summarization began to reflect notable improvements in its ability to distill key information from large volumes of text (Liu and Lapata, 2019; Lewis et al., 2019; Raffel et al., 2020), mostly in the general domain. Medical document summarization is, however, somewhat different as it poses several challenges, including handling complex medical terminology, high accuracy expectations, and preserving patient privacy and confidentiality. What come as natural tasks are the summarization of medical notes (Landes et al., 2023), medical research (Devaraj et al., 2021; Singhal et al., 2023) as well doctor-patient conversations (Abacha et al., 2023).

From the perspective of a clinician, a discharge summary represents a concise overview of the patient's course of hospitalization, treatment, and follow-up care plan that can serve as a communication tool that facilitates continuity of

care between the hospital and outpatient settings. It also represents a demanding, time-consuming administrative activity based on abundant medical documentation that is oftentimes challenging to digest. Therefore, the existing work mostly tries to alleviate this setup (Shing et al., 2021; Searle et al., 2023), the latest one being the *Discharge Me<sup>β</sup>*, a BioNLP ACL'24 Shared Task on Streamlining Discharge Documentation.

Although there are publications addressing the patient aspect of medical summarization (Zaretsky et al., 2024), they are less present. We hope that our work can help fill in the gap by combining the imperatives of both sides into a unified goal.

## 3. Experiment

As stated, our goal was to examine the capacity of publicly available mainstream LLMs to generate SIP-like medication-oriented lists easily accessible by patients. For that purpose, we collected a number of discharge summaries in Serbian, prompted ChatGPT and Gemini to generate SIPs, and manually evaluated the results we obtained.

### 3.1 Dataset

We started our work with a small collection of discharge summaries in Serbian, in total 13, provided by the Liver Transplant Unit of the Clinic for Gastroenterology and Hepatology at the University Clinical Centre of Serbia. All discharge summaries are read by one medical professional and anonymized according to the local privacy regulations by masking patient-related information, dates, ambulance names, names of practicing doctors, names of doctors who are meant to perform additional examinations, and phone numbers for scheduling examinations and obtaining information.

Due to the complexity of cases, discharge summaries were very diverse in terms of medication instructions. In total, 65 medications are covered, of which 38 are unique, including different dosage forms (tablets, capsules, droplets, sprays) and routes of administration. The average number of medications per discharge summary was 5.

The average length of the section of discharge summaries comprising prescribed medications and follow-up care instructions was 97 tokens, indicating short and condensed directions to the patients. We used Latin script as it was used in the original discharge summaries. We did not use any preprocessing step prior to utilizing language models.

---

<sup>1</sup> <https://chat.openai.com/>

<sup>2</sup> <https://gemini.google.com/>

<sup>3</sup> <https://stanford-aimi.github.io/discharge-me/>



### 3.2 Prompts

To generate SIP-like lists, we prompted the models with the template written in Serbian in a zero-shot manner. The appropriate prompt translation in English is given below.

*I will forward the patient's discharge summary from the Clinic for Gastroenterology and Hepatology. You should single out each medication, its dosage, its method of administration, its frequency, and a short explanation of what the drug is used for. In addition, single out notes related to further examinations or controls. If abbreviations appear in the result, please provide the corresponding meanings.*

Figure 2: Initial prompt translated into English.

The medication-related information such as medication name, dosage, and frequency was part of the discharge summary and easily accessible to LLMs. The method of administration depended on the medication form and was supposed to be concluded by LLMs. The same held for medication purposes and short descriptions that were to be generated based on LLMs' medical knowledge.

In cases where medication instructions vary for different days of the week, we prompted LLMs additionally for day-dependant SIP-like lists by utilizing the template below.

*Can you now create a list with appropriate medications for each day of the week?*

Figure 3: Day-dependant prompt translated into English.

## 4. Results

For each discharge summary, we prompted ChatGPT and Gemini using prepared templates with the goal of generating a medication-oriented SIP-like list. All results were manually evaluated by one medical expert by carefully comparing the original discharge summary and generated SIP lists. For each medication, the evaluator scored if the medication was present on the SIP list, if dosage, dosage form, route of administration, its frequency, and duration (when stated) were appropriate, and if a short description of the medication's purpose was correct. The total number of evaluated medications in the dataset was 65. Table 1. summarizes our main quantitative findings related to medication inconsistencies.

Both models correctly extracted medication names from discharge summaries. The exceptions were medications *Entyvio* and *Zometa*, not directly prescribed by the doctor but

mentioned as a part of the patient's existing medication protocol. However, not all medication descriptions and purposes were appropriate. For example, ChatGPT explained that *Oglition* is a cholesterol-lowering medication, while Gemini explained it is an immunosuppressor. None of these is correct, as *Oglition* is primarily used as an antidiabetic. In order to validate the claims, we relied on the expert opinion and package leaflets available on the official website of the Agency for Medicines and Medical Devices of Serbia.

	ChatGPT	Gemini
Medication omission	1	2
Inappropriate description	2	7
Inappropriate frequency	9	11

Table 1: The type and frequency of noticed inconsistencies between original discharge summaries and SIPs. The total number of revised medications is 65.

As precise dosages and dosage forms were present for each medication in discharge summaries (for example, *Advagraf caps. a 1mg 1x2*), we did not record any inconsistencies related to these parts. However, instructions related to medication frequency were the most challenging for the models to interpret and verbalize. Within discharge summaries, there were two different ways of specifying frequencies: *frequency x dosage* and *morning dosage + noon dosage + evening dosage*. Simple instructions, such as *Pravacor tabl. a 20mg 1x1* containing *1x1* form, were successfully interpreted in all cases. Instructions containing forms such as *1x2* or *2x1* often led to swapping the frequency and dosage in the generated narratives. For example, instruction *Imuran tabl. a 50mg 1x3* was interpreted as taking one tablet three times a day, every 8 hours, instead of taking one dose (consisting) of three tablets. The interpretation was even less successful in the case of fractions, for example, with frequency forms such as *1x1/2* or *2x1/4*. Gemini could not interpret these instructions at all, as denominators were excluded from the generated descriptions. Therefore, the instruction such as *Propranolol tabl. a 40mg 2x1/4* was interpreted as *Propranolol tabl. a 40mg 2x1*, leading to a much higher dosage. ChatGPT was partially successful but inconsistent within sessions. Instructions in the form *morning dosage + noon dosage + evening dosage* were correctly interpreted by both models.

Frequency instructions, such as *three times a week* or with an explicit list of the days (Monday/Wednesday/Friday) were correctly extracted by the models but partially utilized. For example, when prompted to generate SIP lists for

each day of the week, *Vigantol* droplets originally prescribed with the instruction *10 drops three times a week*, were repeated for all days by ChatGPT. Gemini could interpret this instruction correctly and even visualize the table with weekday names as headers.

Through the examples, we noticed that models can point to unspecified medical instructions. For example, the instruction to take medicine at 8h was ambiguous, as it was unclear if 8h relates to the morning or evening hours. We found these rare cases of ambiguity particularly important as they can also cause patients to feel unsure and hesitant to act.

Both models demonstrated accuracy in extracting information concerning future appointments, primarily pertaining to additional analyses, biopsies, and scans. They proved particularly useful in clarifying abbreviations associated with medical procedures and dietary regimens.

The content generated by ChatGPT was grammatical and of satisfactory quality. On the other hand, Gemini often code-switched between Serbian and English and even mixed Latin and Cyrillic script. This required additional post-processing and content validation.

When prompted several times within one session, ChatGPT started combining discharge summaries and generating improper, hallucinated, medications. Therefore, each experiment was performed within a unique session. We did not notice similar behavior while using Gemini.

## 5. Conclusion

Our study aims to improve communication between healthcare professionals and their patients. Simplified discharge summaries should translate complex medical information into a more comprehensive language, which should positively impact patient literacy in general. Patient literacy is an ever-growing concept medical experts use to denote levels by which individuals perceive or learn to comprehend health information within the decision-making processes. In confronting the existing challenges concerning the low understanding of discharge summaries, we qualitatively analyze a small dataset in the Serbian language - a language not equally covered in international protocols. Therefore, disseminating these summaries, knowledge gaps, and improvements will lead us to relevant statistics and possible solutions. Hopefully, these results will encourage scholars, stakeholders, and members of the healthcare system to strive to find more accessible paths for delivering better quality care.

Presently accessible general-purpose LLMs exhibit promise in producing simplified discharge summaries, even for languages with limited resources like Serbian. Nonetheless, these summaries do not consistently align with the original summaries concerning critical medical elements such as medication frequency or purpose, thereby compromising their reliability.

As shared results represent only a fraction of our ongoing research, the list of forthcoming activities needed for deeper validation and improvements is extensive. We plan to experiment with a larger dataset that includes other medical subfields and discharge summary writing styles. In order to alleviate the influence of prompts, we plan to design and perform additional behavior consistency experiments. Further, we plan to prepare a supporting dataset for the training of RAG architecture (Lewis et al., 2020) with the aspiration of addressing the previously highlighted accuracy and trustworthy-related observations. Finally, we plan to conduct an on-site evaluation of SIP lists with patients and medical personnel to obtain a qualitative assessment of the proposed methodology in terms of improved medical literacy and instruction comprehension.

## 6. Bibliographical References

- Abacha, Asma Ben, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. (2023). *Overview of the MEDIQA-Chat 2023 Shared Tasks on the Summarization & Generation of Doctor-Patient Conversations*. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, Toronto, Canada. Association for Computational Linguistics.
- Alqenae F. A., Steinke, D., Keers, R. N. (2020). *Prevalence and nature of medication errors and medication-related harm following discharge from hospital to community settings: a systematic review*. *Drug Saf*, 43(6):517–537.
- Bhati D., Deogade M. S., Kanyal D. (2023). *Improving Patient Outcomes Through Effective Hospital Administration: A Comprehensive Review*. *Cureus*, 15(10):e47731.
- DeSai, C., Janowiak, K., Secheli, B., Phelps, E., McDonald, S., Reed, G., & Blomkalns, A. (2021). Empowering patients: Simplifying discharge instructions. *BMJ Open Quality*, 10(3).
- Devaraj, Ashwin, Iain Marshall, Byron Wallace, and Junyi Jessy Li. (2021). *Paragraph-level Simplification of Medical Texts*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4972–4984.
- Geese, F., Schmitt, K. U. (2023). *Interprofessional Collaboration in Complex Patient Care Transition: A Qualitative Multi-Perspective Analysis*. *Healthcare*, 11, 359.

- Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. (2021). *Domain-specific language model pretraining for biomedical natural language processing*. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, no. 1: 1-23.
- Huang, Kexin, Jaan Altosaar, and Rajesh Ranganath. (2019). *Clinicalbert: Modeling clinical notes and predicting hospital readmission*. *arXiv preprint arXiv:1904.05342*.
- Kang, Tian, Adler Perotte, Youlan Tang, Casey Ta, and Chunhua Weng. (2021). *UMLS-based data augmentation for natural language processing of clinical research literature*. *Journal of the American Medical Informatics Association* 28, no. 4: 812-823.
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. (2019). *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. *arXiv preprint arXiv:1910.13461*.
- Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. *Advances in Neural Information Processing Systems* 33: 9459-9474.
- Murugesu L, Heijmans M., Rademakers J., Fransen M. P. (2022). *Challenges and solutions in communication with patients with low health literacy: Perspectives of healthcare providers*. *PLoS One*, 17(5):e0267782.
- Paul Landes, Aaron Chaise, Kunal Patel, Sean Huang, and Barbara Di Eugenio. (2023). *Hospital Discharge Summarization Data Provenance*. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 439–448, Toronto, Canada. Association for Computational Linguistics.
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. (2020). *Exploring the limits of transfer learning with a unified text-to-text transformer*. *Journal of Machine Learning Research* 21, no. 140: 1-67.
- Searle, T., Ibrahim, Z., Teo, J. and Dobson, R.J., (2023). *Discharge summary hospital course summarisation of in patient Electronic Health Record text with clinical concept guided deep pre-trained Transformer models*. *Journal of Biomedical Informatics*, 141, p.104358.
- Shaib, C., Li, M.L., Joseph, S., Marshall, I.J., Li, J.J. and Wallace, B.C., (2023). *Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success)*. *arXiv preprint arXiv:2305.06299*.
- Shing, H.C., Shivade, C., Pourdanghani, N., Nan, F., Resnik, P., Oard, D. and Bhatia, P., (2021). *Towards clinical encounter summarization: Learning to compose discharge summaries from prior notes*. *arXiv preprint arXiv:2104.13498*.
- Singhal, Karan, Shekoofeh Azizi, Tao Tu, S. S. Mahdavi, Jason Wei, Hyung W. Chung, Nathan Scales et al. (2023). *Large Language Models Encode Clinical Knowledge*. *Nature* 620, no. 7972: 172-180.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017). *Attention is all you need*. *Advances in neural information processing systems*, 30.
- Weetman, K., Spencer, R., Dale, J. et al. (2021). *What makes a "successful" or "unsuccessful" discharge letter? Hospital clinician and General Practitioner assessments of the quality of discharge letters*. *BMC Health Serv Res*. 2021 Apr 15;21(1):349.
- Zaretsky J, Kim JM, Baskharoun S, et al. (2024). *Generative Artificial Intelligence to Transform Inpatient Discharge Summaries to Patient-Friendly Language and Format*. *JAMA Netw Open*. 7(3):e240357.

# Medical-FLAVORS: A Figurative Language And Vocabulary Open Repository for Spanish in the Medical Domain

Lucia Pitarch<sup>1</sup>, Emma Herrero-Angles<sup>1</sup>, Yufeng Liu<sup>2</sup>, Daisy Monika Lal<sup>2</sup>,  
Jorge Gracia<sup>1</sup>, Paul Rayson<sup>2</sup>, Judith Rietjens<sup>3</sup>; On Behalf of the 4D PICTURE Consortium

<sup>1</sup>University of Zaragoza, Spain, {lpitarch, jgracia}@unizar.es, emmaa.herrero@gmail.com

<sup>2</sup>UCREL Research Centre, Lancaster University, UK, {yufengliu, d.m.lal, p.rayson}@lancaster.ac.uk

<sup>3</sup>Erasmus MC & TU Delft, Netherlands, j.rietjens@erasmusmc.nl

## Abstract

Metaphors shape the way we think by enabling the expression of one concept in terms of another one. For instance, cancer can be understood as a place from which one can go in and out, as a journey that one can traverse, or as a battle. Giving patients awareness of the way they refer to cancer and different narratives in which they can reframe it has been proven to be a key aspect when experiencing the disease. In this work, we propose a preliminary identification and representation of Spanish cancer metaphors using MIP (Metaphor Identification Procedure) and MetaNet. The created resource is the first openly available dataset for medical metaphors in Spanish. Thus, in the future, we expect to use it as the gold standard in automatic metaphor processing tasks, which will also serve to further populate the resource and understand how cancer is experienced and narrated.

**Keywords:** Metaphors, MetaNet, MIP, Medical NLP, Linked Data

## 1. Introduction

Cancer has been broadly envisioned as a battle. However, this aggressive framing of the disease is not always useful, as it might put unnecessary pressure on the patient (Semino et al., 2017). To facilitate the communication of the experience in different ways that can adapt to the patient's needs, the Metaphor Menu was created (Semino and Demjen, 2017). The Metaphor Menu is a compilation of different narratives to talk about cancer. While this tool has been proven to be useful (Semino and Demjen, 2017), its creation, adaptation to different people and scenarios, and expansion to other languages require a huge amount of manual effort. Advances in the development of computational methods for metaphor processing provide us with a promising scenario for the automation of metaphor search and interpretation. The mechanisation of metaphor processing would enable flexible and faster analysis of particular discourses and people. However, while great developments have been made for English and the general domain (Ge et al., 2022), little has been done in Spanish and the medical domain. To bridge this gap, we propose Medical-FLAVOURS<sup>1</sup>: the first Figurative Language And Vocabulary Open Repository for Spanish in the medical domain.

As for the methodology, we have used Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007) for the detection of metaphors from discourse, and MetaNet (Dodge et al., 2015), through its implementation in Framester (Gangemi et al., 2016), for the structured representation of the metaphors. MIP is the most commonly used method for the identification of metaphors and creation of datasets for automatic metaphor identification (Ge et al., 2022), and, MetaNet is a repository that encodes a wide range of metaphors

(mostly in English) with relevant information about them as their target and source domains, usage examples, or links between different metaphors. Framester provides a Linguistic Linked Data Hub where MetaNet is linked to other resources such as Framenet (Baker et al., 1998) or Ontolex (McCrae et al., 2017). Linkage to different resources in Framester enables a deeper description of metaphors.

We expect Medical-FLAVORS to be, not only a repository where to document different metaphors used to refer to cancer experience, but also, to serve as gold standard in automatic metaphor identification and interpretation tasks, and, as a basis for further studies which analyze the usage of metaphors in the medical domain.

In this work, we present the first version of Medical-FLAVOURS. In Section 2, similar works are discussed and the creation of our resource is motivated. Section 3 describes our work pipeline and the characteristics of the first version of our repository. Section 4 shows preliminary usage examples of our repository. Sections 4.1 and 4.2 aim to show the convenience of representing data as linguistic linked data by easily enabling the comparison of data from Reddit forums to data available in philosophy and philology literature in Section 4.1, and by facilitating multilingual comparison with English and Spanish examples in Section 4.2.

## 2. Background

Metaphor identification is one of the most basic and explored tasks in Computational Metaphor Processing. In such tasks, the most used dataset is VUA (Krennmayr and Steen, 2017), created using the Metaphor Identification Procedure (MIP) (Pragglejaz Group, 2007). MIP takes a discourse-based approach, in which the text is initially read and understood to annotate as metaphors those utterances that differ with the literal interpretation of the text. In MIP methodology this is named as the basic versus metaphorical

<sup>1</sup>Openly available at: <https://github.com/4dpicture/Medical-FLAVORS>

meaning comparison. Basic meaning as described in MIP tend to be: "a) More concrete; what they evoke is easier to imagine, see, hear, feel, smell, and taste. b) Related to bodily action. c) More precise (as opposed to vague) d) Historically older". MIP has already been used in Spanish to create the COMETA dataset (Sanchez-Bayona and Agerri, 2022). We take inspiration from this work and expand it by analyzing different Spanish varieties other than the peninsular one, and, by focusing on the medical domain.

The thorough guidelines, the need for professional annotators, and word-by-word analysis allow replicability, exhaustivity, and quality of the dataset. However, such a deep and human-focused method is very time-consuming and subjective. Further, metaphor identification alone does not provide a clear understanding of how an overall idea is being expressed, nor what conceptual patterns are being conveyed. Symbolic methods such as taxonomies and ontologies used to organize and structure knowledge can aid in the interpretation of metaphors.

Conceptual metaphors are the vertebral column of MetaNet. Other characteristics of the metaphor and other metaphors are linked to conceptual metaphors to further describe them. For example, in MetaNet<sup>2</sup>, the conceptual metaphor LIFE IS A JOURNEY is linked to verbalisations as 'She hasn't gone as far in life as her parents had hoped', defined by its linked source domain *JOURNEY* and target domain *LIFE*, and related to other metaphors as *CANCER IS A JOURNEY*. By providing not only the verbalisations but also other relevant properties of the metaphors, MetaNet makes them comparable conceptually and cross-linguistically (see Section 4.2) and provides cues to metaphors that are only conceptualized but not yet verbalized (see Section 4.1). Further, the modelisation of MetaNet via Framester Hub (Gangemi et al., 2016) as Linguistic Linked Data, provides more sophisticated tools for a deeper analysis and identification, and linkage of metaphors to other lexical resources<sup>3</sup>.

Works as Amnestic Forgery (Gangemi et al., 2018) and ImageSchemaNet (Giorgis et al., 2022) have explored metaphors as linked data and provided very interesting insights. However, they have been only explored for English. And, while in Magaña (2020), they claim to work with their data for Spanish oncological metaphors in MetaNet format, their resource is not openly available. Our resource contributes by expanding these resources to Spanish, fostering cross-lingual and domain-focused approaches to metaphor research.

### 3. Resource Description

For the metaphor annotation we started by scraping the Reddit forums (Section 3.1), once we had the relevant posts we identified the metaphors in the texts

<sup>2</sup>see [https://metaphor.icsi.berkeley.edu/pub/en/index.php/Metaphor:LIFE\\_IS\\_A\\_JOURNEY](https://metaphor.icsi.berkeley.edu/pub/en/index.php/Metaphor:LIFE_IS_A_JOURNEY)

<sup>3</sup>Framester can be accessed and queried through <https://framester.github.io/>

using MIP (Section 3.2) and modeled them using MetaNet as base (Section 3.3). As a result, we end up with a dataset in tabular format compatible with State of the Art algorithms for automated metaphor identification (Ge et al., 2022), and with a turtle file that can be inspected and linked to other relevant lexical resources.

#### 3.1. Data Selection and Scraping

Using PRAW<sup>2</sup> (Python Reddit API Wrapper), we retrieved a small sample of Spanish Reddit posts from five subreddits ('r/argentina', 'r/chile', 'r/columbia', 'r/espanol', and 'r/mexico'), using the search term *cáncer*. The relevant threads were manually selected using the selection criteria of excluding texts that used cancer as a metaphor to describe something else, such as poverty or crime, and only choosing the ones entailing the cancer experience. We finally selected 11 threads with around 700 comments and 50,000 words. The threads were written in different Spanish varieties: 3 in Mexican Spanish, 1 in Chilean, 1 in Colombian, and 6 in Argentinian. Some threads focus on different kinds of cancer: testicular in the case of the Chilean thread, cerebral cancer and leukemia in the case of the Mexican threads, while others focus on other issues related to cancer as the funding of medicines and governmental aid, and others are rather general. Given the scarcity of threads and comments, we were unable to gather sufficient threads for different aspects of cancer, such as treatment phases, stages, and people involved (patients, healthcare professionals, or carers).

#### 3.2. Metaphor Identification

For the Metaphor Identification, we used MIP procedure, described in Section 2. First the whole reddit thread was read to get the general meaning of it. Then, the text was split into words using the PyMUSAS rule based tagger for Spanish (Rayson et al., 2004). Thirdly, the comment is inspected word by word comparing its most basic meaning to its contextual meaning, if the word's meaning in context is different from its most basic meaning, then, it is marked as metaphorical. As suggested in MIP, dictionaries were used to support the basic meaning identification: namely, Diccionario de la Real Academia de la Lengua (Real Academia de la Lengua, 2023) and Diccionario de Americanismos (Asociacion de Academias de la Lengua Española, 2010) for the non-peninsular varieties.

Annotations were made by two Spanish linguists trained in MIP and doubtful annotations were then discussed with a third annotator expert in MIP and the analysis of metaphors in the medical domain. The third annotator is not a Spanish speaker, however, translating the challenging examples to English by the main annotators was beneficial for clarifying the meaning and usage of the analyzed words.

When performing the manual annotation, we encountered several problems or borderline cases in

<sup>2</sup>PRAW available at <https://praw.readthedocs.io/en/latest/>.

which the decisions taken had to be agreed upon.

1. We decided to focus on words considered open class: verbs, nouns, adjectives, and adverbs.
2. When the Reddit comment was insufficient to elucidate what the actual meaning was (see Example 1), we first tried reading the original post (see Example 2)<sup>3</sup>, if even with the original post the interpretation was not possible no annotation was made.

(1) Amigo, qué bien que todo **salió**  
Friend how good that everything **went-out**  
bien  
well  
'Dude, I'm glad everything went well'

(2) Me operaron y me pusieron una prótesis (que real, ni se nota) y la biopsia salió que no tenía ningún rastro de cáncer (tumor benigno).  
'I had an operation and a prosthesis (which is not even noticeable) and the biopsy showed that I had no trace of cancer (benign tumor).'

3. Another issue had to do with the limits that the topic of cancer has. For instance, in example 3, the metaphor is about money rather than about cancer, however, the money needed for medicines or the paperwork that must be done to get into a hospital also influences the cancer experience. Discussions about where to establish the limits on what resonates with the cancer experience and what not, were the main cause of disagreement between the annotators.

(3) Quizá te podemos ayudar con eso si teniendo algo de plata es más **rápido**.  
'Maybe we can help you with that if having some money makes it **faster**.'

4. In some cases, Reddit comments had some figurative speech, but it is not enough to affirm that those are metaphors 4. In the example below, first there is a metonymy, where the balls are being used to refer to the patient. Secondly, there is a personification of the testicles as they are attributed the feeling of sadness. While this could be interpreted as having balls with tumors is having sad balls, or as the person with cancer is a sad person, several interpretations appear. With this example we were sure there was some figurative speech, but not how to mark it, thus, we left it unannotated.

(4) El bolas tristes  
The balls sad  
'The person with testicle cancer has sad testicles'

---

<sup>3</sup>All orthographic mistakes have been corrected by the annotators

5. Regarding multi-word expressions we followed the decision taken in (Sanchez-Bayona and Agerri, 2022): if a word had its own entry in the dictionary and could be annotated as metaphor, so was the case.

In this initial effort we have completed the annotation of 3 Reddit threads, 79 comments, and 5565 words. From the annotated data, just 1.5% of the words were annotated as metaphorical. This ratio of metaphor/literal expressions is lower than ratios presented in other general domain datasets (Krennmayr and Steen, 2017; Sanchez-Bayona and Agerri, 2022), we hypothesize the decrease is due to the constraint of focusing on a very particular domain as the medical domain instead of looking for metaphors on any domain. The low number of metaphors and imbalance with literal words further highlight the need to find computational methods for metaphor processing that adapt to low-data scenarios. Our current Inter Annotator Agreement was 0.37 using Cohen's K, usually interpreted as fair agreement in the literature. Similar works on metaphor annotation using MIP report kappas between 0.6 and 0.8 (Pragglejaz Group, 2007; Steen, 2010). We expect to increase our Inter Annotator Agreement in future annotations, after the discussions and training for our first results.

A sample of the data with the annotated metaphors can be found in Table 3.2.

### 3.3. Metaphor Modeling as Linked Data

Once metaphorically used words were annotated with MIP, the evoked conceptual metaphor was also annotated. E.g., taking the example in Table 3.2 '¿en cuanto tiempo te mata un cancer testicular?' which means 'how much time does it take to testicular cancer to kill you?', once the word *mata* (*kill* in English) is identified as metaphorically used, the conceptual metaphor being evoked in the sentence is annotated, in this case CANCER IS A MURDERER. To verbalize the conceptual metaphors, English was used as metalanguage, which is the language used in MetaNet. By using English as metalanguage we could reuse and compare our data to the one in MetaNet (through Framester). To avoid slightly different verbalisations of the same conceptual metaphor (e.g., CANCER IS ASSASAIN instead of CANCER IS MURDERER) both annotators shared a common list of identified conceptual metaphors. This shared list has the metaphors in Framesters version of MetaNet as a base and was further populated by the annotators. In cases in which the same word can be associated with different conceptual metaphors (e.g., CANCER IS MURDERER and CANCER IS PERSON) both were annotated. As in MIP, conceptual metaphors were initially annotated individually by two different linguists, then results were compared and discussed. The Interannotator agreement in this case was unanimous, and in the cases where both annotators had difficulties verbalizing the metaphors a third annotator counseled them.

With the MetaNet procedure, 79 comments from 3 different Reddit threads have been analyzed. Among

DocID	SentenceID	WordIndex	Sentence	ConceptualMetaphor
12yxw6r	51	5	¿en cuanto tiempo te mata un cáncer testicular?	CANCER_IS_MURDERER

Table 1: Sample of FLAVORS metaphor annotations in tabular format compatible with SoTA automatic metaphor identification methods. DocID references the Reddit thread where the sentence is found, SentenceID is the particular comment from the thread, WordIndex is the index of the metaphorically marked word from the sentence.

them, 94 metaphoric expressions were identified and linked to 59 different conceptual metaphors (54 novel metaphors found by the annotators and 5 reused from MetaNet). Conceptual metaphors belong to 41 different source frames and 38 different target frames.

Once the linguistic metaphors in the Reddit posts were annotated with MIP and their relative conceptual metaphors, the data was translated into Resource Description Framework (RDF) structured format, following Framester’s model, and with the properties shown in Figure 1.

In Figure 2 a brief quantitative analysis of the attested conceptual metaphors can be seen. While some of the metaphors in Figure 2 are very frequent given the particular focus of the thread (e.g., TESTICLES ARE COCONUTS was only attested on the thread about testicular cancer) other metaphors were common among the different threads (e.g., CANCER IS OBJECT OR STRENGTH IS TREATMENT). Similarly, the bias of domains given the thread’s focus can also be observed on the target and source domain frequencies. We expect to decrease such biases when adding more examples from different threads.

## 4. Usage Examples

FLAVORS dataset could be used to train and test SoTA models on automated metaphor identification, where progress has been made for English and the general domain, but less for Spanish and the medical domain. Improving these models could aid in finding relevant narratives about the experience of cancer.

Further, modeling metaphors as linked data can structure the information encoded in metaphors (e.g., which semantic domains are being mostly used to refer to particular experiences of cancer), what metaphors are being expressed, and which ones could be potentially expressed but have not been yet verbalized in daily discourse and just in critical literature revisiting cancer experience or allow multilingual comparison. Further, the linkage of our FLAVORS to other resources such as FrameNet (Baker et al., 1998) could elucidate what elements of the cancer experience are being highlighted (e.g., the patient’s role as in CANCER IS WAR in sentences such as ‘I am tired of fighting cancer’) or it could be linked to other lexical resources as WordNets to group metaphors and find their hypernymic and hyponymic metaphors (e.g., CANCER IS JOURNEY by using WordNets could be linked to CANCER IS MOVEMENT).

In this section we introduce two brief examples to illustrate how modeling metaphors as MetaNet can be beneficial to show gaps and potential metaphors both monolingually (Section 4.1) and crosslingually (Section 4.2).

### 4.1. With Manual Efforts in Spanish Data

By modelling metaphors as Linked Data we can compare the ones elicited in the Reddit posts with the ones described in literature which might not be yet verbalized.

We reviewed the following works: Chacón (2008), Paituví (2019), Taylor and Velázquez (2020), Barrera (2016), Barbosa and Coll-Planas (2015). At first sight, we could observe that many metaphors used by patients on Reddit were very common and also represented in the more scientific literature: CANCER IS A JOURNEY (Paituví, 2019), CANCER IS A MURDERER (Chacón, 2008), CANCER IS A BATTLE (Chacón, 2008), CANCER IS A STOP IN LIFE (Barbosa and Coll-Planas, 2015), LOSE TESTICLES IS LOSING MANHOOD (Barrera, 2016). However, other metaphors are not usually spoken by the patients or families of cancer patients, while still having significance to understanding how cancer is perceived by general society. The “Pink Ribbon Culture” as it is explained in Paituví (2019) related to breast cancer women makes a metaphor that could be something like: women with breast cancer are like children. Another example is the concept of silence related to cancer showed in Taylor and Velázquez (2020) article. The cancer is seen as something taboo, something to avoid or something shameful, some of the metaphors we could find are: BEING SILENT IS BEING ASHAMED (Paituví, 2019), SILENCE IS INVISIBILIZATION (Paituví, 2019). This can have serious consequences, such as patients not voicing their concerns for fear of not being taken seriously. Other kind of examples are those that from a metaphor like CANCER IS A BATTLE (Chacón, 2008) entails other metaphors like DYING IS LOSING THE BATTLE (Chacón, 2008), so that the patient with cancer that dies is drawn in society as a loser, or people might think that they haven’t fought enough to get ride of cancer.

Modelling metaphors as Linked Data allowed us to see not only what is being said, but also what is being silent and gives us cues to see where new metaphors might be verbalized to explore other framings of cancer.

### 4.2. With English Data

In comparison to MetaNet, where only seven mappings are cancer-related, our Medical-FLAVORS provides a more nuanced differentiation of different aspects of the cancer experience, including cancer itself (e.g., CANCER IS MURDERER), cancer patients (e.g., BODY IS CONTAINER), their general experience of being ill with cancer (e.g., HAVING CANCER IS BEING DEPRIVED), their psychological wellbeing (e.g., BEING PSYCHOLOGICALLY BAD IS BEING MADE OF SHIT), their relationships with other key stakeholders,

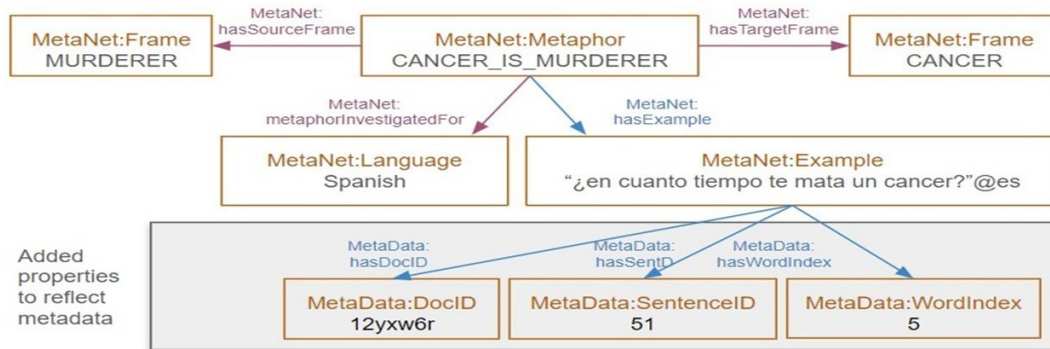


Figure 1: Example of our adaptation of MetaNet model to our examples

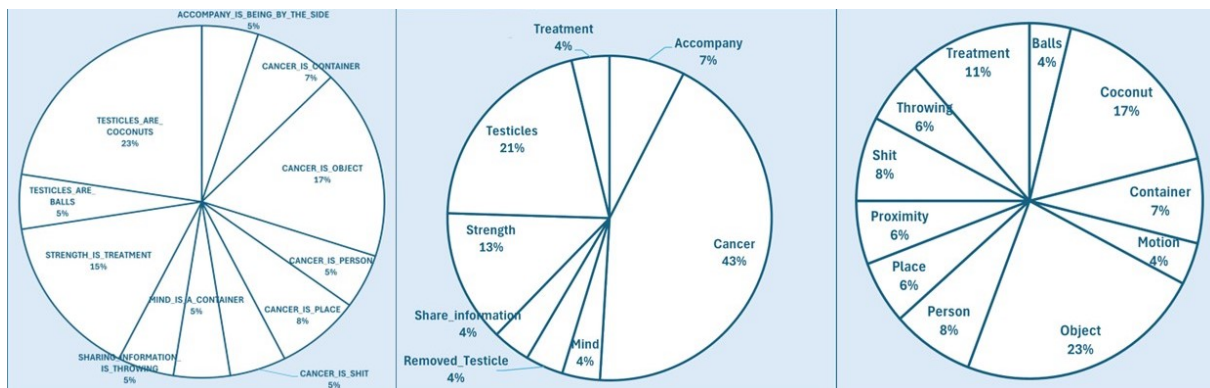


Figure 2: Frequency count of most attested conceptual metaphors, target domains, and source domains in the annotated Reddit threads

such as families and friends (e.g., ACCOMPANY IS BEING BY THE SIDE) and health professionals (e.g., BAD DOCTORS ARE SHIT), as well as treatment (e.g., TREATMENT IS A BALANCE). Nevertheless, while our study is more contextualized in a cancer scenario, it also shares or builds up on the more generic mappings in MetaNet, especially in terms of the Orientation and Container metaphors that are common in human thinking (cf. CANCER IS CONTAINER, IMPROVEMENT IN MOOD IS UPWARD MOTION). On the other hand, our project’s contextualization is also illustrated through the use of more specific mappings under a generic mapping, such as the specific mapping CANCER IS A STOP IN LIFE in Medical-FLAVORS under the generic mapping LIFE IS A JOURNEY in both our project and the MetaNet project.

However, the current datasets are quite limited, as many generic but essential mappings in the MetaNet were not found in the current study. For instance, mappings related to “Action”, “Change-of-state”, “Communication” and “Control” are not commonly witnessed in our current Spanish data, even though these elements are indispensable in a cancer scenario regardless of the country contexts. Moreover, as we have found for the mappings above, the language used in the cancer domain is usually more specific but also closely linked to some generic mappings. Hence, it is likely to find the specific versions of those mappings related to “Action”, “Emotion”, “Change-of-state”, “Com-

munication” and “Control” in our Medical-FLAVORS project, if more data is analyzed and trained.

## 5. Conclusions and Future Work

In this paper we present the first version of Medical FLAVORS, the first openly available resource for the computational and manual analysis of Spanish metaphors in the medical domain. We present our data both in VUA dataset format which can be exploited with state-of-the-art automated metaphor identification algorithms to further populate the dataset, and in MetaNet Linked Data format which enables the further linking of the resource and inference-making on the metaphors with other relevant resources as FrameNet or Ontolex based lexical resources.

## 6. Acknowledgements

We are grateful for Elena Semino’s valuable feedback and for the reviewer’s comments. Supported by the EU research and innovation program HORIZON Europe in the “4D PICTURE” project under grant agreement 101057332, by the Innovate UK Horizon Europe Guarantee Programme, UKRI Reference Number 10041120, by DGA/FEDER, the European Social Fund through the “Ramón y Cajal” program (RYC2019-028112-I), by DGA Government predoctoral fellowship, and by the Spanish project PID2020-113903RB-I00 (AEI/FEDER, UE).



## 7. Bibliographical References

- Asociación de Academias de la Lengua Española. 2010. *Diccionario de Americanismos*. Santillana.
- Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Annual Meeting of the Association for Computational Linguistics*.
- Mariona Visa Barbosa and Gerard Coll-Planas. 2015. Compartir la enfermedad on-line: Narrativas de restitución y búsqueda en blogs de mujeres con cáncer de mama. *Zer: Revista de estudios de comunicación = Komunikazio ikasketen aldizkaria*, 20:195–210.
- Lucero Fuentes Barrera. 2016. ¿es posible sentirse cómodos con el cuerpo que habitamos? narrando la experiencia del padecer en hombres con cáncer de testículo. *CONAMED*, 21:109–112.
- Inmaculada Chacón. 2008. El cáncer y su metáfora. *Eidon: revista de la fundación de ciencias de la salud*, 28:40–44.
- Ellen Dodge, Jisup Hong, and Elise Stickles. 2015. [MetaNet: Deep semantic automatic metaphor analysis](#). In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 40–49, Denver, Colorado. Association for Computational Linguistics.
- Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti, and Diego Reforgiato Recupero. 2016. Framester: A wide coverage linguistic linked data hub. In *International Conference Knowledge Engineering and Knowledge Management*.
- Aldo Gangemi, Mehwish Alam, and Valentina Presutti. 2018. Linked metaphors. In *International Workshop on the Semantic Web*.
- Mengshi Ge, Rui Mao, and Erik Cambria. 2022. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application.
- Stefano De Giorgis, Aldo Gangemi, and Dagmar Gromann. 2022. Imageschemanet: A framester graph for embodied commonsense knowledge. *Semantic Web*.
- Tina Krennmayr and Gerard Steen. 2017. *VU Amsterdam Metaphor Corpus*, pages 1053–1071. Springer Netherlands, Dordrecht.
- Dalia Magaña. 2020. Local voices on health care communication issues and insights on latino cultural constructs. *Hispanic Journal of Behavioral Sciences*, 42(3):300–323.
- John P McCrae, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21.
- Marisa Paituví. 2019. Oncogrrrls, narrativas colaborativas sobre el cáncer de mama. autoetnografía de una micro-utopía. pages 239–256.
- Pragglejaz Group. 2007. [MIP: A method for identifying metaphorically used words in discourse](#). *Metaphor and Symbol*, 22(1):1–39.
- Paul Rayson, Dawn Archer, Scott Piao, and Tony McEnery. 2004. The UCREL semantic analysis system. In *Proceedings of the Beyond Named Entity Recognition Semantic Labelling for NLP tasks workshop, Lisbon, Portugal, 2004*, pages 7–12.
- Real Academia de la Lengua. 2023. *Diccionario*. Real Academia Española.
- Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and crosslingual metaphor detection. *arXiv preprint arXiv:2210.10358*.
- Elena Semino and Zsofia Demjen. 2017. The cancer card: Metaphor, intimacy, and humor in online interactions about the experience of cancer. *Metaphor: Embodied cognition and discourse*, page 181.
- Elena Semino, Zsófia Demjén, Andrew Hardie, Sheila Payne, and Paul Rayson. 2017. *Metaphor, cancer and the end of life: A corpus-based study*. Routledge.
- Gerard Steen. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. Converging evidence in language and communication research. John Benjamins Publishing Company.
- Tomas Loza Taylor and Josefina Ramírez Velázquez. 2020. El cuerpo silenciado: Reflexiones en torno a la experiencia de personas con cáncer terminal y sus metáforas. *Corpo-grafías: Estudios críticos de y desde los cuerpos*, 7:131–144.

# Generating English Synthetic Documents with Clinical Keywords: A Privacy-Sensitive Methodology

Simon Meoni, Éric de la Clergerie, Théo Ryffel

Inria/Arkhn (Paris, France), Inria (Paris, France), Arkhn (Paris, France)  
simon.meoni@arkhn.com, theo@arkhn.com, Eric.De\_La\_Clergerie@inria.fr

## Abstract

Electronic Health Records (EHR) store valuable patient-staff interaction data. These notes, often unstructured to save healthcare personnel time, can be challenging to analyze manually. Proprietary online LLMs have demonstrated impressive results in analyzing EHR notes. However, Clinical NLP faces unique challenges due to the sensitive and specialized nature of the data. Sending patient information via external APIs poses privacy risks, and hospitals require customized NLP systems to align with their practices. Developing customized LLMs using specific training datasets is crucial to address these challenges. We propose generating synthetic training data using keywords extracted without confidential information. Furthermore, we introduce a reward mechanism that iteratively refines the quality of synthetic documents. This involves scoring synthetic candidates against real clinical reports using a semantic textual similarity score and performing an alignment step to align the model with its best-scored utterances.

**Keywords:** Style Transfer, Data Generation, LLM, Reinforcement Learning, Data Privacy

## 1. Introduction

Electronic Health Records (EHR) contain patient and healthcare staff interactions. Professionals record their impressions, observations, and various medical procedures performed. Despite the computerization of clinical documents, notes remain fairly expressive and in a free format to save time for healthcare personnel and allow for the description of unusual situations (Rosenbloom et al., 2011; Wu et al., 2022). These notes can be handy for medical professionals, but analyzing them manually is daunting. Natural Language Processing (NLP) techniques come here, as they speed up the decision processes (Zhou et al., 2022; Wu et al., 2022). In recent years, Proprietary Online Large Language Models (LLMs) such as ChatGPT have shown impressive results using zero or few-shot techniques in analyzing these notes (Agrawal et al., 2022; Meoni et al., 2023; Hu et al., 2024). However, clinical NLP faces challenges that arise from the sensitive, confidential, and specialized nature of its data—sending such information through an external API risks patient privacy. Hospitals must maintain control over their NLP systems due to their unique practices and environments. Creating customized LLMs is an important issue.

A specific training dataset is required to develop such a model with clinical skills. Accessing real clinical data to constitute this dataset remains very complex and requires anonymization, which is time-consuming and expensive. Another option is to generate synthetic clinical notes that resemble real data and do not contain any patient identifiers (Melamud and Shivade, 2019; Ive et al., 2020). This approach reduces human intervention and is more compliant with regulation laws.

## 2. Contributions

This work introduces a novel method for generating synthetic documents, enforcing privacy preservation by design, only using sparsely pseudo-anonymised data. Our key contributions include:

**Privacy-conscious Synthetic Document Generation:** We propose a methodology that utilizes a small amount of manually anonymized data to generate synthetic documents. These documents are then used to supervise fine-tuned generators, as illustrated in Figure 1.

**Incorporating Clinical Keywords:** We enhance synthetic document generation by enriching prompts with privacy-safe keywords as illustrated in Figure 2. Using QuickUMLS (Soldaini and Goharian), we generate candidate documents based on keywords extracted from real Clinical reports (CR). The keywords guide the model to produce text that closely aligns with specified content and style criteria.

**Reward Mechanism:** We introduce an iterative refinement process for enhancing the quality of the synthetic documents generated by the seeded model. This method involves two main key steps:

- Scoring the synthetic candidates through comparison with private or real CR using an SEM-SCORE evaluator model in the private side returning only scores to the public side;
- Aligning the model with its best utterances using Direct Preference Optimization (DPO) (Rafailov et al., 2023).



---

**Algorithm 1: Reward Training Algorithm**

---

```
Input :  $D_{\text{train}}$  = train dataset;  $ratio_{\text{sft}}$  = sft ratio;  $ratio_{\text{gen}}$  = gen ratio;  $M_{\text{gen}}$  = generative model;  
          $M_{\text{score}}$  = evaluator model;  $p$  = percentile filter value;  $N$  = number of candidates to  
         generate;  
Output :  $M_{\text{gen}}$   
 $K_{\text{train}} \leftarrow \text{ExtractKeywords}(D_{\text{train}})$   
 $D_{\text{sft}}, K_{\text{sft}} \leftarrow \text{Anonymize}(\text{Sample}(D_{\text{train}}, K_{\text{train}}, ratio_{\text{sft}}))$   
 $D_{\text{gen}}, K_{\text{gen}} \leftarrow \text{Sample}(D_{\text{train}}, K_{\text{train}}, ratio_{\text{gen}})$   
// Seed Step  
 $M_{\text{gen}} \leftarrow \text{Supervised fine-tune } M_{\text{gen}} \text{ on pairs in } (K_{\text{sft}}, D_{\text{sft}})$   
for  $step = 1$  to  $steps$  do  
  // Generation Step  
   $D_{\text{gen}}^* \leftarrow \text{generate new } N \text{ candidates with } M_{\text{gen}} \text{ per } k \in K_{\text{gen}}$   
  // Scoring Step  
  if  $step = 1$  then  
    // Building the evaluator model  
     $D_{\text{contr}}^*, D_{\text{contr}} \leftarrow \text{Sample}(D_{\text{gen}}^*, D_{\text{gen}}, ratio_{\text{contr}})$   
     $M_{\text{score}} \leftarrow \text{ContrastiveTrain}(M_{\text{score}}, \text{neg}=D_{\text{contr}}^*, \text{pos}=D_{\text{contr}})$   
  endif  
   $D_{\text{score}} \leftarrow \text{score } D_{\text{gen}}^* \text{ over } D_{\text{gen}} \text{ with } M_{\text{score}} \text{ over the candidates generated}$   
   $D_{\text{dpo}} \leftarrow \text{for each data point in } D_{\text{score}}, \text{ keep a pair of candidates, then filter pairs on percentile } p$   
  // Alignment Step  
   $M_{\text{gen}} \leftarrow \text{DPO Alignment } M_{\text{gen}} \text{ on } (K_{\text{gen}}, D_{\text{dpo}})$   
endfor
```

---

<s> [INST] As a doctor, you must write an original 'History of Present Illness' (HPI) section for a discharge summary. Your response should capture the essence of a patient's health journey and recent medical experiences while strictly using all the provided keywords, preserving the order. You must adopt a medical telegraphic style, abbreviated, characterized by concise and direct language.

**Keywords:** *cirrhosis c, portal, esophageal varices, SBP, angioectasias, gout, liver, note, fractured, left wrist, hip, note, admissions, asymptomatic, range, PRBCs, angioectasias, estrogen, bleeding, hospital course, SBP, guaiac, stool, L wrist, L hip, consulted, L wrist, leg, said, surgical, pantoprazole, gtt, morphine, hip pain, PRBCs, transfer, sat, hip pain, esp, feeling, note, iron, stools, stools* [/INST]

Figure 2: Example of prompt with injected keywords

1. *pre-processing*: we extract from Mimic-III the clinical notes from the clinical event row. We select only the *Discharge Summaries* from these clinical notes and parse them to retrieve the *History of Patient Illness* section. we use

them as data points for our  $D_{\text{train}}$ . On average, the data points consist of a 248-word excerpt.

2. *keywords extraction*: We project UMLS concepts using QuickUMLS over  $D_{\text{train}}$ . QuickUMLS is an unsupervised biomedical concept extraction based on pattern matching that guarantees only medical concepts are extracted and no identifying information. We obtain  $K_{\text{train}}$  (cf. Section 4) used to enrich the prompts, as illustrated in Figure 2. On average, we extract 58 keywords per data point.
3. *post-processing*: We filter out data points without keywords. We keep the keywords ordered to force the model to follow the same narrative as the ground truth. In this way, we constitute a dataset of 5602 excerpts as data points. We use 70% (either 3921 data points) of these data points as a train set ( $D_{\text{train}}$ ) and 30% (either 1680 data points) as a test set ( $D_{\text{test}}$ ).

**Evaluation:** To monitor and evaluate  $M_{\text{gen}}$  progression, we also train a model ( $M_{\text{ref}}$ ) supervised fine-tune overall  $D_{\text{train}}$ .  $M_{\text{ref}}$  is used as a witness and reference, trained without privacy concerns. We compare the performance of  $M_{\text{gen}}$  and  $M_{\text{ref}}$  along the different  $step$  as described in Algorithm 1. Additionally, we calculate a baseline where we compute SEMSCORE between the real  $D_{\text{test}}$  and  $K_{\text{test}}$  as illustrated in Table 1.

## 6. Results and Discussion

Our experimental setup aimed to evaluate the performance of our model trained with the method described in section 4 with different  $ratio_{sft} \in \{4\%, 6\%\}$  (i.e 4% and 6% is equal to 156 and 235 data points, respectively) against  $M_{ref}$ , a reference model fine-tuned with the full  $D_{train}$ . To gauge the different fine-tuned scenarios, we use two  $M_{score}$  fine-tuned as described in Section 4.3 on  $D_{test}$ .

We observe monotonous score improvements over steps.  $M_{gen}^{6\%}$  model even outperforms at step 2 the score of  $M_{ref}$ , highlighting the relative efficiency of alignment in refining the generated documents' quality over successive iterations. Moreover,  $M_{score}^{4\%}$  trained on lower-quality synthetic data tends to overestimate the higher-quality generated documents. This overestimation is observed in both  $M_{ref}^{100\%}$  and  $M_{gen}^{6\%}$ . However, the same trends have been observed with any evaluator.

These improvements can be attributed to various factors. The scoring mechanism allows for a focused learning approach, where a model iteratively learns from the chosen examples and adjusts away from the rejected ones. Such a dynamic refinement process effectively distills the desired style and content characteristics along the steps.

Comparing different data ratios further reveals the nuanced impact of training data volume on model performance. It underscores the efficiency of DPO in leveraging available data regardless of the seed dataset size to achieve superior outcomes.

	steps	$M_{score}^{4\%}$	$M_{score}^{6\%}$
baseline	-	48.43	49.35
$M_{ref}^{100\%}$	-	<b>74.48</b>	72.48
$M_{gen}^{4\%}$	0	67.95	65.94
	1	71.53	69.18
	2	72.25	70.12
$M_{gen}^{6\%}$	0	70.78	67.26
	1	72.54	70.78
	2	76.10	<b>74.37</b>

Table 1: SEMSCORE evaluation for models  $M_{gen}^a$  with  $a = r_{sft} \in \{4\%, 6\%, 100\%\}$  using the different evaluators  $M_{score}^b$  with  $b = r_{sft} \in \{4\%, 6\%\}$ . The grey scores denote cross-evaluation where  $a \neq b$ .

## 7. Related Works

**Synthetic Data Generation:** Recent works tend to generate synthetic data with privacy concerns (Li et al., 2023a; Hiebel et al., 2023; Xie et al., 2024; Li et al., 2024). For instance, (Kweon et al., 2023) proposes to train LLMs for different purposes using

synthetic clinical data generated by online LLMs. This way, (Xie et al., 2024) has developed AUG-PE, a high-quality differential privacy synthetic text generation method leveraging API access. Furthermore, the work by (Li et al., 2024) introduces Generalized Instruction Tuning (GLAN). Unlike previous approaches that rely on seed or existing datasets, GLAN uses a pre-curated taxonomy of human knowledge and capabilities as input to generate instructions across all disciplines.

**Self-Rewarding:** Reinforced Self-Training is an offline RL algorithm proposed by (Gulcehre et al., 2023) for self-align LLMs generating a dataset from the initial LLM policy and using it to improve the policy via offline RL. Instruction back translation, proposed by (Li et al., 2023b), is a scalable method that automatically labels human-written text with corresponding instructions by finetuning a language model on a small seed dataset and a web corpus to generate and selecting high-quality examples for further finetuning. (Yuan et al., 2024) use the trained LLM to provide rewards via LLM-as-a-Judge prompting, leading to improvements in both instruction following and reward provision.

Our method differs from the methods described above in these differences :

- only the score is accessible to the learner, preserving the privacy of real EHR.
- only public medical keywords extracted from EHRs are used to generate synthetic data
- the SEMSCORE evaluator can be easily hosted in a clinical environment and the generator LLM may be shared with external actors.

## 8. Future Directions

This study has laid the groundwork for generating synthetic documents enforcing privacy protection. It leverages a small anonymized seed dataset for supervised fine-tuning alongside keyword-augmented prompts and refinement steps based on synthetic candidates to reduce human intervention. Despite its promise, shortcomings and openings need to be addressed.

As we can annotate privacy-free generated documents using online models for NER and EL tasks, we can train models for downstream tasks using the generated data and compare them with models trained on real data to reinforce our evaluation. Moreover, We envision advancing our methodology by exploring a mixture of evaluation metrics incorporating more sophisticated evaluators and

experimenting with alternative reinforcement learning such as KTO (Ethayarajh et al., 2024), or IPO (Azar et al., 2023). These would rely on classical metrics in style transfer and embrace notions of document quality (Jin et al., 2022). Such advancements could streamline the generation process, reduce the computational cost, and enhance synthetic documents' overall quality and applicability in privacy-sensitive applications.

## 9. Bibliographical References

- Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. [Large Language Models are Few-Shot Clinical Information Extractors](#). ArXiv: 2205.12689.
- Ansar Aynedinov and Alan Akbik. 2024. [Sem-Score: Automated Evaluation of Instruction-Tuned LLMs based on Semantic Textual Similarity](#). ArXiv:2401.17072 [cs].
- Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. [A General Theoretical Paradigm to Understand Learning from Human Preferences](#). ArXiv:2310.12036 [cs, stat].
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient Fine-tuning of Quantized LLMs](#). ArXiv:2305.14314 [cs].
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. [KTO: Model Alignment as Prospect Theoretic Optimization](#). ArXiv:2402.01306 [cs].
- Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. [Reinforced Self-Training \(ReST\) for Language Modeling](#). ArXiv:2308.08998 [cs].
- Nicolas Hiebel, Olivier Ferret, Karen Fort, and Aurélie Névéal. 2023. [Can Synthetic Text Help Clinical Named Entity Recognition? A Study of Electronic Health Records in French](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2320–2338, Dubrovnik, Croatia. Association for Computational Linguistics.
- Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. [Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering](#). ArXiv:2303.16416 [cs].
- Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, So-main Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. [Generation and evaluation of artificial mental health records for Natural Language Processing](#). *npj Digital Medicine*, 3(1):1–9. Publisher: Nature Publishing Group.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7B](#). ArXiv:2310.06825 [cs].
- Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. [Deep Learning for Text Style Transfer: A Survey](#). *Computational Linguistics*, 48(1):155–205. Place: Cambridge, MA Publisher: MIT Press.
- Sunjun Kweon, Junu Kim, Jiyou Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2023. [Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes](#). ArXiv:2309.00237 [cs].
- Loïc Lanelongue, Jason Grealey, and Michael Inouye. 2021. [Green Algorithms: Quantifying the Carbon Footprint of Computation](#). *Advanced Science*, 8(12):2100707.
- Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. 2024. [Synthetic Data \(Almost\) from Scratch: Generalized Instruction Tuning for Language Models](#). ArXiv:2402.13064 [cs].
- Rumeng Li, Xun Wang, and Hong Yu. 2023a. [Two Directions for Clinical Data Generation with Large Language Models: Data-to-Label and Label-to-Data](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7129–7143, Singapore. Association for Computational Linguistics.

- Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023b. [Self-Alignment with Instruction Backtranslation](#). ArXiv:2308.06259 [cs].
- Oren Melamud and Chaitanya Shivade. 2019. [Towards Automatic Generation of Shareable Synthetic Clinical Notes Using Neural Language Models](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 35–45, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Simon Meoni, Eric De la Clergerie, and Theo Ryffel. 2023. [Large Language Models as Instructors: A Study on Multilingual Clinical Entity Extraction](#). In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 178–190, Toronto, Canada. Association for Computational Linguistics.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. [Direct Preference Optimization: Your Language Model is Secretly a Reward Model](#). ArXiv:2305.18290 [cs].
- S. Trent Rosenbloom, Joshua C. Denny, Hua Xu, Nancy Lorenzi, William W. Stead, and Kevin B. Johnson. 2011. [Data from clinical notes: A perspective on the tension between structure and flexible documentation](#). *Journal of the American Medical Informatics Association*, 18(2):181–186.
- Luca Soldaini and Nazli Goharian. QuickUMLS: a fast, unsupervised approach for medical concept extraction.
- Honghan Wu, Minhong Wang, Jinge Wu, Farah Francis, Yun-Hsuan Chang, Alex Shavick, Hang Dong, Michael T. C. Poon, Natalie Fitzpatrick, Adam P. Levine, Luke T. Slater, Alex Handy, Andreas Karwath, Georgios V. Gkoutos, Claude Chelala, Anoop Dinesh Shah, Robert Stewart, Nigel Collier, Beatrice Alex, William Whiteley, Cathie Sudlow, Angus Roberts, and Richard J. B. Dobson. 2022. [A survey on clinical natural language processing in the United Kingdom from 2007 to 2022](#). *npj Digital Medicine*, 5(1):1–15. Publisher: Nature Publishing Group.
- Chulin Xie, Zinan Lin, Arturs Backurs, Sivakanth Gopi, Da Yu, Huseyin A. Inan, Harsha Nori, Hao-tian Jiang, Huishuai Zhang, Yin Tat Lee, Bo Li, and Sergey Yekhanin. 2024. [Differentially Private Synthetic Data via Foundation Model APIs 2: Text](#). ArXiv:2403.01749 [cs].
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. [Self-Rewarding Language Models](#). ArXiv:2401.10020 [cs].
- Nina Zhou, Qiucheng Wu, Zewen Wu, Simeone Marino, and Ivo D. Dinov. 2022. [DataSifterText: Partially Synthetic Text Generation for Sensitive Clinical Notes](#). *Journal of Medical Systems*, 46(12):96.

## Appendix A. Carbon Footprint

The algorithm 1 with the experimentation protocol detailed in 5 runs in 40h on 4 GPUs NVIDIA A100 PCIe, and draws 46.34 kWh. Based in France, this has a carbon footprint of 2.38 kg CO<sub>2</sub>e, which is equivalent to 2.59 tree-months (calculated using green-algorithms.org v2.2 (Lannelongue et al., 2021)).

## Appendix B. Outputs

```
<s>[INST]As a doctor, you must write an original 'History of Present
  ↳ Illness' (HPI) section for a discharge summary.
Your response should capture the essence of a patient's health journey and
  ↳ recent medical experiences, while strictly using all the provided
  ↳ keywords conserving the order.
You must adopt a medical telegraphic style, abbreviated, characterized by
  ↳ concise and direct language.
Keywords: metastatic, RCC, pancreas, reports, chills, tylenol, reports,
  ↳ rib pain, lying, chills, reports, dark stools, fever, zosyn, headache,
  ↳ contacts, anxious, pain, dysuria, joint pain, rash, hypotensive,
  ↳ asymptomatic, given, lactate, baseline, guaiac, stool, saw, stent,
  ↳ pancreatic, blood, tomorrow, treated, cholangitis, given, sat, ARF,
  ↳ reports, anxious[/INST]
```

Figure 3: An example of prompt for the Figures 4,5,6

```
This is a 67 y.o male with h.o metastatic RCC to the pancreas,
recent ICU course for UGIB (12units pRBCs) who reports sudden
intermittent chills since wednesday for which he took tylenol.
Pt also reports R.side gnawing rib pain, while lying in bed
before the onset of chills. In addition, pt reports dark stools
for the last few days which started after taking "iron pills".
Pt states he went to [**Hospital1 2436**] ED because of a fever of 101.3,
however he felt better and did not want to wait to be seen. He
returned to [**Hospital1 2436**] today and was transferred to [**Hospital1
  ↳ 18**] after
a dose of zosyn, HCT 25. Pt denies headache/dizziness/blurred
vision, URI/cough, sick contacts, CP, +palp when anxious, -abd
pain/n/v/d/brbpr, dysuria/hematuria, joint pain, rash,
paresthesias.
.
At [**Hospital1 18**], pt found to be hypotensive to 75/40, asymptomatic.
  ↳ He
was given 3L IVF, lactate 6.8. HCT 22.8 from a baseline of 35 a
few weeks ago. He was found to have black, guaiac +stool. GI saw
pt, pt s/p stent to pancreatic ampulla, ?blocked from blood.
Plan is to transfuse, ERCP tomorrow. ED also treated for
possible cholangitis/sepsis and pt was given dose of vanco.
Vitals 99.2, BP 99/66 HR 88 sat 98% on RA. Access 3PIV's 2,
20's, 18. Pt also found to be in ARF.
.
Currently, pt reports that he is anxious.
.
```

Figure 4: An example of a ground truth associated with the prompt 3



Ms. [\*\*Known lastname \*\*] is a 69 year old woman with a history of  
→ metastatic renal cell carcinoma (RCC) with distant metastases to the  
→ pancreas  
who presents with fever, chills, abdominal pain since [\*\*2119-1-7\*\*]. She  
→ reports taking 5 grams of tylenol daily at home  
which she also reports taking in the past week. She also reports having  
rib pain while lying in the [\*\*location unclear \*\*]. She reports chills  
and having dark stools, but denies any associated fever or rigors.  
She was recently started on zosyn for headache and nausea and  
vomiting. She denies any contact with sick person, recent travel  
or animals, or recent change in her work or home environment.

She is an anxious, non-smoker with a past medical history of pancreatic  
insufficiency and hypertension who presents with abdominal pain and  
lower extremity edema. She denies dysuria, joint pain or rash. She  
is not hypotensive with a BP of 125/80 at her arrival to the  
Emergency Department. She is asymptomatic for lactate at the time of  
this admission to the floor. On review of her baseline admission guaiac  
stool testing is not available. On her last outpatient visit to her  
oncologist at [\*\*Hospital1 \*\*] on [\*\*1-7\*\*], she was seen to have a stent  
→ placed for  
cholangitis and was given an ERCP. She was treated for cholangitis  
at the time and was given some ARF on [\*\*1-7\*\*].

Her oxygen saturation was 97% on 2L. As a work up she was given  
→ antibiotics  
for possible E.Coli pneumonia. She was found to have hypotension on  
admission to the floor and received dobutamine.

Figure 5: An example of synthetic data generated with Figure 3 with a SEMScore at 52

77 yo M w/ hx of metastatic RCC of the pancreas (pancreo-duodenal mass reports R/o chills since starting [\*\*6-2\*\*] on [\*\*2160-5-18\*\*] (day → 7-10 from  
Hx. He has had fevers with taking tylenol at that time (Ibuprofen 400mg, acetaminophen 650mg PRN for R/o rib pain and steal). On [\*\*2160-5-22\*\*] he was lying in bed and became extremely ill with chills. He reports this started a few minutes ago. Also reports dark stools. His fever is 102F (38.3C), he is on 2g IV zosyn (given IVF 2L NS, as he is V/5) for pneumonia with HC reported by the Wife. He reports severe headache. He is not compliant with PCP but has noticed many recent contacts with his S/M. He is very anxious for relief of pain. He has some dysuria and no joint pain. He has no rash. He is in the ED actively pacing. He is hypotensive. He is BP 90/45 with 3+ pti on his left hand. PRI all his vitals are stable otherwise at BP 92/44 HR 81 RR 14 96 O2 Sat 99% RA. On ABG: pH 7.31, PaCO2 28.1, PaO2 113. As a baseline (was done in ED), guaiac in stool was positive. He was sent to CT with ortho in ED. Saw his PCR [\*\*Last Name (Only) \*\*] [\*\*Date\*\*] [\*\*Time (only) \*\*] 5:15 and → showed a 6.2mm x 5.1mm pancreatic tail mass (blood in the head of the pancreas with no dilation distally). He is scheduled for pancreatic stent placement tomorrow. He was treated for acute cholangitis (e.g. given 3L NS and 1g of IVF) and was given 1L NS to help with ARF. He reports that he is more anxious for relief of pain.

Figure 6: An example of synthetic data generated with Figure 3 with a SEMScore at 79

# Building Certified Medical Chatbots: Overcoming Unstructured Data Limitations with Modular RAG

Leonardo Sanna\*, Patrizio Bellan\*, Simone Magnolini\*,  
Marina Segala\*, Saba Ghanbari Haez\*<sup>†</sup>, Monica Consolandi\*,  
Mauro Dragoni\*

\*Fondazione Bruno Kessler, Trento (ITALY)

[lsanna, pbellan, magnolini, msegala, sghanbarihaez, mconsolandi, dragoni]@fbk.eu

<sup>†</sup>Free University of Bozen, Bozen (ITALY)

## Abstract

Creating a certified conversational agent poses several issues. The need to manage fine-grained information delivery and the necessity to provide reliable medical information requires a notable effort, especially in dataset preparation. In this paper, we investigate the challenges of building a certified medical chatbot in Italian that provides information about pregnancy and early childhood. We show some negative initial results regarding the possibility of creating a certified conversational agent within the RASA framework starting from unstructured data. Finally, we propose a modular RAG model to implement a Large Language Model in a certified context, overcoming data limitations and enabling data collection on actual conversations.

**Keywords:** Conversational Agent, Digital Health, Retrieval-Augmented Generation

## 1. Introduction

In recent research, the demonstrated effectiveness of conversational agents and Large Language Models (LLMs) has expanded to include tasks that were once thought unlikely, marking a notable advancement in their capabilities. For instance, within the digital health area, it has been shown that conversational agents can provide emotional support to patients, possibly more efficiently than a standard interaction between a physician and a patient (Supadungsuk et al., 2023; Ayers et al., 2023; Fadhil and Gabrielli, 2017).

In this paper, we present the work-in-progress of a project to create a conversational agent capable of providing certified medical information regarding pregnancy and the first thousand days of a child's life. With the expression "*certified information*" we mean textual content generated or validated by healthcare professionals, ensuring its verifiability and alignment with the current scientific knowledge in the respective domain. In addition, an essential attribute of "*certified information*" is its predictability, indicating that, given a specific question the response would always be the same. The agent will be implemented initially in Italian only.

To the best of our knowledge, there are no examples in the literature where conversational agents have been employed to aid patients in this particular field. Likewise, there are no examples of an Italian medical conversational solution capable of delivering certified medical advice. Current applications of conversational agents within the healthcare industry suffer problems of data certification and accuracy (Srivastava and Singh, 2020; Jungmann et al., 2019; Swick, 2021); consequently, there is

a lack of evidence of their efficacy in clinical contexts (Bibault et al., 2019). Therefore, medical conversational agents are often limited to assisting medical staff rather than patients (Minutolo et al., 2022), or used as a tool to help diagnostics (Ni et al., 2017; Verma et al., 2022) and integrate the search for medical assistance (Soprano et al., 2023; Polignano et al., 2020). Also, the trust towards deploying this kind of technology is an aspect that needs to be addressed, as it directly impacts the potential efficacy (Seitz et al., 2022; Martens et al., 2024; Laumer et al., 2019). Creating a certified medical conversational agent would address some of these significant issues, especially when deploying these agents in the public sector.

In the following sections, we outline the main issues we have found in our workflow so far, summarize some text insights, and explore the possible solutions for the upcoming steps.

## 2. Dataset and Conversational Design

Our current corpus contains approximately 1300 texts sourced from verified medical channels <sup>1</sup>, focusing predominantly on *informational cards*. These cards offer brief yet detailed medical information on various topics, providing verified advice on conditions, treatments, and procedures. They are commonly used in FAQ sections, offering patients reliable information without direct interaction with healthcare professionals.

However, working with certified information

<sup>1</sup>The content is sourced from texts curated by the Obstetrician Department of the Hospital of Trento and from UPPA, a reputable child care website <https://www.uppa.it/>

poses challenges, particularly when adapting it for conversational use. Indeed, our dataset is not designed for integration into a conversational framework. One of the main challenges is that editing options are severely limited when dealing with certified medical information. The optimal approach would be to use the texts in their original form to preserve their certification. Yet, they often tend to be excessively lengthy and informationally dense for effective conversation use.

Moreover, we must consider that extracting information from these texts is complicated due to their highly discursive nature. Automatic segmentation often results in imprecise responses, occasionally leading to grammatical inaccuracies since segments are extracted from an existing discursive context. There is also a notable risk of encountering information gaps, despite the fact they are densely packed with information. In fact, in a certified context, all the deliverable information must be present explicitly in the text; even the simplest inferences are impossible since they would require certification, ensuring that they correspond to correct medical knowledge.

Lastly, our informational cards come from specialized sites and are meant to be instructive, so they often use medical vocabulary. This characteristic complicates the process of generating additional data, especially when generating questions for training a conversational agent. Medical jargon is indeed quite influential in affecting question generation, often leading to the creation of improbable examples.

While using an LLM could compensate for the lack of conversational data, our requirement to provide reliable information without any changes prevents us from directly using an LLM for user interaction. LLMs' erratic nature doesn't align with the need for stable and predictable output in certified information contexts.

### 3. Workflow: Creating a RASA Chatbot

We began with an existing COVID-19 FAQ chatbot (Lucianer et al., 2022) named *Covibot*. Since this agent was realized within the RASA framework<sup>2</sup>, we used RASA to create our first test conversational agent, focusing our efforts on the Natural Language Understanding (NLU) module, as its performance significantly impacts the overall conversation flow. This first experiment was therefore only focused on a simple classification pipeline, with the goal of associating each intent with a specific reply.

Using our data, we automated the generation of example questions with GPT-4 via the OpenAI

<sup>2</sup><https://github.com/RasaHQ/rasa>

ChatGPT API. We segmented the texts into shorter paragraphs using GPT-4 to generate the briefest meaningful paragraphs while considering the textual excerpt's topic. We then prompted the model to generate three simple questions for each text. These questions were then associated with specific intents linked to their corresponding answers.

Since RASA intent classifier<sup>3</sup> also supports custom word embeddings, we created a model (Le and Mikolov, 2014) from our data. While RASA supports various embedding techniques, support for highly specific domains, like ours, is limited<sup>4</sup>.

Our custom embedding model showed promising results in improving the conversational agent's performance in an initial sample of around 50 intents and 1500 total examples. Performance assessment was conducted by partitioning the dataset into 80% for training and 20% for testing, progressively increasing the number of examples during the training phase. In the graph shown in Figure 1, the *UPPA* configuration uses the embeddings of our dataset; the *Spacy* configuration uses pre-trained Spacy embeddings<sup>5</sup> for Italian, whereas the *Base* configuration uses no pre-trained embeddings.

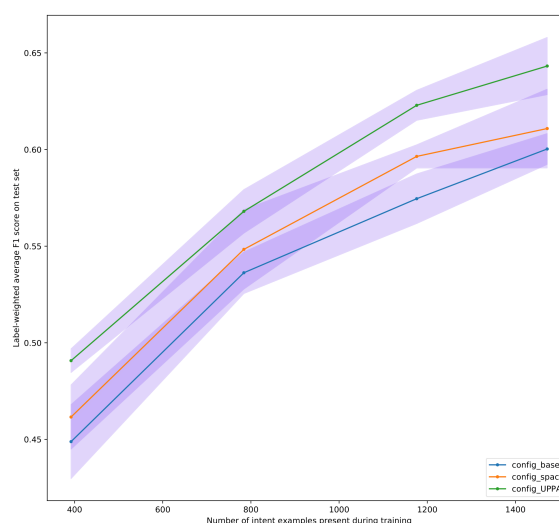


Figure 1: Comparison of custom word embedding impact on our first trained model.

Subsequently, we expanded our dataset to include 4500 intents and their corresponding answers. However, this dataset extension resulted in a noticeable decline in the RASA model's perfor-

<sup>3</sup><https://rasa.com/blog/introducing-dual-intent-and-entity-transformer-diet-state-of-the-art-performance-on-a-lightweight-architecture/>

<sup>4</sup>Support is limited to Gensim embeddings: <https://rasa.com/blog/custom-gensim-embeddings-in-rasa/>

<sup>5</sup><https://spacy.io/usage/models>

mance. This second evaluation assessed RASA's capacity for predicting the right intent class and, consequently, giving the right answer for each of the main topics in our dataset. Figure 2 illustrates the model's performance, which has been proven to be below acceptable standards.

Our RASA chatbot could classify correctly only an average of 28% of intents. Moreover, the model is quite sparse, with an average confidence on correct predictions of 0.27. Also, our custom embeddings lost their relevance in enhancing the training; the model proved indeed highly sensitive to minor rephrasing operations, where even a small alteration in a training sentence could easily cause the model to fail.

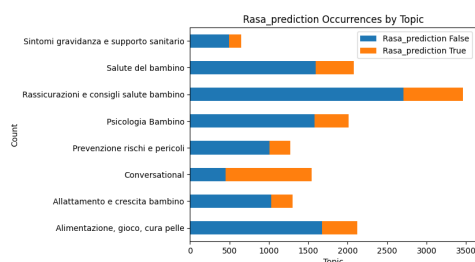


Figure 2: RASA performance across the main topics with 4500 intents. In orange, the correct replies.

#### 4. Data Limitations

Considering the outcome of the first test, some additional considerations on data quality are necessary. The data that we have is all unstructured text. These texts have great stylistic heterogeneity, even within the same source, combined with great semantic homogeneity, all being part of a specific medical domain. This dual characteristic makes topic modeling problematic; we have currently tried different types of approaches, ranging from the more classic Latent Dirichlet Allocation (LDA) (Blei et al., 2003), keywords (Bondi and Scott, 2010; Gabrielatos and Marchi, 2011), and BERTopic<sup>6</sup>, which has recently been shown as one of the most effective topic modeling techniques (Gan et al., 2023; Egger and Yu, 2022). Regardless of the method we used, we found that semantic areas in our data are always rather fragmented because of the great ramifications of sub-topics, even within the same thematic areas. For instance, in Figure 3 we show the topics found using BERTopic. The two main semantic macro-areas consist of one encompassing documents related to the newborn and another containing documents regarding pregnancy. Nevertheless, the extensive thematic fragmentation within these areas poses a significant challenge in

<sup>6</sup><https://doi.org/10.48550/arXiv.2203.05794>

training conversational agents to effectively associate intents with their respective topics.



Figure 3: Visualization of the topics found using BERTopic.

We would need fine-grained annotation on topics and other relevant linguistic aspects to effectively deliver certified information. Yet, since our semantic areas frequently overlap, automatic topic extraction does not produce qualitatively acceptable document groups. This means an in-depth qualitative analysis of the automatic topic extraction is required before annotation, also to highlight other elements like named entities and hardly quantifiable textual features (Hunston, 2004) such as relevant pragmatic aspects for medical conversations.

Moreover, having only unstructured texts is a substantial problem for RASA, since its intent classifier is designed to work with Named Entity Recognition. The existing state-of-the-art approaches such as MedBert (Egger and Yu, 2022) are also not focused on question answering nor entity recognition on unstructured texts like ours. Also, we have to consider that most of the approaches regarding medical conversational agents, especially for question answering (Kacupaj, 2022) have a knowledge-based approach (Dayal et al., 2023; Minutolo et al., 2017), which also requires annotated data.

#### 5. Future Work: Annotation and RAG

In our case, the data quality is a major issue that might have different solutions. Looking at previous approaches, it becomes evident that using certified sources in a conversational context, even a basic one, necessitates a considerable amount of contextual information (Kadariya et al., 2019; Fenza et al., 2023; Alloatti et al., 2021). Hence, developing an annotation methodology is essential to improve the

performance of the conversational agent, irrespective of the chosen framework. Certain information required for building our knowledge base can only be obtained through fine-grained annotation. However, this process proves to be time-consuming, and its success remains uncertain.

Alternatively, an immediately implementable strategy could involve using an LLM to address the discursive aspects, while incorporating certified sources from our database. LLMs, especially ChatGPT, have proven to be reasonably reliable, at least on basic questions about medical care (Mihalache et al., 2024; Cheong et al., 2023; Cascella et al., 2023). In addition to this, techniques such as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Karpukhin et al., 2020) can be used to increase the LLM’s ability to correctly answer a question, minimizing hallucinations (Martino et al., 2023). Essentially, the user’s request and the additional knowledge work together to guide the Language Model’s response. This prevents the model from giving inaccurate information when it does not have it readily available. However, as we said before, in a certified context we cannot rely on an LLM to provide the information to a patient, since it is impossible to certify the model output because of its stochastic nature.

Furthermore, a key issue in the standard RAG approach is the possible mismatch between the user’s query and the correct documents. Typically, RAG involves the transformation of a user query into a vector embedding representation, which is then used to assess semantic similarity among the repository of documents. However, the vector of the query and documents’ vectors might be significantly different within the semantic space; this discrepancy introduces a consequential constraint, as it may lead to the exclusion of relevant documents during the retrieval process.

**Modular RAG with HyDE** We are working within the Hypothetical Document Embeddings (HyDE) framework (Gao et al., 2023) to address these two limitations. HyDE is a novel approach recently introduced that operates unsupervised. In a nutshell, HyDE uses an LLM to produce a hypothetical document (HyDoc) based on input queries and then it uses the HyDoc to retrieve the information from the certified repository. Despite the hallucinations that might be present in the HyDoc, the generated text should lie in the semantic space in a neighborhood of similar real documents that contain the correct and certified answer to provide to the user.

In the pipeline that we are implementing, given a specific question, we generate a hypothetical document that is used to query the certified document repository. Then, the *paraphrase-multilingual-mpnet-base-v2* Bi-Encoder model (Reimers and

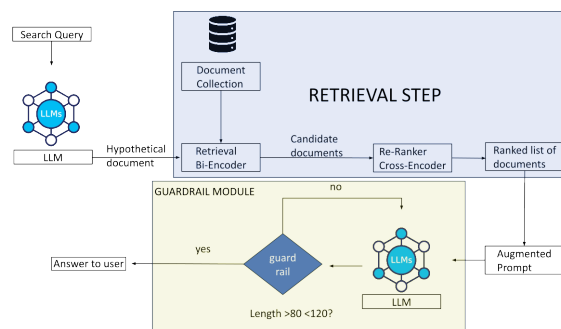


Figure 4: An overview of the RAG model we are implementing.

Gurevych, 2019) is used to retrieve the documents. However, the Bi-Encoder performs optimally when estimating similarity between documents of similar sizes. Given that our HyDoc and the certified documents may differ significantly in length, we use a cross-encoder, i.e. *ms-marco-MiniLM-L-6-v2*<sup>7</sup>, to re-rank the retrieved documents and refine the list. Finally, the selected documents are used to augment the initial prompt, and a *Guard-Rail* module<sup>8</sup> ensures that the LLM reply is short enough. As shown in Figure 4, the conversational agent’s final answer contains the documents’ textual summary (80-120 words) and the pointers to the original certified sources. Although our RAG model represents a compromise, it facilitates testing in a production environment, enabling data collection from authentic conversations and facilitating data augmentation.

Preliminary testing with GPT-4-turbo on 100 user-generated questions yielded promising results, retrieving relevant documents in over 85% of cases. On the same test set, the RASA model achieved only 13% correct answers, with approximately on-topic responses in 25% of cases and off-topic replies in over 60% of cases. In terms of HyDoc generation, GPT-4-turbo demonstrated the ability to produce pertinent responses in over 95% of examples. Given that the initial module impacts the entire model, additional investigation is required to assess open-source LLMs<sup>9</sup> performance, both in generating HyDocs and in the quality of document summarization.

## 6. Acknowledgments

This paper is part of the project TrustAlert which has received funding from the Fondazione Compagnia

<sup>7</sup><https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2>

<sup>8</sup><https://doi.org/10.48550/arXiv.2402.15911>

<sup>9</sup>For instance: <https://huggingface.co/swap-uniba/LLaMAntino-2-70b-hf-UltraChat-ITA>

San Paolo and Fondazione CDP under the “Artificial Intelligence” call. We acknowledge the support of the PNRR project INEST - Interconnected North-East Innovation Ecosystem (ECS00000043), under the NRRP MUR program funded by the NextGenerationEU. We acknowledge the support of the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

## 7. Bibliographical References

- Francesca Alloatti, Alessio Bosca, Luigi Di Caro, and Fabrizio Pieraccini. 2021. Diabetes and conversational agents: the aida project case study. *Discover Artificial Intelligence*, 1:1–21.
- JW Ayers, A Poliak, M Dredze, et al. 2023. [Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum](#). *JAMA Intern Med*, 183(6):589–596.
- Jean-Emmanuel Bibault, Benjamin Chaix, Pierre Nectoux, Arthur Pienkowski, Arthur Guillemasé, and Benoît Brouard. 2019. Healthcare ex machina: Are conversational agents ready for prime time in oncology? *Clinical and translational radiation oncology*, 16:55–59.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Marina Bondi and Mike Scott, editors. 2010. *Keyness in Texts*, volume 41. John Benjamins Publishing.
- Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 47(1):33.
- Ryan Chin Taw Cheong, Kenny Peter Pang, Samit Unadkat, Venkata Mcneillis, Andrew Williamson, Jonathan Joseph, Premjit Randhawa, Peter Andrews, and Vinidh Paleri. 2023. Performance of artificial intelligence chatbots in sleep medicine certification board exams: Chatgpt versus google bard. *European Archives of Otorhinolaryngology*, pages 1–7.
- Raghav Dayal, Parv Nangia, Surbhi Vijh, Sumit Kumar, Saurabh Agarwal, and Shivank Saxena. 2023. Development of chatbot retrieving fact-based information using knowledge graph. In *Proceedings of International Conference on Recent Innovations in Computing: ICRIC 2022, Volume 1*, pages 153–164. Springer.
- Roman Egger and Joanne Yu. 2022. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in sociology*, 7:886498.
- Ahmed Fadhil and Silvia Gabrielli. 2017. Addressing challenges in promoting healthy lifestyles: the ai-chatbot approach. In *Proceedings of the 11th EAI international conference on pervasive computing technologies for healthcare*, pages 261–265.
- Giuseppe Fenza, Francesco Orciuoli, Angela Peduto, and Alberto Postiglione. 2023. Healthcare conversational agents: Chatbot for improving patient-reported outcomes. In *International Conference on Advanced Information Networking and Applications*, pages 137–148. Springer.
- Costas Gabrielatos and Anna Marchi. 2011. Keyness: Matching metrics to definitions. In *Theoretical-methodological Challenges in Corpus Approaches to Discourse Studies and Some Ways of Addressing Them*.
- Lin Gan, Tao Yang, Yifan Huang, Boxiong Yang, Yami Yanwen Luo, Lui Wing Cheung Richard, and Dabo Guo. 2023. Experimental comparison of three topic modeling methods with lda, top2vec and bertopic. In *International Symposium on Artificial Intelligence and Robotics*, pages 376–391. Springer.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. [Precise zero-shot dense retrieval without relevance labels](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1762–1777. Association for Computational Linguistics.
- Susan Hunston. 2004. Counting the uncountable: Problems of identifying evaluation in a text and in a corpus. *Corpora and discourse*, 9:157–188.
- Stefanie Maria Jungmann, Timo Klan, Sebastian Kuhn, and Florian Jungmann. 2019. Accuracy of a chatbot (ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. *JMIR formative research*, 3(4):e13863.
- Endri Kacupaj. 2022. *Conversational Question Answering over Knowledge Graphs with Answer Verbalization*. Ph.D. thesis, Universitäts- und Landesbibliothek Bonn.

- Dipesh Kadariya, Revathy Venkataramanan, Hong Yung Yip, Maninder Kalra, Krishnaprasad Thirunarayanan, and Amit Sheth. 2019. kbot: knowledge-enabled personalized chatbot for asthma self-management. In *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 138–143. IEEE.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Sven Laumer, Christian Maier, and Fabian Tobias Gubler. 2019. [Chatbot acceptance in health-care: Explaining user adoption of conversational agents for disease diagnosis](#). In *Proceedings of the 27th European Conference on Information Systems (ECIS)*, Stockholm & Uppsala, Sweden.
- Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Marzia Lucianer, Francesca Perini, Giulia Malfatti, Lorenzo Gios, Alessandro Bacchiega, Claudio Giuliano, Andrea Nicolini, Stefano Forti, Roberta Corazza, Veronica Tretter, et al. 2022. A technology-enabled, public-driven, and multi-channel communication strategy during covid19 pandemic in the province of trento, italy. *JOURNAL OF MEDICAL INTERNET RESEARCH*.
- Marijn Martens, Ralf De Wolf, and Lieven De Marez. 2024. Trust in algorithmic decision-making systems in health: A comparison between ada health and ibm watson oncology. *Cyberpsychology*, 18(1).
- Ariana Martino, Michael Iannelli, and Coleen Truong. 2023. Knowledge injection to counter large language model (llm) hallucination. In *The Semantic Web: ESWC 2023 Satellite Events*, pages 182–185, Cham. Springer Nature Switzerland.
- Andrew Mihalache, Ryan S Huang, Marko M Popovic, and Rajeev H Muni. 2024. Chatgpt-4: an assessment of an upgraded artificial intelligence chatbot in the united states medical licensing examination. *Medical Teacher*, 46(3):366–372.
- Aniello Minutolo, Emanuele Damiano, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2022. A conversational agent for querying italian patient information leaflets and improving health literacy. *Computers in Biology and Medicine*, 141:105004.
- Aniello Minutolo, Massimo Esposito, and Giuseppe De Pietro. 2017. A conversational chatbot based on knowledge-graphs for factoid medical questions. In *SoMeT*, pages 139–152.
- Lin Ni, Chenhao Lu, Niu Liu, and Jiamou Liu. 2017. Mandy: Towards a smart primary care chatbot application. In *International symposium on knowledge and systems sciences*, pages 38–52. Springer.
- Marco Polignano, Fedelucio Narducci, Andrea Iovine, Cataldo Musto, Marco De Gemmis, and Giovanni Semeraro. 2020. Healthassistantbot: A personal health assistant for the italian language. *IEEE Access*, 8:107479–107497.
- Nils Reimers and Iryna Gurevych. 2019. [Sentencebert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.
- Lennart Seitz, Sigrid Bekmeier-Feuerhahn, and Krutika Gohil. 2022. Can we trust a chatbot like a physician? a qualitative study on understanding the emergence of trust toward diagnostic chatbots. *International Journal of Human-Computer Studies*, 165:102848.
- Michael Soprano, Kevin Roitero, Vincenzo Della Mea, Stefano Mizzaro, et al. 2023. Towards a conversational-based agent for health services. In *Proceedings of the Italia Intelligenza Artificiale-Thematic Workshops co-located with the 3rd CINI National Lab AIIS Conference on Artificial Intelligence (Ital IA 2023)*, pages 278–283.
- Prakhar Srivastava and Nishant Singh. 2020. Automated medical chatbot (medibot). In *2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)*, pages 351–354. IEEE.



S. Suppadungsuk, C. Thongprayoon, J. Miao, et al. 2023. Exploring the potential of chatbots in critical care nephrology. *Medicines*, 10:58.

Robert K Swick. 2021. The accuracy of artificial intelligence (ai) chatbots in telemedicine. *Journal of the South Carolina Academy of Science*, 19(2):17.

Shreya Verma, Mansi Singh, Ishita Tiwari, and BK Tripathy. 2022. An approach to medical diagnosis using smart chatbot. In *International Conference on Computational Intelligence in Pattern Recognition*, pages 43–56. Springer.

# Towards using Automatically Enhanced Knowledge Graphs to Aid Temporal Relation Extraction

Timotej Knez, Slavko Žitnik

University of Ljubljana, Faculty of Computer and Information Science  
Večna pot 113  
{timotej.knez, slavko.zitnik}@fri.uni-lj.si

## Abstract

Temporal relation extraction in medical document analysis is crucial for understanding patient histories and treatment outcomes. This paper introduces a novel approach leveraging a bimodal model integrating textual content and a knowledge graph to enhance temporal relation extraction. The paper presents ongoing research on constructing an optimal knowledge graph by augmenting PrimeKG with dynamically expanded information using a language model-generated knowledge graph. It also further personalizes the information with patient-specific graphs tailored for relation prediction. The pipeline for constructing this enriched knowledge graph is detailed, aiming to improve the capabilities of temporal relation extraction models. The preliminary results show that adding a simple knowledge graph to the temporal relation extraction model can significantly increase the performance, achieving new state-of-the-art results. While research on enhanced knowledge graphs is ongoing, this paper lays the groundwork for leveraging common knowledge to advance temporal relation extraction in medical contexts. This approach holds promise for enhancing the understanding of patient histories and treatment outcomes, potentially leading to improved healthcare decision-making and patient care.

**Keywords:** Information extraction, Knowledge graph building, Large language models

## 1. Introduction

In medical document analysis, extracting temporal relations is pivotal in enhancing our understanding of patient histories and treatment outcomes. In our preliminary research, we showed that we can improve temporal relation extraction by using a bimodal model that integrates information not only from the textual content of medical documents but also from a knowledge graph containing information about the patient's treatment from a discharge summary. We suggest further improving the results by introducing common knowledge from multiple sources to the model. Our experiments use general medical knowledge from a knowledge graph named PrimeKG (Chandak et al., 2023), providing a broad information foundation. As this is a broad human-curated graph, it does not contain information about all the concepts appearing in the analyzed documents. Therefore, augmenting the information and adding some common sense knowledge about the missing concepts is important. To address this, we use a knowledge graph automatically constructed using a large language model (Wang et al., 2023), allowing for dynamic expansion to include details about each entity in the dataset. Lastly, we introduce a patient-specific graph, individually tailored for each relation prediction. This graph contains other events from the document we are analyzing and the relations between them. Such information can give the relation prediction model temporal context about the events it analyzes. We test the proposed methods on the

i2b2 2012 dataset (Sun et al., 2013) containing medical discharge summaries.

## 2. Related Work

Temporal relation extraction aims to discern the temporal association between events within a text document, offering insights into the narrative behind the document. This paper focuses on extracting such relations from medical documents. Early approaches relied on rules (Gaizauskas et al., 2006; Dorr and Gaasterland, 2007) and traditional feature-based machine learning (Mani et al., 2006; Bethard, 2013; Chang et al., 2013), while recent years have seen a shift towards deep neural network-based models. Researchers have used two main architectures of deep neural networks. The first are long short-term memory networks (LSTM) exemplified by Tourille et al. (2017), Cheng and Miyao (2017), and Leeuwenberg and Moens (2018). Such networks build sentence embeddings by applying the LSTM layer over the tokens to generate token and sentence embeddings. They predict the relations based on the embeddings of the tokens corresponding to the events. Another approach is the use of pre-trained language models (PLM). Lin et al. (2019) mark events with special tokens and use a BERT network to encode the text into a vector representation. They classify the relation based on the sentence embedding that the PLM produces. Zhou et al. (2021) use a similar approach but enhance the results with soft logic regularization.

## 2.1. Use of Common Knowledge

One of the improvements for temporal relation extraction that has been proposed recently is the introduction of common sense knowledge. Such knowledge can help the model in cases where relations are not explicitly mentioned in the text, as it can be used to reason about how events might be related. Ning et al. (2018) developed a statistical resource for temporal relation extraction, which includes statistics about common relations between events. These statistics are then used in their model to improve predictions. They managed to improve the prediction accuracy of their model by 3% when introducing the resource (Ning et al., 2019). The idea was later expanded by Han et al. (2020). They enhance predictions using domain knowledge, which includes statistics, event types, and structured constraints based on the relationships between events.

Some researchers propose using knowledge graphs as an additional source of common knowledge to help with tasks similar to temporal relation extraction. Lin et al. (2023) propose a model for extracting disease relations from text using a bimodal architecture where a SciBERT model encodes text. In contrast, a heterogeneous graph attention network encodes additional information from a knowledge graph. They make the final relation prediction based on both sets of encodings. Similarly, Yasunaga et al. (2022) propose a pre-trained DRAGON model, which simultaneously encodes and combines text and graph information. Such a network can be used to improve language processing using a knowledge graph.

## 2.2. Automatic Knowledge Graph Construction

When using knowledge graphs as a source of common knowledge to help information extraction models, an important problem that needs to be addressed is where to get a knowledge graph that contains all the necessary information. As the abilities of large language models to express common sense advanced, the idea of automatically constructing knowledge graphs was presented. Jiang et al. (2023) propose to use ChatGPT to generate knowledge graph triplets about drugs, conditions, and procedures. Such an approach enables the creation of specialized knowledge graphs containing much information about the relevant concepts.

## 3. Using Common Knowledge in Temporal Relation Extraction

Significant advancements in natural language processing (NLP) tasks have been observed in recent years, driven by the introduction of large pre-

trained language models. These models achieve high performance due to their extensive training on vast datasets, providing them with common sense knowledge. However, employing such models has drawbacks, including the need for large computational resources and lengthy training times. We propose that equipping models with specialized common sense knowledge tailored to specific tasks could yield comparable or superior performance using simpler models. To realize this concept, we present a bimodal model for temporal relation extraction. This model operates on two inputs: the text describing events and a knowledge graph offering general knowledge about these events. By encoding general knowledge in the knowledge graph, we provide essential information aiding the model in determining temporal relations. This knowledge inclusion is crucial, especially in medical contexts, where predictions often demand specialized knowledge that is not adequately represented in general datasets. Furthermore, predictions in the medical domain may rely on patient-specific histories, which must be incorporated at inference time.

Our preliminary experiments (Knez and Žitnik, 2024) show that introducing general knowledge to such a model can improve temporal relation prediction results. We used a knowledge graph containing automatically predicted temporal relations between other events in a document. Such relations give the model temporal context about the events we are observing. In our tests, the model managed to achieve new state-of-the-art results on the i2b2 2012 temporal relation extraction dataset (Sun et al., 2013) using this method (see Table 1). Based on this result, we believe that introducing additional general knowledge would improve the results even further.

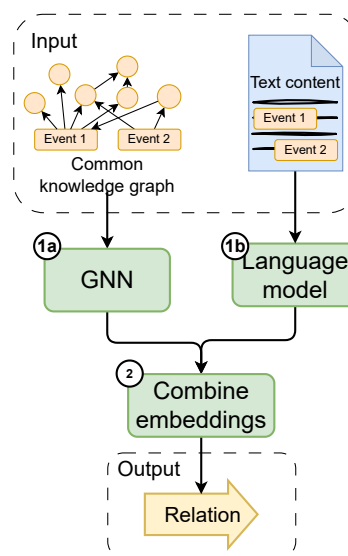


Figure 1: The architecture of the temporal relation extraction model.

### 3.1. Model Architecture

The temporal relation extraction model comprises two event encoders labeled 1a and 1b in Figure 1. The graph encoder (1a) utilizes a convolutional graph neural network to aggregate information from the local knowledge graph centered around a node representing an event from the text, thus computing an event embedding. Similarly, the text encoder (1b) employs the EntityBERT model to generate embeddings for tokens corresponding to the events in the text, capturing sentence meaning. In the final phase of the model (2 in Figure 1), both types of embeddings are merged to compute the final temporal relation prediction.

## 4. Building Knowledge Graphs

The proposed approach’s main advantage is the model’s ability to use common knowledge from a knowledge graph. We must determine how to get the knowledge graph to include such knowledge. In our research, we combine three separate kinds of knowledge graphs to provide as much relevant information as possible, shown in Figure 2. While PrimeKG and automatically generated graphs contain general relations between medical concepts, the patient-specific temporal knowledge graph contains temporal relations that determine in what order the medical events occurred to the patient. We can combine the graphs by linking events from the text to all three knowledge graphs.

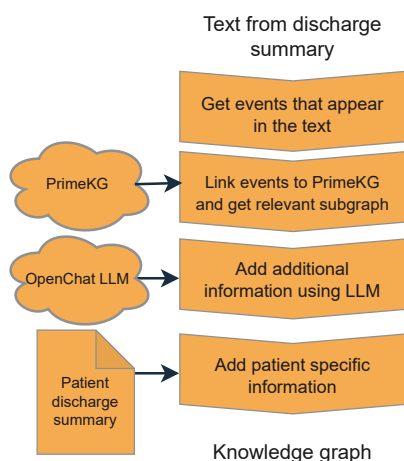


Figure 2: The proposed process of constructing a knowledge graph.

### 4.1. PrimeKG Knowledge Graph

Our first source of common knowledge is PrimeKG, a knowledge graph developed by Chandak et al. (2023). PrimeKG focuses on precision medicine

and integrates data from 14 sources into a structured graph format. To leverage PrimeKG, we employ the SciSpacy library to identify UMLS concepts within event mentions. 85% of the nodes in PrimeKG are linked to their corresponding UMLS concepts. We utilize these connections to determine the PrimeKG node that best represents the medical event mentioned in the text using the UMLS identifiers that the SciSpacy library provides. We use the most likely identifier that appears in the PrimeKG knowledge graph. Subsequently, we extract the neighborhood of this node from PrimeKG and incorporate it as a subgraph in our model.

### 4.2. Automatically Generated Knowledge Graph

A problem with using the PrimeKG as a source of common knowledge is that a limited number of medical concepts are present in such graphs. We observed that 78% of the events described in clinical discharge summaries do not correspond to any node in PrimeKG. We propose generating larger common knowledge graphs using large language models to improve that. Our method is based on the one used by Jiang et al. (2023).

We used the OpenChat pre-trained large language model (Wang et al., 2023), which we prompt to generate knowledge graph triplets related to the concept in question. We generate prompts detailing that the model should prepare as many relations as possible about a medical concept in the format [entity 1, relation, entity 2]. We found that the model performs better if we include a small number of examples of responses that we expect in the prompt. In this way, we were able to create a knowledge graph that contains all of the events that occur in the dataset. While generating such a graph requires a large amount of resources, making it unpractical for some applications, it is possible to generate the graph in advance and use it at inference time.

### 4.3. Patient-Specific Knowledge Graph

In understanding medical notes concerning patients, their medical history holds significant importance. We aim to incorporate this information into our knowledge graph by automatically generating an additional graph based on patient data extracted from clinical discharge summaries. Utilizing a discharge summary as a base, our model predicts relations between different events within the document. These predicted relations serve as the basis for constructing a knowledge graph, providing temporal context for the events under consideration. This approach enhances our model’s ability to comprehend the chronological sequence of events. We construct the final graph to be used by the model by combining nodes from all three graphs.

Table 1: Comparison of temporal relation extraction models on the i2b2 dataset without any common sense knowledge graph to the model with information from a patient-specific knowledge graph.

Model	Text
EntityBERT encoder	73.11%
EntityBERT + patient-specific KG	<b>74.16%</b>
BERT base (Ul Haq et al., 2022)	72.41%
BioBERT (Ul Haq et al., 2022)	73.60%
Alpaca model	35.01%

## 5. Results

We tested our approaches on the i2b2 2012 (Sun et al., 2013) dataset for temporal relation extraction from clinical discharge summaries. The dataset contains 14,256 training relations and 11,752 test relations.

### 5.1. Improving Temporal Relation Prediction

When predicting a temporal relation in our preliminary tests, we used knowledge graphs containing predicted temporal relations between other events to give the model some temporal context as described in Section 4.3. While constructing such graphs increases the computational requirements of the model, it shows how the prediction accuracy can be improved using additional knowledge.

The F1 scores of the predictions in our experiments are shown in Table 1. We can see that adding a patient-specific knowledge graph improves prediction results by 1 pt. We rerun the training 20 times and found that the improvement is statistically significant, with a P-value of 0.039. We compare our model to two models proposed by Ul Haq et al. (2022) and a prompt-based approach using the Alpaca large language model. Our proposed bimodal model surpasses other state-of-the-art models on the same dataset.

### 5.2. Knowledge Graph Construction

Table 2: Comparing knowledge graphs based on PrimeKG with knowledge graphs generated using large language models.

	PrimeKG	LLM
i2b2 events present	22%	100%
Average number of nodes	133.7	27.6
Average node degree	2.00	1.78

We analyzed the graphs constructed using the procedure described in Section 4. We found that only 22% of the concepts from the i2b2 dataset are successfully linked to PrimeKG. This happens

because we need a very broad knowledge graph that contains concepts from a large variety of areas. As a result, such a knowledge graph cannot contain all of the concepts from each area. This represents a large problem for our relation extraction model, as it has no common sense information for most of the events it encounters in a clinical document. For this reason, we enrich the knowledge graph using nodes and relations generated using a large language model.

Using a large language model to construct the knowledge graphs, we generated a knowledge graph about each concept from the i2b2 dataset. While generated graphs can contain errors as humans did not curate them, we believe they can be much more useful as they contain information about each concept.

We compare the knowledge graphs extracted from PrimeKG to those automatically generated using a large language model in Table 2. We found that the subgraphs we extracted from PrimeKG are generally larger than those created by a large language model. While a larger graph is generally beneficial, the graphs created using LLMs might be more useful for helping a machine learning model as they contain only the highly relevant information. The average degree of nodes in both graphs is quite similar at around two. Based on the analysis results, we recommend using a manually curated knowledge graph like PrimeKG enriched using automatically determined relations.

## 6. Conclusion

In our study, we demonstrated that adding additional knowledge to the model for temporal relations extraction can improve its performance, allowing it to achieve state-of-the-art results. While we only tested the performance gains when using knowledge graphs containing information from the active document, we believe including additional information could improve the results even further. When constructing knowledge graphs, we propose to enrich information from a large knowledge graph like PrimeKG by automatically generating relations using a large language model. Our results show that adding such relations greatly increases the coverage of a knowledge graph for the relation extraction task. The research is ongoing, and we expect to evaluate the model’s performance using the proposed knowledge graph in future work.

## 7. Acknowledgements

This work was supported by Slovenian Research Agency grant number H017002 (Young Researchers grant).

## 8. Bibliographical References

- Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second joint conference on lexical and computational semantics (\*SEM), volume 2: proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, pages 10–14.
- Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. [Building a knowledge graph to enable precision medicine](#). *Scientific Data*, 10(1):67.
- Yung-Chun Chang, Hong-Jie Dai, Johnny Chi-Yang Wu, Jian-Ming Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2013. Tempting system: a hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. *Journal of Biomedical Informatics*, 46:S54–S62.
- Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional lstm over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6.
- Bonnie J Dorr and Terry Gaasterland. 2007. Exploiting aspectual features and connecting words for summarization-inspired temporal-relation extraction. *Information processing & management*, 43(6):1681–1704.
- Rob Gaizauskas, Henk Harkema, Mark Hepple, and Andrea Setzer. 2006. Task-oriented extraction of temporal information: The case of clinical narratives. In *Thirteenth International Symposium On Temporal Representation And Reasoning (time'06)*, pages 188–195. IEEE.
- Rujun Han, Yichao Zhou, and Nanyun Peng. 2020. Domain knowledge empowered structured neural net for end-to-end event temporal relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5717–5729.
- Pengcheng Jiang, Cao Xiao, Adam Cross, and Jimeng Sun. 2023. Graphcare: Enhancing healthcare predictions with open-world personalized knowledge graphs. *arXiv preprint arXiv:2305.12788*.
- Timotej Knez and Slavko Žitnik. 2024. [Multi-modal learning for temporal relation extraction in clinical texts](#). *Journal of the American Medical Informatics Association*.
- Artuur Leeuwenberg and Marie Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1237–1246.
- Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A bert-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71.
- Yucong Lin, Keming Lu, Sheng Yu, Tianxi Cai, and Marinka Zitnik. 2023. Multimodal learning on graphs for disease relation extraction. *Journal of Biomedical Informatics*, page 104415.
- Inderjeet Mani, Marc Verhagen, Ben Wellner, Chungmin Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760.
- Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6203–6209.
- Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018. Improving temporal relation extraction with a globally acquired statistical resource. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 841–851.
- Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.
- Julien Tourille, Olivier Ferret, Aurelie Neveol, and Xavier Tannier. 2017. Neural architecture for temporal relation extraction: A bi-lstm approach for detecting narrative containers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–230.
- H UI Haq, V Kocaman, D Talby, APB Veyseh, F Der-noncourt, TH Nguyen, and VD Lai. 2022. Deeper clinical document understanding using relation extraction. In *Proceedings of the workshop on scientific document understanding*. CEUR.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xi-  
angang Li, Sen Song, and Yang Liu. 2023.  
Openchat: Advancing open-source language  
models with mixed-quality data. *arXiv preprint  
arXiv:2309.11235*.

Michihiro Yasunaga, Antoine Bosselut, Hongyu  
Ren, Xikun Zhang, Christopher D Manning,  
Percy S Liang, and Jure Leskovec. 2022. Deep  
bidirectional language-knowledge graph pretrain-  
ing. *Advances in Neural Information Processing  
Systems*, 35:37309–37323.

Yichao Zhou, Yu Yan, Rujun Han, Harry J Caufield,  
Kai-Wei Chang, Yizhou Sun, Peipei Ping, and  
Wei Wang. 2021. Clinical temporal relation ex-  
traction with probabilistic soft logic regularization  
and global inference. In *Proceedings of the AAAI  
Conference on Artificial Intelligence (AAAI'21)*.

# Experiments in Automated Generation of Discharge Summaries in Italian

Lorenzo Ruinelli<sup>1,2</sup>, Amos Colombo<sup>1,2</sup>, Mathilde Rochat<sup>3,4</sup>,  
Sotirios Georgios Popeskou<sup>4,5</sup>, Andrea Franchini<sup>6</sup>, Sandra Mitrović<sup>6</sup>,  
Oscar Lithgow<sup>6</sup>, Joseph Cornelius<sup>6</sup>, Fabio Rinaldi<sup>6</sup>

<sup>1</sup> Team Innovation & Research, Area ICT,

Ente Ospedaliero Cantonale (EOC), Bellinzona, Switzerland

<sup>2</sup> Clinical Trial Unit, Ente Ospedaliero Cantonale (EOC), Bellinzona, Switzerland

<sup>3</sup> Servizio di Medicina Interna, Ospedale Regionale di Lugano,  
Ente Ospedaliero Cantonale (EOC), Bellinzona, Switzerland

<sup>4</sup> Facoltà di Scienze Biomediche, Università della Svizzera Italiana (USI), Lugano, Switzerland

<sup>5</sup> Dipartimento di Chirurgia Viscerale, Ospedale Regionale di Lugano,  
Ente Ospedaliero Cantonale (EOC), Bellinzona, Switzerland

<sup>6</sup> Dalle Molle Institute for AI (IDSIA - USI/SUPSI), Lugano, Switzerland

{lorenzo.ruinelli,amos.colombo,mathilde.rochat,sotiriosgeorgios.popeskou}@eoc.ch

{andrea.franchini,sandra.mitrovic,oscarwilliam.lithgow,joseph.cornelius,fabio.rinaldi}@idsia.ch

## Abstract

Hospital discharge letters are a fundamental component of patient management, as they provide the crucial information needed for patient post-hospital care. However their creation is very demanding and resource intensive, as it requires consultation of several reports documenting the patient's journey throughout their hospital stay. Given the increasing pressures on doctor's time, tools that can draft a reasonable discharge summary, to be then reviewed and finalized by the experts, would be welcome. In this paper we present a comparative study exploring the possibility of automatic generation of discharge summaries within the context of an hospital in an Italian-speaking region and discuss quantitative and qualitative results. Despite some shortcomings, the obtained results show that a generic generative system such as ChatGPT is capable of producing discharge summaries which are relatively close to the human generated ones, even in Italian.

**Keywords:** Generating Discharge Summaries, LLMs, NLP

## 1. Introduction

The management of an hospitalization foresees the preparation of a Discharge Letter (DL) to summarize important information about the patient's diagnosis, treatment, medications, follow-up care, and any additional instructions or recommendations for the patient's ongoing health management. The primary goal of a DL is to convey critical information regarding a patient's care and treatment throughout their hospitalization to their general practitioner or primary care provider. The redaction of DLs is a resource-intensive process, both for the caretaker and the hospital (Golder et al., 2011; Cocco, 2012). The process often involves junior physicians who initially compose the first draft, which is then reviewed and validated by senior physicians before finalization. Physicians incur high risks of burnout (Hartman et al., 2023), which has been correlated to the bureaucratic tasks involved in their daily activities (Reith, 2018). While certain sections of the letter necessitate straightforward data extraction from the clinical records, others call for the capacity to distill and summarize complex clinical notes effectively. To fully or partially automate this process

would imply a reduction in the time investment from the physician (Reith, 2018).

This paper explores the potential of large language models (LLMs) in enhancing the summarization of clinical records, written in Italian. In particular, we present an experiment aimed at validating the effectiveness of utilizing LLMs for supporting the summarizing of clinical diaries to be integrated into the discharge letter. The experiment is grounded in real-world clinical diaries correlated with their associated discharge letters, which are provided by our partner hospital. The evaluation process involves expert knowledge assessment and similarity-based metrics, with the aim of comparing the quality of the summaries generated by the LLM against the manually generated summaries (i.e. the DL).

## 2. Related Work

The interest and relevance of the task of automated generation of discharge summaries is shown by several publications and initiatives such as the BioNLP ACL'24 Shared Task on Streamlining Discharge Documentation (Xu, 2024). The generation of discharge summaries specifically tailored



to the needs of the patient, aiming to maximize readability and understandability without sacrificing correctness, is discussed by (Zaretsky et al., 2024; Eppler et al., 2023). Other projects, as in (Ando et al., 2022; Hartman et al., 2023), research better strategies for summarizing structured or unstructured medical notes while still maintaining the domain's expert terminology, akin to our own goal. Given the recent advancements in transformer-based architecture and their performance in text summarization, recent studies almost exclusively rely on transformer-based neural network architecture for their experiments, such as (Ando et al., 2022) with BERT, (Hartman et al., 2023) with BERT and BART. Studies such as (Zaretsky et al., 2024; Eppler et al., 2023) approach the problem of text summarization through the use of readily available LLMs, specifically GPT-4.0. The problem then revolves around enhancing the language generation model by providing instructions to the LLM about the task, also known as prompt-engineering. The latter is a heuristic process highly specific to its target model. The use of more tailored prompts in these studies has shown measurable improvements in most metrics.

The typical evaluation strategies we find in the literature often involve the following metrics: ROUGE (Lin and Hovy, 2003), BLEU (Papineni et al., 2002), BertScore (Zhang et al., 2019), BLEURT (Sellam et al., 2020) and MoverScore (Zhao et al., 2019), which score the similarity between documents, usually between the reference, written by a physician, and the generated one. Some studies employ ROUGE and BLEU (Ando et al., 2022; Hartman et al., 2023), now considered less sophisticated than their neural network alternatives, which offer a more human-like judgment. Neural network-based metrics usually consider semantic and contextual information, thus providing more reliable insight into the generated text when comparing it against the reference, as employed by (Ando et al., 2022). Some authors, as (Hartman et al., 2023; Zaretsky et al., 2024; Eppler et al., 2023) supplement their evaluations by involving one or more domain's experts to review the generated document and provide a correctness measure based on human judgment.

Related literature involving the usage of LLMs in the medical context with Italian language seems to be quite restricted, studying for example the capacity of LLMs (including ChatGPT-3.5 and ChatGPT-4) to answer the questions and provide templates related to structured reports in radiology (Mallio et al., 2023). Another study investigated ChatGPT potential in generating and annotating goal-oriented dialogues, and used as one of the use cases a scenario when doctor needs to explain the diagnosis and treatment to a patient (Labruna et al., 2023).

In (Montagna et al., 2023), a comprehensive framework for creating an LLM-based chatbot system that assists chronic patients is introduced. To the best of our knowledge, this is the first study focusing on discharge letters/summaries in Italian.

### 3. Methodology

We screened the hospital database and collected both discharge letters and the corresponding clinical notes utilized in their composition. Clinical notes are written by nurses and doctors during the patient's stay, describing the current status of the patient and the future steps in the patients care. The timeframe was restricted to a recent six-month period. The language of discharge letters and corresponding clinical notes is Italian. Our focus was on simple cases, defined as clinical notes with a character length ranging between 3400 and 4000. This character length was chosen in order to not exceed the ChatGPT character limit and is close to the mean length of the clinical notes. Additionally, we targeted two medical specialties: surgery and medicine, sampling 30 cases from each group. Discharge letters from medicine cases tend to be more complex in nature compared to those of surgical cases. Clinical notes and discharge letters were de-identified using an internally developed tool capable of removing patient names and ages, contacts, locations and organizations. We produced two summaries for each case: one utilizing ChatGPT-3.5 (denoted as AI<sub>3.5</sub>) and the other using ChatGPT-4 (denoted as AI<sub>4</sub>). As the purpose of this experiment was only to test the feasibility of the idea, we used a prompt composed by a simple request (*"Crea un riassunto del seguente decorso clinico da includere nella lettera di uscita"*)<sup>1</sup> followed by the clinical notes in JSON format<sup>2</sup>. Each case had then 4 documents: the prompt including the clinical notes (P), the summary written by medical doctors (MD), and the two summaries generated by ChatGPT (AI<sub>3.5</sub> and AI<sub>4</sub>).

The similarities between each document pairs were evaluated the using following metrics:

- ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Hovy, 2003), a recall-oriented metric based on longest shared common subsequence in the documents.
- BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), a precision-based metric quantifying the overlap of n-grams between the documents

<sup>1</sup>In English: "Create a summary of the following clinical discourse to include in the discharge letter"

<sup>2</sup>An example is provided in the Appendix A.

- BERTscore (Zhang et al., 2019), which harnesses contextual embeddings from BERT to compute the similarity between the documents
- BLEURT (Sellam et al., 2020), a BERT-based text similarity evaluation metric modeled to mimic human judgment and optimized for generality. It is designed to compare sentences, so it might not be effective on entire documents.

We use ROUGE-L as the chosen ROUGE metric computed with the rouge\_score python library. The BLEURT score was computed using the code from the official BLEURT GitHub repository. The standard scorer uses the BERT-Tiny<sup>3</sup> model.

Additionally, we conducted an expert evaluation of the ChatGPT-generated summaries with the assistance of two medical doctors.

## 4. Results

Metric	Medicine		Surgery	
	$AI_{3.5}$	$AI_4$	$AI_{3.5}$	$AI_4$
BERTscore	0.8890	0.9050	0.8900	0.8950
BLEU	0.0003	0.0002	0.0006	0.0006
BLEURT	-0.3710	-0.3740	-0.3420	-0.3240
ROUGE	0.1130	0.1280	0.1370	0.1360

Table 1: Comparison of DLs generated by doctors (MD) versus AI generated ones ( $AI_{3.5}$  and  $AI_4$ ) in the Medicine & Surgery specialties (30 cases each)

Table 1 illustrates the average of the numerical comparison for the 30 samples in each of the two specialties. It shows that ChatGPT-4 generally outperforms ChatGPT-3.5 across all metrics except for BLEU. Specifically, ChatGPT-4 has higher BERTscore and ROUGE scores for general medicine, indicating better semantic similarity and n-gram overlap with reference texts. Additionally, ChatGPT-4 achieves better BERTscore for surgery than ChatGPT-3.5. BLEU scores are very low for both models, but slightly better for ChatGPT-3.5 in one instance and equal in another. BLEURT scores, while negative for both, are slightly higher for ChatGPT-4, suggesting a slight improvement in semantic quality. Overall, ChatGPT-4 demonstrates a marginal but consistent improvement in text generation quality over ChatGPT-3.5.

The qualitative evaluation was conducted with the collaboration of two medical doctors, one specialized in Medicine and the other in Surgery. Each doctor evaluated cases from their respective specialty. Both doctors found the AI-generated summaries well-done and potentially useful, expressing a preference for those generated by ChatGPT-4

<sup>3</sup><https://github.com/google-research/bert>

over ChatGPT-3.5. The following section aims to illustrate the problems that have been identified, using one case for each of the two specialties. Figure 1 refers to a general medicine case, while Figure 2 refers to a surgery case. Both figures show on top the original human-generated summary, and on the bottom the summary generated using GPT-4.

In the medicine case (Figure 1), we observe that a significant portion of the medical doctor’s summary (highlighted in yellow) reports information that was not present in the clinical notes processed by the AI models. This is because this information comes from the notes collected in the emergency room, which were not used in our experiment. The AI-generated summary begins by stating when the patient was discharged (see 1 in Figure 1)<sup>4</sup>. While this information is factually correct, it does not follow the typical style of discharge letters, which typically do not begin in this manner. The AI-generated summaries also included a series of stay-related information that are not relevant in this discharge letter, namely: fever episodes (2), infusion treatment (3), addressing hypokalemia (6), conducting regular laboratory tests (8), planning the return home after the hospital stay (9), and treating with Sintrom due to INR values (10). The AI-generated summary includes a sentence stating that the patient was treated with azithromycin because of a positive result on the Legionella test (4). However, this is not entirely accurate, as the treatment decision was based on the positive test result, along with the patient’s medical history and other diagnostic investigations. The AI models incorrectly interpreted the Italian acronym for vital parameters (PV) as venous pressure (5).

In the surgery case (Figure 2), we notice that the yellow-highlighted portion is smaller compared to the medicine case. This indicates that the notes used in our experiment include a larger portion of the necessary information. The AI models omitted two important pieces of information: that the patient had an intraductal papillary mucinous neoplasm (IPMN), as well as the result of the cholangiography. Similarly to the medicine case, the AI models included a series of stay-related information that are not relevant in a discharge letter: comprehensive blood tests (see 2 in Figure 2), a Cholangiography performed and report pending (3, 5), fasting blood glucose test (4), and a summary of the patient’s status (7). The AI-generated summary also included a sentence (6) that is not entirely correct from a clinical perspective, and it also hallucinated about colestasi (1).

<sup>4</sup>Notice that specific text segments in the figures are identified by a superscript, which we use from here on to refer to them.

<p>MD Trattasi di un paziente di XX anni, noto per le diagnosi sopracitate, in particolare per cardiopatia ischemica e uno stato post PTA-stenting dell'arteria femorale superficiale e poplitea destra, che giunge in data DD.MM.YYYY presso il Pronto Soccorso OCL inviato da medico curante per tosse e febbre persistenti da una settimana con associato rialzo degli indici di flogosi all'esame ematochimico. Ci confrontiamo con un paziente vigile, cosciente, ed orientato nei 3 domini, normoteso (PA 132/73 mmHg), normocardico (FC 79 bpm), eupnoico in aria ambiente (SpO2 95%) e apiretico (T 37.3°C). All'obiettività cardiaca si apprezzano toni validi e ritmici con pause libere. Gli arti inferiori sono ben perfusi e non edematosi. Al torace si nota un murmure vescicolare ubiquitario con crepitii medio basali a destra. L'addome è trattabile, non dolente alla palpazione superficiale e profonda, con peristalsi normoviva e segni di Murphy e Blumberg negativi. L'esame neurologico cursorio risulta nella norma. Gli esami ematochimici documentano un rialzo dei parametri infiammatori (PCR 162 mg/L) con leucocitosi neutrofila e una deflessione degli indici di funzionalità renale (creatinina 117 μmol/L, GFR 58 mL/min/1.74m<sup>2</sup>, urea 9 mmol/L). Lo striscio nasofaringeo per Sars-CoV-2/Influenza/RSV è negativo. Gli antigeni urinari risultano positivi per Legionella e negativi per Pneumococco. Il sedimento urinario non è patologico. L'ECG traccia un ritmo sinusale alla frequenza di 78 bpm, con deviazione assiale sinistra e senza alterazioni della ripolarizzazione. La radiografia del torace mostra un addensamento parenchimale polmonare al campo medio di destra di aspetto flogistico. In considerazione dei dati anamnestici, della clinica e dell'esito delle indagini paracliniche effettuate, il quadro depone per una polmonite destra a Legionella. Dal punto di vista terapeutico si instaura un'antibioterapia con azitromicina e si ricovera il signor XXXX nel reparto di Medicina Interna per prosecuzione delle cure. Durante la degenza, si nota un progressivo miglioramento della clinica a respiratoria con svezzamento dall'ossigenoterapia, in paziente emodinamicamente stabile ed afebrile. Dato il decorso clinico e laboratoristico favorevole, possiamo dimettere in data DD.MM.YYYY il signor XXXX a domicilio, in condizioni cliniche stabili.</p> <p>Parametri vitali alla dimissione : TAS/TAD 128/62 mmHg, FC 74 bpm, SaO2 93%, T 36.5°C, Peso 81.8 kg</p> <p>Not in P Omitted</p>
<p>AI4 Il paziente è stato dimesso il DD MON YYYY alle ore HH:MM<sup>1</sup>, dopo un ricovero iniziato il DD MON YYYY per polmonite ad acquisizione comunitaria con insufficienza renale acuta di grado I su base pre- renale causata da disidratazione. Durante il ricovero, il paziente ha presentato episodi febbrili, raggiungendo una temperatura di 38.8°C<sup>2</sup>, e pressione arteriosa ipotensiva gestita con somministrazione di liquidi<sup>3</sup>. È stato trattato con azitromicina per 10 giorni a causa di una positività all'antigene di Legionella<sup>4</sup>, mantenendo una buona saturazione di ossigeno senza necessità di ossigenoterapia dopo i primi giorni. Si sono osservati valori stabili di pressione venosa (PV)<sup>5</sup> e assenza di picchi febbrili verso il termine del ricovero. È stata inoltre gestita un'ipokaliemia<sup>6</sup>. Il paziente è stato monitorato per la febbre<sup>7</sup> e sottoposto a controlli laboratoristici regolari<sup>8</sup>, con un piano di rientro al domicilio senza aiuti previsto per l'inizio della settimana successiva alla dimissione<sup>9</sup>. Durante il soggiorno è stata effettuata anche una dichiarazione di Legionella e adattamenti del trattamento con Sintrom in base ai valori di INR<sup>10</sup>.</p> <p>Wrong discharge letter style Not relevant in a discharge letter Not entirely correct Hallucination</p>

Figure 1: Comparison of the MD summary with the AI-generated summary for a medicine case. Color coding: yellow - information not present in clinical notes and consequently neither in the prompt; grey - information provided in the prompt but omitted in the generated summary; olive - wrong discharge letter style; cyan - information not relevant in a discharge letter; magenta - information not entirely correct; red - hallucination.

<p>MD Il bilancio radiologico eseguito dai colleghi in Pronto Soccorso evidenzia colecisti sovradistesa, con sludge biliare e ispessimento di parete e falda fluida periviscerale, in assenza di dilatazione delle vie biliari intra- od extraepatiche, in quadro compatibile con colecistite acuta. Gli esami emato-chimici eseguiti presso l'ambulatorio del medico curante documentano sindrome infiammatoria, iperbilirubinemia e rialzo degli indici di colestasi. Completiamo le indagini con esecuzione di colangio-RM che conferma la diagnosi di colecistite acuta litiasica senza immagini riferibili a coledocolitiasi. Come reperto occasionale si segnala verosimile IPMN di tipo 2 e cisti ranel sinistra (Bosniak II). Si pone indicazione ad esecuzione di intervento di colecistectomia, che si svolge in assenza di complicanze mediche o chirurgiche da segnalare. Alla colangiografia intraoperatoria vi è un sospetto di litiasi residua a livello del coledoco distale. Gli esami emato-chimici post-procedurali documentano comunque una graduale normalizzazione del pannello epatobiliopancreatico. Il decorso post-operatorio è gravato da ritenzione urinaria acuta per cui posizioniamo CV, poi rimosso con ripresa della diuresi spontanea. Il controllo urologico dopo la rimozione del presidio documenta residuo post-minzionale di circa 100cc e prostata di volume moderatamente aumentato (100cc). La degenza risulta altrimenti regolare con buona tolleranza dell'alimentazione e algie palliate con terapia antalgica per os. È quindi possibile dimettere il paziente al domicilio in buone condizioni generali ed apiretico, con ferite chirurgiche in ordine.</p> <p>AI4 Il paziente è stato ricoverato per una colecistite acuta litiasica, trattato inizialmente con la sospensione di Eliquis e l'introduzione di Clexane 40 a partire dal MM.DD. La terapia antibiotica adottata comprendeva Ceftriaxone 2gr/24h e Metronidazolo 500mg/8h. Inizialmente, il paziente presentava condizioni generali nella norma, con cute rosea e un addome globoso ma morbido, senza segni di peritonismo. La PCR era elevata a 147, con iperbilirubinemia mista e colestasi<sup>1</sup>, mentre le lipasi risultavano negative. Erano stati programmati esami ematochimici completi<sup>2</sup>, una ColangioRM per il giorno seguente<sup>3</sup> e il monitoraggio della glicemia a digiuno<sup>4</sup>. Il MM.DD, dopo la sospensione di Eliquis e l'introduzione di Clexane, il paziente mostrava miglioramenti con un addome molle e non dolente<sup>5</sup>. La Colangio-MR era stata eseguita, in attesa di referto, e i valori di coagulazione evidenziavano un INR di 1.3, con introduzione di Konaktion a causa di un anti-Xa inferiore a 0.5<sup>6</sup>. Il MM.DD, la Colangio-RM confermava la diagnosi di colecistite acuta, con la colecisti piena di sludge e calcoli, senza evidenze di calcoli nel coledoco. Era stata programmata una colecistectomia per quel giorno. Veniva anticipata la necessità di una successiva RM per il controllo del decorso di un IPMN già noto. Il MM.DD, l'intervento di colecistectomia si era svolto senza complicazioni, sebbene fosse stata osservata una possibile presenza di un piccolo calcolo nel coledoco e un passaggio di contrasto in duodeno rallentato. Seguivano lievi segni di colestasi e rialzo della citolisi epatica, con bilirubina a 23 (da un precedente valore di 70)<sup>7</sup>. A seguito di ritenzione urinaria acuta, era stato posizionato un catetere vescicale, con previsione di rimozione il giorno successivo e valutazione urologica programmata. In sintesi, il paziente ha mostrato un progressivo miglioramento clinico dopo il trattamento iniziale per colecistite acuta litiasica e successiva colecistectomia, con attenzione a complicanze post-operatorie minime e monitoraggio continuo delle condizioni ematochimiche e della funzionalità epatica<sup>8</sup>.</p>
---

Figure 2: Comparison of the MD summary with the AI-generated summary for a surgery case. Color coding: yellow - information not present in the prompt; grey - information provided in the prompt but omitted in the generated summary; cyan - information not relevant in a discharge letter; magenta - information not entirely correct; red - hallucination.

## 5. Discussion

Our experiment shows that LLMs offer considerable potential for improving the summarization of clinical records in healthcare, particularly for the preparation of discharge letters. However, their adoption faces challenges.

First and foremost, our experiment setup, which relies on a publicly available ChatGPT model,

raises significant privacy concerns, as the de-identification tool may fail to completely remove sensitive parts of the data. In our case, since the number of cases was limited, the risk was mitigated by manual validation of complete de-identification before submission. Secondly, it remains challenging to ensure that the generated summaries consistently maintain a high level of accuracy. The most common quantitative metrics excel at capturing the

broad meaning of a text, but they cannot capture specific details crucial in clinical practice. Qualitative evaluations, while providing more specific indications, rely on expert evaluation, which is often subjective, and also extremely expensive to obtain. The observation that the versions of the DLs generated by ChatGPT-4 were considered better than those generated by ChatGPT-3.5, in particular with enhanced understanding of temporal aspects, is a positive signal that indicates further improvements can be expected. From a quantitative standpoint, it's interesting to observe that ChatGPT-4 produces longer summaries compared to ChatGPT-3.5, with an increase of around 20%.

We would like to add several observations regarding the obtained quantitative results. First, as was evidenced in yellow coded parts in Figure 1 and Figure 2, given that AI models operated exclusively based on the information provided in the prompts, which were missing some of the extra information available to doctors, the content of expert summary (MD) remarkably extends that of the AI-generated one. This clearly drastically reduces the overlapping parts of AI-generated DL and MD summaries. Given that ROUGE-L is based on the longest common sequence of words (not necessarily consecutive, but still in order) shared between AI-generated DL and ground truth (MD summaries), it is thus not surprising that the obtained scores are very low for both ChatGPT-3.5 and ChatGPT-4. The same problem reflects even more drastically on BLEU scores, since they exploit *consecutive* sequences of words (in our case, up to 3-grams were considered). Finally, there are at least three reasons for obtaining somewhat unexpected negative BLEURT scores: 1) using BERT-Tiny as checkpoint was probably not the best option since although very light is also known to be very inaccurate<sup>5</sup>; 2) the more stable BLEURT checkpoint BLEURT-20 was not tested on Italian language; 3) BLEURT scores heavily depend on the quality and representatives of the training data and may not fully capture the nuances of language quality across different domains or contexts. We thus recommend to consider BLEURT scores with caution.

Clinical notes are very detailed in nature, as they must contain all the information utilized for patient management during hospitalization. In our experimental setup, the AI models appeared unable to accurately filter relevant information to be included in the discharge letter. To address this gap, we could modify the structure of the clinical notes (e.g. by implementing a more structured reporting format for the information), or enhance the prompt, or try different models.

Given the positive outcome of the feasibility study described in this paper (as corroborated also by

medical experts), we are now setting up a larger and more advanced experiment which will enable us to tackle some of the shortcomings previously described. The first crucial step will be to use a local installation of an advanced open-source domain-specific model such as (Chen et al.; Jin et al.; Li et al., 2023), which were specifically trained on medical terminology and context. These specialized models can better capture the intricacies of medical causality, enhance the coherence and reduce errors in term interpretation. Additionally, the local installation will enable larger experiments, while at the same time mitigating privacy risks. The experimental strategy will involve a combination of prompt engineering techniques, including knowledge-infused prompting, chained inference, and corrective retrieval-augmented generation (Yan et al.). During *prompt engineering* (Brown et al., 2020), we can enrich the model's prompt with specific information about guidelines governing the generation process. In a *chained inference process* the AI model self-reflects and critiques its initial answer, subsequently generating a refined response based on this introspection. Finally, by contextualizing prompts with clinical topics from reputable sources, potentially obtained through *retrieval augmented generation*, we aim to provide more relevant and grounded knowledge to the LLM, enabling it to accurately correlate medical information.

## 6. Conclusion

In this paper we presented the results of a preliminary experiment aimed at testing the feasibility of automatic generation of discharge summaries in Italian. The setting of our experiment is deliberately oversimplified, in order to enable the validation of the idea, before attempting experiments that would require larger investments, such as the in-house installation and usage of an open-source LLM.

The results show that a generic generative system such as ChatGPT is capable of producing discharge summaries which are relatively close to the human generated ones, even in Italian. We have however noticed some shortcomings, which will need to be addressed in order for the system to be used in production. These observations have been collected and will guide the development of strategies to overcome them, such as enhanced prompting and retrieval-augmented generation.

---

<sup>5</sup><https://github.com/google-research/bleurt>

## 7. Limitations and Ethical Considerations

We are aware that this work has several limitations. First, we operate with limited number of clinical notes and consider only two medical specialities. Second, we consider only Italian language hence the obtained insights might not be transferable to other languages.

In accordance with ethical principles, this scientific study exploits data de-identification to safeguard the privacy and confidentiality of patients, thus aiming to minimize the risk of potential harm or identification. All clinical notes were manually revised after de-identification, to make sure that no instance of personally identifiable information was left in them.

We also contacted the ethical committee and they confirmed that this type of research did not require their authorization.

## 8. Bibliographical References

- Kenichiro Ando, Mamoru Komachi, Takashi Okumura, Hiromasa Horiguchi, and Yuji Matsumoto. 2022. Is in-hospital meta-information useful for abstractive discharge summary generation? In *2022 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 143–148. IEEE.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. *MEDITRON-70B: Scaling Medical Pretraining for Large Language Models*.
- Giuseppe Cocco. 2012. Bureaucracy and medicine, an unholy marriage. *Cardiovasc Med*, 15:243–244.
- Michael B Eppler, Conner Ganjavi, J Everett Knudsen, Ryan J Davis, Oluwatobiloba Ayo-Ajibola, Aditya Desai, Lorenzo Storino Ramacciotti, Andrew Chen, Andre De Castro Abreu, Mihir M Desai, et al. 2023. Bridging the gap between urological research and patient understanding: the role of large language models in automated generation of layperson’s summaries. *Urology practice*, 10(5):436–443.
- Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis. 2023. *Abstractive vs. extractive summarization: An experimental review*. *Applied Sciences*, 13(13).
- Lukas Golder, Claude Longchamp, Martina Imfeld, Silvia Ratelband-Pally, Stephan Tschöpe, Andreas Stettler, and Jonas Ph. Kocher. 2011. Drg: Befürchtungen einer zunehmenden bürokratisierung der medizin. Technical report, Gfs.bern, Hirschengraben 5, Postfach 6323, 3001 Bern, Switzerland.
- Vince C Hartman, Sanika S Bapat, Mark G Weiner, Babak B Navi, Evan T Sholle, and Thomas R Campion Jr. 2023. A method to automate the discharge summary hospital course for neurology patients. *Journal of the American Medical Informatics Association*, 30(12):1995–2003.
- Mingyu Jin, Qinkai Yu, Chong Zhang, Dong Shu, Suiyuan Zhu, Mengnan Du, Yongfeng Zhang, and Yanda Meng. 2024. Health-llm: Personalized retrieval-augmented disease prediction model. *arXiv preprint arXiv:2402.00746*.
- Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. *MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval*. 39(11):btad651.
- Tiziano Labruna, Sofia Brenna, Andrea Zaninello, and Bernardo Magnini. 2023. Unraveling chatgpt: A critical analysis of ai-generated goal-oriented dialogues and annotations. In *International Conference of the Italian Association for Artificial Intelligence*, pages 151–171. Springer.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day.
- Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, pages 150–157.
- Carlo A Mallio, Andrea C Sertorio, Caterina Berneti, and Bruno Beomonte Zobel. 2023. Large language models for structured reporting in

- radiology: performance of gpt-4, chatgpt-3.5, perplexity and bing. La radiologia medica, 128(7):808–812.
- Sara Montagna, Stefano Ferretti, Lorenz Cuno Klopfenstein, Antonio Florio, and Martino Francesco Pengo. 2023. Data decentralisation of llm-based chatbot systems in chronic disease self-management. In Proceedings of the 2023 ACM Conference on Information Technology for Social Good, pages 205–212.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.
- Thomas P Reith. 2018. Burnout in united states healthcare professionals: a narrative review. Cureus, 10(12).
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. [Bleurt: Learning robust metrics for text generation](#). In ACL.
- J. Xu. 2024. [Discharge me: Bionlp acl'24 shared task on streamlining discharge documentation \(version 1.2\)](#).
- Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, Wei Jin, Joyce Ho, and Carl Yang. 2023. Knowledge-Infused Prompting: Assessing and Advancing Clinical Text Data Generation with Large Language Models.
- Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. [Corrective Retrieval Augmented Generation](#).
- Jonah Zaretsky, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan Austrian, Yindalon Aphinyanaphongs, Ravi Gupta, Saul B Blecker, and Jonah Feldman. 2024. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Network Open, 7(3):e240357–e240357.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.
- Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. arXiv preprint arXiv:1909.02622.

## A. An Example of a Prompt

```
P Crea un riassunto del seguente decorso clinico da includere nella lettera di uscita:
[{"CONTENUTO": "Dimesso alle 14:00", "DATA": "2023-01-09 11:21:49"},
{"CONTENUTO": "Paziente stabile, PV nella norma. Apiretico."
  "Procedere:"
  "- sorveglianza clinica e febbre. Se persiste, valutare switch a levofloxacina"
  "- labor controllo domani (K, INR)"
  "- rientro al domicilio senza aiuti inizio settimana prossima",
"DATA": "2023-01-08 12:29:57"},
{"CONTENUTO": "Paziente stabile, PV nella norma. "
  "Assenza di picchi febbrili. Buone Spo2 senza ossigenoterapia."
  "Procedere:"
  "- sorveglianza clinica e febbre. Se persiste, valutare switch a levofloxacina"
  "- labor controllo lunedì (K, INR)"
  "- rientro al domicilio senza aiuti inizio settimana prossima",
"DATA": "2023-01-07 10:23:32"},
{"CONTENUTO": "Paziente stabile, PV nella norma. Assenza di picchi febbrili. "
  "Buone Spo2 senza ossigenoterapia. Per ipokaliemia in corso sostituzione."
  "Procedere WEEKEND:"
  "- sorveglianza clinica e febbre. Se persiste, valutare switch a levofloxacina"
  "- labor controllo lunedì e RAD se stabilità clinica",
"DATA": "2023-01-06 10:01:10"},
{"CONTENUTO": "Febbrile durante la notte a 38°C, afebrile durante il giorno. "
  "Resto parametri nella notte. L: PCR stagnante, ipokaliemia. Non indicazione a switch antibioterapia"
  "Procedere WEEKEND:"
  "- sorveglianza clinica e febbre. Se persiste, valutare switch a levofloxacina"
  "Procedere:"
  "- RAD lunedì"
  "- labor controllo lunedì",
"DATA": "2023-01-05 12:11:11"},
{"CONTENUTO": "Addendum 04.01.2022 "
  "- Dichiarazione Legionella al xxxxxxxxxxxxxxxx fatta oggi"
  "- Sintrom da adattare secondo INR domani",
"DATA": "2023-01-04 18:33:03"},
{"CONTENUTO": "PV nella norma, afebrile. Questa mattina saturava a 96% con 2L di O2, lo svezziamo e in giornata
  "satura a 92% senza O2. "
  "Cercare di mantenere senza O2."
  "Procedere:"
  "- Domani labor"
  "- Aritromicina per 10 giorni"
  "- RAD lunedì",
"DATA": "2023-01-04 14:57:55"},
{"CONTENUTO": "PV stabili, febbrile a 38.8°. No tosse. Crepitii mediobasali a destra. AG legionella positivi --> Th: Azitromicina
  "10 giorni secondo evoluzione clinica "
  "(per il momento impostata fino all'11, rivalutare). Schema sintron reimpostato"
  "Procedere:"
  "- Sorveglianza clinica e laboratoristica"
  "- Se tutto Ok, RAD giovedì",
"DATA": "2023-01-03 12:37:34"},
{"CONTENUTO": "Paziente ricoverato per polmonite ad acquisizione comunitaria "
  "con IRA AKIN I di origine pre renale su disidratazione. Vengo chiamata in serata per valori pressori ipotensivi "
  "(PA 85/55 mmHg) in un paziente asintomatico. "
  "Si somministrano 250 ml di liquidi in 30 minuti con risposta sulla pressione. Alla visita paziente vigile ed orientato. "
  "Diaforetico e febbrile. Mucose secche."
  "Si velocizza la somministrazione dei liquidi prescrivendo 1000 ml in 12 ore.",
"DATA": "2023-01-03 00:02:17"}]
```

Figure 3: Example of a prompt (P)

# Evaluating LLMs for Temporal Entity Extraction from Pediatric Clinical Text in Rare Diseases Context

Judith Jeyafreeda Andrew, Marc Vincent, Anita Burgun, Nicolas Garcelon

Université de Paris, Imagine Institute, Data Science Platform, INSERM UMR 1163

PaRis Artificial Intelligence Research InstitutE (PRAIRIE)

Bd de Montparnasse, Paris, France

{judith.jeyafreeda, marc.vincent, nicolas.garcelon}@institutimagine.org

anita.burgun@aphp.fr

## Abstract

The aim of this work is to extract Temporal Entities from patients' EHR from pediatric hospital specialising in Rare Diseases, thus allowing to create a patient timeline relative to diagnosis. We aim to perform an evaluation of NLP tools and Large Language Models (LLM) to test their application in the field of clinical study where data is limited and sensitive. We present a short annotation guideline for temporal entity identification. We then use the tool EDS-NLP, the Language Model CamemBERT-with-Dates and the LLM Vicuna to extract temporal entities. We perform experiments using three different prompting techniques on the LLM Vicuna to evaluate the model thoroughly. We use a small dataset of 50 EHR describing the evolution of rare diseases in patients to perform our experiments. We show that among the different methods to prompt a LLM, using a decomposed structure of prompting method on the LLM Vicuna produces the best results for temporal entity recognition. The LLM learns from examples in the prompt and decomposing one prompt to several prompts allows the model to avoid confusions between the different entity types. Identifying the temporal entities in EHRs helps to build the timeline of a patient and to learn the evolution of a diseases. This is specifically important in the case of rare diseases due to the availability of limited examples. In this paper, we show that this can be made possible with the use of Language Models and LLM in a secure environment, thus preserving the privacy of the patient.

**Keywords:** Temporal Entities, Vicuna, Prompt-based learning, rare diseases

## 1. Introduction

Electronic Health Records (EHR) contain several valuable information that help in advancing clinical research. Automatic extraction of information from EHRs has evolved greatly overtime with the development of Machine Learning and Natural Language Processing (NLP) techniques. In the present article we focus on a sub-task of NLP: Named Entity Recognition (NER) of temporal entities. In particular, we aim at extracting temporal entities from EHRs of patients with Rare Diseases. Identifying the temporal Entities in such texts allows to build the timeline of a patient, allowing for the analysis of patient history, prediction of next steps in the process of diagnosing a disease and the evolution of a patient after a therapeutic decision has been taken. This is a very important application in the field of rare diseases where the data is limited.

There have been several research works for the automatic extraction of information from clinical texts. These works have enabled building several novel methods and models for the extraction of useful information within the clinical texts such as drugs, treatments, diagnosis, symptoms, etc. However, to be able to create a timeline of a patient, the relations between these entities and temporal entities such as date, time, duration etc must be established. For this purpose, the extraction of

temporal entities is also essential.

Although there have been considerable efforts in making de-identified EHRs publicly available, accessible after considerable ethical training, the language and format of the EHRs influence greatly the development of Large Language Models for Information Extraction. Models and methods that perform well for the English Language do not necessarily have the same performance on the French language. Also, the format of EHR used in a clinic might not be the same as the format used in another clinic, this also affects the performance of a model. Thus external validation of LLMs with local datasets is essential.

Thus in this work, we use EHR from patients with rare disease particular to the Necker Hospital in Paris <sup>1 2</sup> for the extraction of temporal entities. Our contributions in this paper are as follows: (i) a short annotation guideline that has been used for a manual annotation. (ii) using existing tools and Large Language models for temporal entity extraction to study their performance and re-usability in a secure environment. (iii) applied to real hospital data.

<sup>1</sup>Hospital located at 149 Rue de Sèvres, 75015 Paris

<sup>2</sup>The dataset is private and cannot be distributed



## 2. Related Works

(Bose et al., 2021) gives a detailed study on all NER methods and models available in the clinical context until the year 2020. The entities that are often used in the clinical context are drugs, diagnosis, treatment, dosage, family history etc. The methods of NER used include dictionary-based approach, rule-based approach, CRF, Machine Learning based approach, Deep Learning-Based Methods and some hybrid approaches. The authors show the several models that are available in different language, most being in Chinese and some in English. Although this study does not mention extraction of temporal entities, there has been several works done in the field of temporal relation extraction in clinical text in the English Language. (Alfattni et al., 2020) points to the general approach used in Temporal Relation Extraction which include pre-processing, NER of EVENTS and TIMEX entities, TLINK candidate extraction, TLINK classification and post-processing, thus indicating the importance of having an efficient temporal entity recognition method for the task of Temporal Relation Extraction. Within the context of clinical texts in the French language, (Tourille, 2018), has studied various approaches for NER within the clinical context and presented the results on publicly available French corpora. The author uses an LSTM approach with inspiration from sequence labelling for the purpose of NER, while the temporal relation extraction relies solely on LSTM. Lastly, in (Vincent et al., 2022) and (Faviez et al., 2022) the authors use deep learning and hybrid NER methods to perform deep phenotyping on a specialised rare disease dataset, using the resulting models and information extraction to augment the UMLS metathesaurus with specific and previously not included terms.

### 2.1. Prompt Learning for NER

Prompt learning has gained increasing popularity with the development of LLM and they have been used successfully for several NLP applications (Brown et al., 2020). Prompt learning involves using prompts which are injected to the input into a designed template. This converts the downstream task into a fill-in-the-blank task, then allows the language model to predict the slots in the prompts and eventually deduce the final output. This is often used for text generation and classification tasks. There have been several research works on the several prompting techniques such as discrete and continuous prompt templates (Jiang et al., 2020), (Shin et al., 2020), (Liu et al., 2023), (Li and Liang, 2021), (Lester et al., 2021), (Qin and Eisner, 2021). (Cui et al., 2021) is one of the first attempts in using prompt learning for NER. (Ashok

and Lipton, 2023) introduces PromptNER, where a text and a task description is given along with the question for the prediction of entities. This has been tested on the biomedical dataset GENIA (Kim et al., 2003) for NER and outperforms competing models like GPT 3.5. In (Liu et al., 2022), the authors present QaNER, which is a prompt-based learning NER method with Question Answering. The authors of (Ye et al., 2023) propose a decomposed two-stage prompt learning framework for few-shot named entity recognition, which include the entity location and entity typing stages. (Shen et al., 2023) unify entity locating and entity typing in prompt learning for NER with a dual-slot multi-prompt template. (Huang et al., 2022) proposes a few-shot NER approach named COPNER, which combines contrastive learning and prompt guiding, where the prompt is concatenated with the sentence and is then fed to a pre-trained language model.

In this work, we use three different prompts with the "Vicuna" large language model (LLM). The first prompt is a basic question which asks the LLM to identify all the temporal entities in a given clinical text. The second prompt, is a definition based prompt where the entities are defined as part of the prompt which helps the LLM understand the entities that are to be identified. For the third prompt, we decompose the prompt into different prompts (one for each entity).

## 3. Dataset

As mentioned previously, the language and format of clinical text have a great deal of influence to the performance of large language models. (Youssef A, 2023) has stressed the need for external evaluation in the setting where the LLM models are to be deployed. The selection of testing dataset would depend on the setting of the deployment environment. In this work, we focus on clinical texts in the French language. Our dataset is a collection of patients' EHRs from The Necker pediatric Hospital in Paris, specialised in Rare diseases. (Garcelon et al., 2018) describes Dr Warehouse, which is a database used at the Necker Children's Hospital. The features and capabilities of this database enables efficient use of NLP techniques in a secure environment.

DATE	AGE	DURATION	FREQ	TIME
213	47	12	58	81

Table 1: Number of each entity in the Gold Standard

### 3.1. Annotation Guidelines

Defining temporal entities within the clinical context can be a difficult task, we build on previous works to do so, most notably the guidelines presented as part of the annotation of the MERLOT corpus (Campillos-Llanos et al., 2018). Broadly, temporal entities can be categorized into the following classes: Dates (including Date of Birth, Date of visit, Date of Report, Date of test, Date of consultation, Date of next scheduled visit), Time, Frequency, Duration and Age. In order to produce reliable and reproducible annotations of the available clinical data, we established the following guidelines, giving precise definitions of each categories as well as informative or borderlines cases that were found by comparing several annotators outputs:

**DATE:** All dates that are presented within the clinical text. This can be any date including the dates representing the history of the patient, date of birth, date of visit, date of creation of the record, date of identification of a diagnosis, date of commencement of medication etc.

Date mentions can be either complete or incomplete. We consider date mentions to be complete if they mention a year (optionally completed by a month and/or a day), while mentions lacking the mention of a year are considered incomplete (i.e. they require extra information to unambiguously determine the 'absolute' date they refer to). Irrespective of complete or incomplete mentions, these entities are annotated as DATE.

*Examples:*

- "Craniopharyngiome type decouvert sur des signes d'HTIC en Aout" → **Aout** annotated as DATE
- "Radiotherapie prevue debut Novembre" → **debut Novembre** annotated as DATE
- "Je propose un rendez-vous de consultation le 20 decembre" → **20 decembre** annotated as DATE
- "Dicte le: 02/02/2021" → **02/02/2021** annotated as DATE
- "Paris le 01/07/2000" → **01/07/2000** annotated as DATE

If the Date is written as a range with the year and/month attached to the second part, a fragment with the day, month and year to complete the DATE  
*Ex:* "Hospitalise(e) du 19 au 29/07/2023": fragment with 19/07/2023 annotated as DATE and another entity 29/07/2023 annotated as DATE (not as DURATION)

If the DATE includes days such as "Lundi 3 Mars 2011", the entire phrase is annotated as DATE, including the day

**AGE:** This refers to the age of the patient presented in the text, his/her parents or relations, age of a fetus. A fetus's age is usually represented in terms of "SA" or as "Age Gestationnel" *Ex1:* IMG à 33SA + 5jours pour immobilisme foetal, Caryotype normale → **33SA + 5jours** annotated as AGE. *Ex2:* Il a 36 ans → **36 ans** annotated as AGE

**DURATION:** This entity reference to a continuous duration of time. *Ex:* "depuis le 20/1/2001", "pendant 2 jours", "depuis plus de 25 ans" etc.

**FREQUENCY:** Any time related quantity repeated at regular intervals. *Ex:* "par jour", "par semaine", "par seconde" "/jour", "/hr", "/le soir", "/le matin", "tout les matins" etc. FREQUENCY also includes visits to the clinic schedules at specific intervals or tests scheduled at/taken at specific intervals.

- KCL 10ml par jour → **par jour** annotated as FREQUENCY

- Heparine 70 mg dans 48 ml, vitesse 5ml/heure → **/heure** annotated as FREQUENCY

**TIME:** This entity refers to the any time relative to a date. (i.e) when the date is unclear, it is TIME. *Ex:* "4 semaines", "4 jours", "toujours", "ce moment", "ce jour", "matin", "midi", "soir" etc

- Any specific time to be marked as time. *Ex:* "9:28"

- A "rendez-vous" made after certain amount of time is to be annotated as Time, without a specific date mentioned. *Ex1:* Nouveau controle endoscopique dans 3 mois → 3 mois annotated as TIME. *Ex2:* Prochain RDV dans 1 semaine → 1 *semaine* annotated as TIME

- Time indicated as J1, J2 ..etc indicate "Jour 1", "jour 2" etc. Thus these should be annotated as time, since they are relative to the date.

### 3.2. Annotation Process

For the purpose of testing our experiments, we annotate 50 clinical notes using the annotation guidelines as mention in section 3.1. Three annotators were asked to annotate the same set of clinical notes to be able to establish a gold standard. They were given the same set of the above mentioned annotation guidelines. The methods and models are tested and evaluated on these 50 notes.

A set of 150 EHRs has been annotated by one annotator using the above mentioned guidelines which can be used for training any language model.

## 4. Experimental Setup

There are indeed several tools that explore temporal entities in the French language. Even if these

tools and models are not particularly tailored for the clinical context, these can be used to identify basic dates and times within the text. In this paper, we perform experiments with 3 existing tools and models on our hospital local dataset. We then evaluate the results to determine how the tools and models perform on our internal dataset.

The experiments are performed using local installations of the tools and models, thus preserving the privacy of patient information.

**EDS-NLP:** (Wajsburt et al., 2022) is a NLP framework that aims at extracting information from French clinical notes. It is a collection of components or pipes, either rule-based functions or deep learning modules. EDS-NLP has a component (*eds.date*) for extracting dates in medical reports. In this paper, we apply EDS NLP's *date component* to detect temporal entities in our dataset. This method is able to identify the dates as an entity, however this method fails to differentiate between the temporal entities such as duration and frequency. We use the 50 clinical texts annotated by the 3 annotators to extract the temporal entities. The results are then used to be compared with the manual annotations.

**CamemBERT-with-Dates:** (Martin et al., 2020) CamemBERT is a state-of-the-art language model for French based on the RoBERTa architecture pre-trained on the French subcorpus of the multilingual corpus OSCAR. CamemBERT-with-dates is an extension of french camembert-ner model with an additional tag for dates. This model was trained on an enriched version of wikiner-fr dataset. This model is able to identify the dates as an entity, however this model fails to differentiate between the temporal entities such as duration and frequency, as the model is not trained for these entities. For the first experiment, we extract the temporal entities from the 50 clinical texts annotated by the 3 annotators. The results are then compared with the manual annotations. For the second experiment, we fine-tune the CamemBERT-with-dates model using the 150 clinical texts that has been annotated by one annotator as stated in section 3.2. The fine-tuned model is then tested on the 50 clinical texts (annotated by the 3 annotators). The results from the fine-tuned model is then used to be compared with the manual annotations.

**Large Language Model:** In this work, we use the Vicuna model (Chiang et al., 2023) for testing the prompt based approach on the dataset. Vicuna is an open-source large Language Model (LLM) with 13 billion parameters. There are several versions of Vicuna available. For experimentation, we use Vicuna v1.5. This model is fine-tuned from Llama2 with supervised instruction fine-tuning and

linear RoPE scaling. The training data is around 125K conversations collected from ShareGPT.com. These conversations are packed into sequences that contain 16K tokens each.

In this work, we setup a local version of the model that is used for experimentation, so as to preserve the privacy of the dataset. This model is prompted with three different kinds of prompts to identify the temporal entities.

We use prompt based methods to query the LLM for the purpose of identifying temporal entities. As mentioned in section 2.1, there have been several works on using various types of templates for prompting LLMs. In this work, we experiment with 3 different prompts to extract temporal entities using the Vicuna LLM. They are as follows:

- Posing a general question to the LLM (Vicuna) to identify the temporal entities (i.e What are the temporal entities in the text "..."?).
- Defining the temporal entities to the LLM before posing the question to the LLM. For example: We define all entities together such as "date: date written in any format. time: time of the day or any time without mention of date. age is the age of the patient or fetus. frequency: time related quantity repeated at regular intervals. Ex: "par jour", "par semaine", "par seconde" "/jour", "/hr", "/le soir", "/le matin", "tout les matins" etc Duration: a continuous duration of time. Ex: "depuis le 20/1/2001", "pendant 2 jours", "depuis plus de 25 ans" etc." and then ask Vicuna to identify all temporal entities defined above
- Decomposing the prompt into several parts. In this part, we split the prompt into 5 different prompts (one for each entity). Each of the prompt has a definition of the entity with examples and a question asking the LLM to identify that particular entity. For example: "time is defined as any time of the day like "matin", "soir", "midi" or any time without mention of date like "ce jour", "ce moment", "aujourd'hui" or time indicated as number of says like "Jour 1", "Jour 2" etc or "J1", "J2" etc. Identify all the mentions of TIME entities in the following text: ..."

For the purpose of evaluation, a certain amount of post-processing is required as comparison to the gold standard annotation requires the outputs from the tools and models to have span (start and end indices) of the entities. As mentioned in (Ashok and Lipton, 2023), one of the limitations of prompting LLMs is the preservation of spans for the entities. As the testing data is small (50 EHR), the post processing of matching the entity with the span was done manually.

## 5. Results and Discussion

The results from our experiments (as mentioned in section 4) are presented in Tables 2 and 3. Table 2 gives the F1 scores of the entities, while table 3 provides a token level evaluation that counts partial token matches of multi-tokens terms as positives.

**EDS-NLP and CamemBERT:** Both EDS-NLP and CamemBERT, do not differentiate dates with frequency, duration, time or age. That is, every temporal entity is labelled as DATE. For example: In the text: "*Depuis Juin 2008, la creatininémie augmentée*", the entity "*Juin 2008*", is marked as DATE by both EDS-NLP and CamemBERT, while according the Gold Standard annotations they should be marked as DURATION. Phrases such as "*Il y a 5 mois*", "*par semaine*" are also marked as DATE by both EDS-NLP and CamemBERT, while according the Gold Standard annotations they should be marked as TIME and FREQUENCY respectively. Thus, to have a fair evaluation of these tools, we mark all temporal entities as DATE in the Gold Standard as well (i.e), all the other entities (AGE, DURATION, FREQUENCY and TIME) are renamed as DATE for the purpose of evaluating EDS-NLP and CamemBERT with our test dataset.

It has to be noted that EDS-NLP has been developed for French Clinical texts, while CamemBERT-with-dates has been trained for the French language but not particularly for clinical texts.

**CamemBERT Finetuned:** For the purpose of fine-tuning a language model, we use the 150 documents annotated by one annotator. All temporal entities in these 150 texts are DATE, (i.e), all the other entities (AGE, DURATION, FREQUENCY and TIME) are renamed as DATE. This will help to fine-tune the CamemBERT-with-Dates model more efficiently as DATE is already a supported entity by the model. The fine-tuned model (dubbed CamemBERT-fit in the results tables) is then tested on the 50 EHRs (annotated by 3 annotators). As seen from Tables 2 and 3, there is definitely improvements in the results when a fine-tuned model is used. However, table 2 shows very low F1 score (0.047) for the DATE entity. This is because of variations in the tokenization used by the model. For example: the text "16.04.1968" is marked as a whole as DATE, however, the model splits the tokens into three different tokens as "16","04","1968" and each of them are labelled as DATE. This is evident from Table 3 where the token level evaluation is presented. This shows a F1 score of 0.758 for the fine-tuned CamemBERT-with-Dates model. It is to be noted that only 150 documents were used to fine-tune the model. The number of Epochs used for fine tuning is 25. Given the improvement in result of a fine-tuned model when compared to the raw model, even while using such a small amount

of data for fine-tuning, it can be envisioned that using a bigger amount of data for fine-tuning could result in a more competitive model.

**LLM - Vicuna:** We have used three different prompts with Vicuna to extract the temporal entities in the text. It has to be noted that Vicuna is not particularly trained for the French Language, nor particularly for clinical texts but positive results on early experiments prompted us to continue testing it.

The first prompt, being a very general prompt demanding the LLM to identify all temporal entities, while performing well for the identification of DATE, AGE and Duration entities, does not perform well for FREQUENCY and TIME (Tables 2 and 3). It has a poor performance specifically for the FREQUENCY entity as the LLM is not able to understand our definition of FREQUENCY. For example: In the text, "*KCL 10 ml par jour*", the entity "*par jour*" is not marked at all, while it has to be marked as frequency. This is because a general question to the LLM demanding the identification of temporal entities is not well understood by the model.

The second prompt, where the definitions of all the entities are given to the LLM before posing a question asking for the identification of the defined entities, the results (Tables 2 and 3) are better. The results for the entity FREQUENCY has improved a lot as the model is now able to understand each entity. The definition of the FREQUENCY and DURATION also includes examples for each entity, thus helping Vicuna to learn from example. For the TIME entity, there seem to be several TIME entities misclassified as DATE like "*ce jour*", "*ce semaine*" etc.

The third prompt, where a prompt is generated for each entity with examples before posing questions to the LLM, performs the best. In particular, the TIME entity improves in performance drastically. Not only does the model learn from examples but by giving individual prompts for each entities, the confusion between DATE and TIME is avoided. Thus entities like "*ce jour*", "*ce matin*", "*aujourd'hui*" etc which are classified as DATE while using the second prompt is correctly classified while using the third prompt.

## 6. Conclusion

In this paper, we performed an external validation for extraction of temporal entities using the NER tool (EDS-NLP), Language model (CamemBERT-with-Dates) and Large Language Model (Vicuna). There are several other LLM, such as described in (Touvron et al., 2023) with Llama models ranging from 7B to 70B parameters. There are also newer models such as Mistral-7B-v0.1 (Jiang et al., 2023), which is a small ( 7-billion parameters) but powerful

Method	DATE	AGE	DURATION	FREQ.	TIME
EDS-NLP	0.560	NA	NA	NA	NA
CamemBERT	0.024	NA	NA	NA	NA
CamemBERT-ft	0.047	NA	NA	NA	NA
Vicuna Prompt1	0.842	0.84	0.857	0.067	0.527
Vicuna Prompt2	0.853	0.854	0.957	0.840	0.615
Vicuna Prompt3	0.862	0.860	0.960	0.848	0.860

Table 2: F1 scores for entities

Method	DATE	AGE	DURATION	FREQ.	TIME
EDS-NLP	0.779	NA	NA	NA	NA
CamemBERT	0.543	NA	NA	NA	NA
CamemBERT-ft	0.758	NA	NA	NA	NA
Vicuna Prompt1	0.830	0.861	0.822	0.097	0.577
Vicuna Prompt2	0.867	0.840	0.938	0.852	0.667
Vicuna Prompt3	0.912	0.881	0.938	0.867	0.90

Table 3: Token wise F1 evaluation

language model adaptable to several down-stream tasks and shown to perform better than Llama 2 13B on all tested benchmarks. (Jiang et al., 2023). We made a choice to use Vicuna for our experiments as we had the computing power and memory to store a Vicuna model (13 billion parameters), and it displayed good performances (Zheng et al.) that our early experiments confirmed. As the set of available LLMs changes rapidly we intend to test further models such as Mistral-7B-v0.1, keeping in mind performance to cost ratio. Indeed, deploying a Language Model (Large or small) locally in a clinic can be difficult as it requires a significantly higher amount of storage space and computing power than smaller deep learning models, proportional to the increase in the number of parameters (assuming comparable implementations - other factors coming into play such as quantization, method for underlying attention, etc...).

Fine-tuning and storing any Language Model locally is expensive, thus the efficiency of the model is an important factor to be considered. We have selected other tools and models to perform a comparison study between tools tailored for clinical texts, models trained for French (not for clinical texts in particular) and an entirely different model without any context to the french language or for clinical text. This gives us a variety of options to consider before deployment.

From Tables 2 and 3, it can be seen that prompting a LLM with question for NER performs better than EDS-NLP and CamemBERT-with-Dates, even-though Vicuna is not specifically trained for French clinical texts. It is important to note that the dataset used for testing is small (50 EHR). This is a small sample size to generalize the results globally, however locally (within the clinic) it is a good

amount to be able to understand the requirements for good performance.

Language Models such as CamemBERT, though trained on fewer parameters, are easier to fine-tune for downstream tasks. While LLMs such as Vicuna, can have a good performance without any fine-tuning which can make them very useful in a context where data is not readily available and costly to produce. Thus choosing a model for extraction of information depends greatly on the local requirements.

The tools and models have been tested for temporal entities in EHRs of patients with rare diseases, however, this could be easily extended to other entities in any type of clinical text. Thus this presents a feasible method for analysing a patient's history, prediction of next steps and the evaluation of decisions taken.

## 7. Acknowledgement

This work was supported by state funding by The French National Research Agency (ANR) under the C'IL-LICO project (ANR-17-RHUS-0002) and as part of the "Investissements d'avenir" program (ANR-19-P3IA-0001) (PRAIRIE 3IA Institute). The authors acknowledge URC-CIC Paris Centre for the implementation of the study.

## 8. Bibliographical References

Ghada Alfattni, Niels Peek, and Goran Nenadic. 2020. [Extraction of temporal relations from clinical free text: A systematic review of current ap-](#)

- proaches. *Journal of Biomedical Informatics*, 108:103488.
- Dhananjay Ashok and Zachary C. Lipton. 2023. [Prompter: Prompting for named entity recognition](#).
- Priyankar Bose, Sriram Srinivasan, William C. Sleeman, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. 2021. [A survey on recent named entity recognition and relationship extraction techniques on clinical texts](#). *Applied Sciences*, 11(18).
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Leonardo Campillos-Llanos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéal. 2018. [A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus \(merlot\)](#). *Language Resources and Evaluation*, 52.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%\\* chatgpt quality](#).
- Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. [Template-based named entity recognition using BART](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845, Online. Association for Computational Linguistics.
- Carole Faviez, Marc Vincent, Nicolas Garcelon, Caroline Michot, Genevieve Baujat, Valerie Cormier-Daire, Sophie Saunier, Xiaoyi Chen, and Anita Burgun. 2022. Enriching umls-based phenotyping of rare diseases using deep-learning: Evaluation on jeune syndrome. In *Challenges of Trustable AI and Added-Value on Health*, pages 844–848. IOS Press.
- Nicolas Garcelon, Antoine Neuraz, Rémi Salomon, Hassan Faour, Vincent Benoit, Arthur Delapalme, Arnold Munnich, Anita Burgun, and Bastien Rance. 2018. [A clinician friendly data warehouse oriented toward narrative reports: Dr. warehouse](#). *Journal of Biomedical Informatics*, 80:52–63.
- Yucheng Huang, Kai He, Yige Wang, Xianli Zhang, Tieliang Gong, Rui Mao, and Chen Li. 2022. Copner: Contrastive learning with prompt guiding for few-shot named entity recognition. In *Proceedings of the 29th International conference on computational linguistics*, pages 2515–2527.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#).
- Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.
- J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun’ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl\_1):i180–i182.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.
- Andy T. Liu, Wei Xiao, Henghui Zhu, Dejiao Zhang, Shang-Wen Li, and Andrew Arnold. 2022. [Qaner: Prompting question answering models for few-shot named entity recognition](#).
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. Gpt understands, too. *AI Open*.
- Zengjian Liu, Ming Yang, Xiaolong Wang, Qingcai Chen, Buzhou Tang, Zhe Wang, and Hua Xu. 2017. Entity recognition from clinical texts via recurrent neural network. *BMC medical informatics and decision making*, 17:53–61.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. [CamemBERT: a tasty French language model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.
- Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. *arXiv preprint arXiv:2104.06599*.
- Yongliang Shen, Zeqi Tan, Shuhui Wu, Wenqi Zhang, Rongsheng Zhang, Yadong Xi, Weiming Lu, and Yueting Zhuang. 2023. Prompter: Prompt locating and typing for named entity recognition. *arXiv preprint arXiv:2305.17104*.

- Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Auto-prompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.
- Julien Tourille. 2018. *Extracting Clinical Event Timelines : Temporal Information Extraction and Coreference Resolution in Electronic Health Records*. Ph.D. thesis.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).
- Marc Vincent, Maxime Douillet, Ivan Lerner, Antoine Neuraz, Anita Burgun, and Nicolas Garcelon. 2022. Using deep learning to improve phenotyping from clinical reports. *Stud Health Technol Inform*, 290:282–6.
- Perceval Wajsburt, Thomas Petit-Jean, Basile Dura, Ariel Cohen, Charline Jean, and Romain Bey. 2022. [Eds-nlp: efficient information extraction from french clinical notes](#).
- Website. 2023. [Temporal entity definition](#).
- Feiyang Ye, Liang Huang, Senjie Liang, and KaiKai Chi. 2023. [Decomposed two-stage prompt learning for few-shot named entity recognition](#). *Information*, 14(5).
- Thakur A Zhu T Clifton D Shah NH Youssef A, Pencina M. 2023. [External validation of ai models in health should be replaced with recurring local validation](#).
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. corr, abs/2306.05685, 2023. doi: 10.48550. *arXiv preprint arXiv.2306.05685*.

# Generating Distributable Surrogate Corpus for Medical Multi-label Classification

Seiji Shimizu, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki

Nara Institute of Science and Technology, Nara, Japan  
{shimizu.seiji, s-yada, wakamiya, aramaki}@is.naist.jp

## Abstract

In medical and social media domains, annotated corpora are often hard to distribute due to copyrights and privacy issues. To overcome this situation, we propose a new method to generate a surrogate corpus for a downstream task by using a text generation model. We chose a medical multi-label classification task, *MedWeb*, in which patient-generated short messages express multiple symptoms. We first fine-tuned text generation models with different prompting designs on the original corpus to obtain synthetic versions of that corpus. To assess the viability of the generated corpora for the downstream task, we compared the performance of multi-label classification models trained either on the original or the surrogate corpora. The results and the error analysis showed the difficulty of generating surrogate corpus in multi-label settings, suggesting text generation under complex conditions is not trivial. On the other hand, our experiment demonstrates that the generated corpus with a sentinel-based prompting is comparatively viable in a single-label (multiclass) classification setting.

**Keywords:** Text Generation, Language Model, Privacy Protection, Social Media

## 1. Introduction

Supervised machine learning, which is the de facto standard in today's natural language processing (NLP), requires annotated corpora. Although sharing corpora with researchers enhances further development in scale, annotated corpora may not be distributed due to privacy policies and copyrights. Especially in the medical domain, this problem arises frequently and critically (Hahn and Oleynik, 2020; Aramaki et al., 2022). Also, social media posts may not only contain some personal information but are also often limited to content-excluding distribution in the platform's terms<sup>1</sup>.

Two major approaches have been taken to tackle the problem of difficulty in corpus distribution. The first approach is to delete personal information in the corpus, that is, de-identification (Sibanda and Uzuner, 2006; Uzuner et al., 2007) or anonymization (Zuo et al., 2021), which is well studied in the medical domain. MIMIC (Johnson et al., 2016) is the most popular de-identified corpus in the medical domain. However, it is costly and difficult to achieve perfect de-identification of arbitrarily large corpora, regardless of whether the method is based on machine learning or human labor.

The second approach is to generate new corpora in which any *real* person's information is not contained. One such corpus is *MedWeb* (Wakamiya et al., 2019), where patients' self-reports of symptoms were composed manually via crowdsourcing. Whereas manually generating data is highly costly, model-based automatic generation enables large-scale and low-cost corpus creation. The recent

advance in text generation (Zhang et al., 2022) promotes such an approach, for example, in the social media domain (Claveau et al., 2021) and in the medical domain (Amin-Nejad et al., 2020). However, existing studies investigate the viability of such generated corpora mainly for data augmentation, which extends the existing *small* datasets to be larger for data-hungry deep learning models. The generated corpora in this approach are to be mainly combined with the original dataset. The remaining question is: *Can a synthetic corpus created by text generation be a surrogate for a downstream task?*

This study aims at generating a distributable surrogate corpus and investigating its viability in the downstream task. We set the downstream task to multi-label classification in the medical domain, i.e., the aforementioned *MedWeb* task: Multiple symptoms (such as runny nose and cough) expressed in patient-generated short messages must be correctly labeled. We first generate synthetic corpora by generation models trained on the original corpus. Then, we evaluate the quality of the generated corpora by solving the task with classification models.

Specifically, in the generation step, we fine-tuned text generation models with different prompting methods (i.e., the sentinel tokens and soft prompts) to obtain different qualities of generated corpora. In the evaluation (classification) step, we trained the classification models on either the generated corpora or the original corpus. The flow of this experiment is outlined in Figure 1.

While a few recent studies (Claveau et al., 2021; Amin-Nejad et al., 2020; Iye et al., 2020) started investigating the viability of generated corpora as a replacement for the original datasets, we tackle

<sup>1</sup><https://twitter.com/en/privacy>



the following challenging settings:

**User-generated text:** Our target corpus to generate is patient-generated text, which depends highly on context. The textual nature, thus, becomes ungrammatical and fragmented.

**Multi-label condition:** A patient-generated message of the target corpus has multiple symptom labels. The generation model must understand the multiple conditions to create a correct message that expresses the corresponding symptoms.

The contributions of this paper are as follows:

- We propose a text generation approach for **patient-generated corpus** using pre-trained text generation models.
- We evaluated the proposed approach using an existing dataset in a **multi-label classification task** in the medical domain (that is, *MedWeb*).

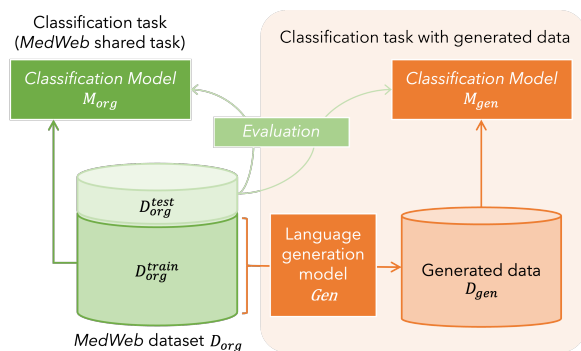


Figure 1: Flow of our experiment. Language generation models ( $Gens$ ) were trained using the original *MedWeb* training dataset  $D_{org}^{train}$  with different prompting methods. Classification models ( $M_{org}$  and  $M_{gen}$ s) are trained on the *MedWeb* training data  $D_{org}^{train}$  or the generated corpora ( $D_{gen}$ s), respectively. The *MedWeb* test data  $D_{org}^{test}$  is used for evaluation of both  $M_{org}$  and  $M_{gen}$ .

## 2. Related Work

### 2.1. Training Corpus Generation

Most studies on corpus generation are motivated by data augmentation rather than the creation of surrogate corpus. On the other hand, this study aims to replace the original corpus. Some studies evaluated surrogate corpus as part of their experiments in the medical and social media domains.

#### 2.1.1. Medical Domain

A few previous studies utilized a pre-trained language model (PLM) for text data generation in the medical domain. The generated text was used as a training corpus.

Amin-Nejad et al. (2020) utilized PLM for Electric Health Record (EHR) generation. They generated texts given the patient’s conditions, including demographic data, diagnosis, procedures, medications, microbiology tests, and laboratory tests. Generated data were evaluated based on the performance of unplanned readmission prediction and phenotype classification. Generated data accomplished comparable results with original data. In addition, this study showed that when combined with original data, using generated data improves the performance of classifiers in downstream tasks.

One of the advantages of automatic text generation is that it can generate a large number of text that are hard to sample from the real world. Motivated by the lack of data for rare symptoms, PLM was used for the generation of symptom definitions alongside with biomedical dictionary in Kim and Nakashole (2022). Given one symptom or two symptoms, definitions were generated. Generated definitions were used in vaccine side effect detection.

Pappas et al. (2022) also applied a pre-trained language generation model for data augmentation. They experimented with different data augmentation approaches for biomedical factoid question answering. As one of the approaches, they utilized question generation using fine-tuned T5 (Rafael et al., 2020). ALBERT (Lan et al., 2019) was used in the downstream task (biomedical factoid question answering). They found that adding generated data to original data results in slightly better performance than only using original data.

#### 2.1.2. Social Media Domain

The generation of social media posts can also be important because it also cannot be distributed for confidentiality reasons. Claveau et al. (2021) utilized a pre-trained language generation model (GPT-2) (Radford et al., 2019) for a surrogate training corpus generation. In downstream tasks, namely sentiment analysis on product reviews and fake news detection, the quality of generated corpora was evaluated. In neural classification approaches, they trained BERT (Devlin et al., 2019) as a classification model using 1) only original corpus, 2) only generated corpus, and 3) a mixture of original corpus and generated corpus. As a result, they found that 1) leads to better classifier performance than 2) and without filtering, 1) tends to perform better than 3).

## 2.2. Controllable Text Generation

Controllable text generation (CTG) is one of the hottest research topics in recent NLP. There could be many possible ways to achieve a patient-generated training corpus.

Zhang et al. (2022) gave a comprehensive survey on different approaches for CTG, which are 1) fine-tuning, 2) retrain/refactor PLMs and 3) post-process. Also, as described in Zhang et al. (2022), there are three major categories of fine-tuning approaches: prompt-based approaches, adapted module approaches, and reinforcement learning inspired approaches.

Jiang et al. (2021) showed that the performance of PLM is sensitive to prompt design modification. Liu et al. (2021a) provided a survey on different prompting approaches and those performances. In terms of methods for designing prompts and their human effort, most of the methods can be categorized into two: hand-crafted and automated search. Hand-crafted is the approach where humans design the suboptimal suitable prompt, while automated search is the approach where a suitable prompt is chosen automatically.

Among automated search approaches, Liu et al. (2021b) and Lester et al. (2021) experimented with soft prompts, tokens with trainable embeddings introduced in the fine-tuning stage. By inserting soft prompts, PLM can automatically search for an optimal prompt in the continuous space of all possible prompts. Wang et al. (2022) applied the soft prompt method for data augmentation in few-shot settings. Generated corpora were used for sequence labeling and sentence classification tasks.

T5 (Raffel et al., 2020) was used in corpus generation and BERT was used for downstream tasks. They found that adding soft prompts is effective in both downstream tasks. Also, Lester et al. (2021) experimented with sentinel tokens used in pre-training of T5. In pre-training of T5, unique sentinel tokens are used for marking masked spans in the input text. The task for T5 is to reconstruct these masked spans. They showed that in some experimental settings, using sentinel tokens in prompts is beneficial for the performance of PLM.

In the context of natural language generation, Schick and Schütze (2021) experimented with combining multiple instructions (prompts) through knowledge distillation. They evaluated the proposed automated search in a summarization task.

Although most prompt designing methods mentioned above are proposed in the context of Natural Language Understanding (NLU) tasks, we applied those methods to experiment with different prompting methods.

## 3. Dataset

*MedWeb* dataset consists of pseudo-posts for multi-label symptom classification.

To avoid privacy issues, the dataset was developed by crowdsourcing and not extracted from the actual X (previously Twitter) platform. In crowdsourcing, data were constructed from the symptom keywords (called “seed words”) that frequently appeared in real-world disease-related posts. Each post includes a description of whether the X (previously Twitter) user is experiencing a combination of symptoms, that is, a combination from a set “Influenza”, “Diarrhea”, “Hay fever”, “Cough”, “Headache”, “Fever”, “Runny nose”, and “Cold”. Examples of pseudo-posts in the dataset are listed in Table 1.

Classifying a combination of symptoms given a post can be considered a multi-label classification task among NLP tasks. In the following sections, we refer to one label pattern as a symptom combination.

*MedWeb* dataset contains 2,560 posts, and the ratio of training to test data is 75% (1,920 posts) to 25% (640 posts). “No symptom”, “Cold”, “Runny nose”, “Fever”, “Headache”, “Cough”, “Cold and Runny nose”, “Hay fever and Runny nose”, “Diarrhea”, and “Influenza and Fever” consist of 1,754 posts in total, which is 91% of all *MedWeb* training data.

## 4. Method

To investigate the viability of automatically generated corpora as surrogate training datasets, we trained; 1) the generation model (Section 4.1), which is utilized to create generated data (denoted with  $D_{gen}$ ) and 2) the classification models (Section 4.2), which are used to evaluate the quality of the generated texts.

### 4.1. Generation Models

We fine-tuned a text generation model  $Gen$  on the original *MedWeb* training dataset  $D_{org}^{train}$ . Following the previous study (Amin-Nejad et al., 2020), we decided to choose a fine-tuning approach among different controllable text generation (CTG) approaches. We used T5 (Raffel et al., 2020) as  $Gen$  in our method. Specifically, we fine-tuned the model pre-trained on Japanese corpora.<sup>2</sup> Since, we are handling a data-to-text generation task, encoder-decoder-type models are suitable. We considered T5 as representative of such models.

$Gen$  is fine-tuned in the following manner: given a symptom combination, it should generate a post

<sup>2</sup><https://huggingface.co/sonoisa/t5-base-japanese>

Post	Influenza	Diarrhea	Hay fever	Cough	Headache	Fever	Runny nose	Cold
風邪をひくと全身がだるくなる。 (The cold makes my whole body weak.)	-	-	-	-	-	-	-	+
花粉症の症状が出てたのは久しぶりだ。 (It's been a while since I've had allergy symptoms.)	-	-	+	-	-	-	+	-
インフルエンザのワクチン打ちに行ってきた。 (I went to get vaccinated for the flu.)	-	-	-	-	-	-	-	-

Table 1: Examples of *MedWeb* pseudo-posts. English translations are shown in the examples. + sign stands for the existence of the corresponding symptom in the user; - sign stands for the absence.

that expresses the corresponding symptoms. For example, when the model is given a combination of “fever and headache”, the generated post should say, for example, “I had a fever today. Bad headache too...”. Among various ways to achieve this conditional text generation, we chose prompting as a method because of its conceptual simplicity and relative efficiency in computational cost. Previous studies (Lester et al. (2021), Jiang et al. (2021), Liu et al. (2021b)) showed that giving appropriate instruction improves the performance of large generative PLMs on multiple tasks. Based on those work, we chose the following prompting methods:

**BASE (bs):** A hand-crafted symptom prompt.

As a baseline prompting method, we designed hand-crafted prompts. We represented a combination of symptoms by symptom name + the description of whether the symptom should appear or not. We put this expression at the beginning of the input sentence. To transform the instruction into the form of a question, we put “このTweetは？ (What is the tweet?)” at the end of the input sentence.

**SENTINEL (st):** A hand-crafted symptom prompt with a sentinel token.

We added a sentinel token (denoted with <X>) to the BASE prompt. Adding the sentinel token makes the task more similar to the task in pre-training of generative PLMs. We expected that catastrophic forgetting of the model could be avoided by making fine-tuning stage more similar to pre-training.

**SOFT (sf):** A hand-crafted symptom prompt with soft prompt tokens.

In our baseline design, we added soft prompt tokens (<s[id]>) that are trained simultaneously with the model parameters, inspired by Liu et al. (2021b) and Lester et al. (2021). This method was originally adopted to solve natural language understanding tasks. We assumed that this method would work for text generation too.

**SENTINEL+SOFT (st+sf):** A hand-crafted symptom prompt with soft prompts and a sentinel token.

We applied two modifications (the sentinel token and soft prompts addition) to the baseline design.

Examples of the four prompt designs given the same symptom combination are listed in Table 2. Finally, we created four *Gens* and 40 *D<sub>gen</sub>*s (10 datasets per prompt design) as we will explain in Section 5.1.

## 4.2. Classification Models

We evaluate the *D<sub>gen</sub>* quality on a classification task, the same as the *MedWeb* shared task.

We compared the evaluation results of different models using  $D_{org}^{test}$ . To do so, we trained the classification models using data created by different *Gens* (mentioned in Section 4.1). We also trained the classification model using  $D_{org}^{train}$ .

We trained *Gens* using the prompt BASE, SENTINEL, SOFT, and SENTINEL+SOFT, and generated posts (*D<sub>gen</sub>*s) from each *Gen*. Then we trained individual classification models on the different *D<sub>gen</sub>*s.  $M_{gen(bs)}$ ,  $M_{gen(st)}$ ,  $M_{gen(sf)}$ , and  $M_{gen(st+sf)}$  respectively denote these classification models.

The classification models used in our experiment are a pre-trained BERT model<sup>3</sup> with a linear transformation layer for the symptom combination classification. We trained our classification models on the task of symptom combination classification. Given a post, the model estimates the probabilities for eight symptom labels. When the output probability of a label surpassed a threshold, we considered the post to contain the corresponding symptom.

## 5. Experimental Setup

We evaluated the generated posts on  $D_{org}^{test}$  by measuring the performance of the different  $M_{gen}$ s compared to  $M_{org}$ . The flow of this experiment is shown in Figure 1.

The hyperparameters for the models  $M_{gen}$  and  $M_{org}$  are as follows:  $M_{gen}$  is trained for 20 epochs with a batch size of 32 using the Adam optimizer with 3e-4 learning rate, while  $M_{org}$  is trained for 24

<sup>3</sup><https://huggingface.co/cl-tohoku/bert-base-japanese>

Prompt design	Example prompt
BASE	インフルエンザの症状なし、...鼻水・鼻づまりの症状あり、風邪の症状なしのTweetは？ (What is the tweet in which the symptom of influenza doesn't appear, ...runny nose appears, and cold doesn't appear?)
SENTINEL	インフルエンザの症状なし、...鼻水・鼻づまりの症状あり、風邪の症状なしのTweetは？<X> (What is the tweet in which the symptom of influenza doesn't appear, ...runny nose appears, and cold doesn't appear?<X>)
SOFT	インフルエンザの症状なし、...鼻水・鼻づまりの症状あり、風邪の症状なしのTweetは？<s1>...<s99> (What is the tweet in which the symptom of influenza doesn't appear, ...runny nose appears, and cold doesn't appear?<s1>...<s99>)
SENTINEL+SOFT	インフルエンザの症状なし、...鼻水・鼻づまりの症状あり、風邪の症状なしのTweetは？<s1>...<s99><X> (What is the tweet in which the symptom of influenza doesn't appear, ...runny nose appears, and cold doesn't appear?<s1>...<s99><X>)

Table 2: Examples of different prompt designs, given “Runny nose” as the input symptom combination. <X> denotes the sentinel token and <s[id]> denotes soft prompt tokens.

epochs with a batch size of 8 using the AdamW optimizer with 1e-8 learning rate. As for the generation process, the hyperparameters include setting the number of beam search and beam groups equal to the number of posts for each label, a diversity penalty of 0.4, and a temperature value of 1.0.

### 5.1. Generation of Data

In order for the generated dataset to be distributable and comparable to  $D_{org}^{train}$ ,  $D_{gen}$  should meet the following conditions. 1) generated post should be de-identified and 2) the distribution of symptom combinations should be the same as that in  $D_{org}^{train}$ . To ensure that the settings of the task are the same for  $M_{gen}$ s and  $M_{org}$ , we made the distribution of conditional labels for a generation the same as that of the distribution of labels in the  $D_{org}^{train}$ . Because of these conditions, additional steps were needed to generate posts. Using the fine-tuned  $Gens$ , we first generated a larger number of posts for each symptom combination than that of  $D_{org}$ <sup>4</sup>. Then, we subtracted the subset of generated posts using an exact match search. The ratios of exactly matched posts generated by  $Gens$  using the prompt BASE, SENTINEL, SOFT, and SENTINEL+SOFT were 2.0%, 1.5%, 1.1%, and 1.1%, respectively<sup>5</sup>.

### 5.2. Evaluation

We evaluated classification models using the basic metrics used in the *MedWeb* shared task, and those are precision (micro average), recall (micro average), F1 score (micro average) and exact match accuracy.  $D_{org}^{test}$  is used for the evaluation of classification models. We created  $D_{gen}$  for each prompt 10 times and trained 10  $M_{gen}$  for each  $D_{gen}$ , resulting in 100 models with different parameters per prompt. Similarly, we trained

<sup>4</sup>Since we have no a priori knowledge about the number of exactly matched posts to be generated, we generated 35% more posts for each symptom combination and then randomly sampled them.

<sup>5</sup> $D_{gen}$  generated in this experiment is available at <https://github.com/seiji-shimizu/medweb-gen>

$M_{org}$  10 times on  $D_{org}^{train}$ . We obtained the scores (precision, recall, F1 score, and exact match accuracy) for each 100 models for  $M_{gen}$  and 10 models for  $M_{org}$ , and present the average scores as the scores for  $M_{gen(bs)}$ ,  $M_{gen(st)}$ ,  $M_{gen(sf)}$ ,  $M_{gen(st+sf)}$ , and  $M_{org}$ .

We evaluated the performance in the following three settings:

- Multi-symptom setting which is a usual multi-label classification (*multi* in short).
- Single-symptom classification using all data including multi-symptom posts for training data (*single\_all* in short).
- Single-symptom classification without using multi-symptom posts for training data (*single\_only* in short).

### 5.3. Manual Evaluation of Fluency

Additionally, we independently evaluated the fluency of the generated corpus with a Turing-test-like evaluation. To do so, we built the mixed 300 test set, which consisted of 150 generated posts and 150 original posts. We asked three Japanese annotators (two of them are nurses with more than 10 years of experience) to label the constructed data. Given a post, the annotators labeled whether the post is from original data or generated data.

## 6. Results

### 6.1. Results of Classification

The results are summarized in Tables 3, 4 and 5. Multiple  $M_{gen}$  with different prompting methods are denoted with  $M_{gen(bs)}$  (prompted by BASE),  $M_{gen(st)}$  (prompted by SENTINEL),  $M_{gen(sf)}$  (prompted by SOFT), and  $M_{gen(st+sf)}$  (prompted by SENTINEL+SOFT).

The results of *multi* are shown in Table 3. Among the four prompting methods,  $M_{gen(st)}$  gave the best result in terms of precision (0.757) and

Model	Prompt	Accuracy	F1 (micro avg.)	Recall (micro avg.)	Precision (micro avg.)
$M_{gen(bs)}$	BASE	0.632 (0.0105)	<b>0.759</b> (0.0120)	<b>0.762 (0.0080)</b>	0.756 (0.0064)
$M_{gen(st)}$	SENTINEL	<b>0.654</b> (0.0105)	0.757 (0.0114)	0.758 (0.0085)	<b>0.757</b> (0.0065)
$M_{gen(sf)}$	SOFT	0.629 (0.0119)	0.716 (0.0095)	0.695 (0.0201)	0.738 (0.0120)
$M_{gen(st+sf)}$	SENTINEL+SOFT	0.615 (0.0062)	0.692 (0.0080)	0.671 (0.0176)	0.716 (0.0078)
$M_{org}$	-	0.855 (0.0325)	0.910 (0.0126)	0.919 (0.0317)	0.901 (0.0170)

Table 3: Scores for *multi*. Values in parentheses represent standard deviations of the scores from 10 models. The highest accuracy and F1 scores are presented in bold.

Model	Prompt	Accuracy	F1 (micro avg.)	Recall (micro avg.)	Precision (micro avg.)
$M_{gen(bs)}$	BASE	0.682 (0.0166)	0.800 (0.0171)	0.867 (0.0134)	0.742 (0.0148)
$M_{gen(st)}$	SENTINEL	<b>0.701</b> (0.0132)	<b>0.807</b> (0.0146)	<b>0.875</b> (0.0096)	<b>0.750</b> (0.0099)
$M_{gen(sf)}$	SOFT	0.677 (0.0120)	0.726 (0.0125)	0.741 (0.0260)	0.712 (0.0145)
$M_{gen(st+sf)}$	SENTINEL+SOFT	0.679 (0.0062)	0.673 (0.0115)	0.770 (0.0200)	0.718 (0.0065)
$M_{org}$	-	0.861 (0.0085)	0.915 (0.0172)	0.938 (0.0110)	0.889 (0.0074)

Table 4: Scores for *single<sub>all</sub>*. Values in parentheses represent standard deviations of the scores from 10 models. The highest accuracy and F1 scores are presented in bold.

exact match accuracy (0.654).  $M_{gen(bs)}$  gave the best F1 score (0.759) and recall (0.762). Since the highest exact match accuracy is the hardest to achieve, we consider  $M_{gen(st)}$  to be the best performing  $M_{gen}$ . Compared to  $M_{org}$ , the best performing  $M_{gen}$  (that is,  $M_{gen(st)}$ ) could not achieve comparable scores.

The results of *single<sub>all</sub>* are shown in Table 4. In this evaluation, we used the same classification models (trained on  $D_{gen}$  and  $D_{org}^{train}$ ) as in *multi* and excluded posts with multiple symptoms only from the test data. Both  $M_{org}$  and  $M_{gen}$  performed slightly better compared to the results of *multi*. The gap between  $M_{org}$  and the best performing  $M_{gen}$  was still relatively large.

The results of *single<sub>only</sub>* are shown in Table 5. We only evaluated  $M_{gen(st)}$ , which was the best performing  $M_{gen}$  model in other experiments. The gap between  $M_{org}$  and  $M_{gen}$  became smaller in this experiment. Compared with the results in Table 4, the scores of the best performing  $M_{gen}$  increased by 0.0407 on average. On the other hand, the scores for  $M_{org}$  increased by 0.0218 on average.

## 6.2. Results of Manual Evaluation of Fluency

The average accuracy of the labeling was 0.648 and average inter-human Cohen’s kappa was 0.355. Both of those scores can be interpreted positively in the context of the Turing-test-like evaluation.

The low accuracy score suggests that the fluency of the generated corpus is relatively indistinguishable from that of the original corpus, and the task of labeling itself was difficult. Also, the low Cohen’s kappa coefficient suggests the subjectivity of labeling. These results can be evidence that the quality of the generated texts is almost equivalent to that of the original.

## 7. Discussions

In Section 6, we found that the gap between  $M_{org}$  and  $M_{gen}$  was prominent. We also found that the scores for  $M_{gen}$  improved, and the gap between  $M_{org}$  and  $M_{gen}$  became less prominent with training data without multiple symptoms (labels). This indicates that the quality of the generated multi-symptom posts is lower than that of single-symptom posts.

### 7.1. Difficulty in Multi-symptom Handling

To find out what is the main cause of the gap between  $M_{org}$  and  $M_{gen}$ , we further analyzed the quality of generated text from different symptom combinations.

We analyze the qualitative difference of generated posts with single-symptom labels and multi-symptom labels. Table 6 shows examples of posts labeled “*Influenza and fever*”, “*Hay fever and Runny nose*”, and “*Cold and Runny nose*” from  $D_{gen}$  generated from the prompt SENTINEL in the upper half of the table. As a comparison to multi-symptom labels mentioned above, we provide examples of posts labeled with “*Fever*”, “*Runny nose*”, and “*Cold*” from  $D_{gen}$  generated from the prompt SENTINEL in the lower half of the table. Note that, **for this qualitative error analysis**, even if only the expression of “*Influenza*” is in the generated post, we consider the generated post correct for “*Influenza and Fever*”. Similarly, for “*Hay fever and Runny nose*”, we consider the generated post correct, even if only the expression of “*Hay fever*” is in the generated post. The reason is that such posts (only including expression of “*Influenza*” or “*Hay fever*” and labeled “*Influenza and Fever*” and “*Hay fever and Runny nose*”) exist in  $D_{org}$ . We provide the correct examples in Table 6 (ids 1, 2, 6, 7, 11, and 12 for multi-symptom post generation and 16, 19, and 22

Model	Prompt	Accuracy	F1 (micro avg.)	Recall (micro avg.)	Precision (micro avg.)
$M_{gen(st)}$	SENTINEL	0.785 (0.0053)	0.837 (0.0111)	0.849 (0.0100)	0.825 (0.0041)
$M_{org}$	-	0.893 (0.0091)	0.932 (0.0122)	0.950 (0.0066)	0.915 (0.0067)

Table 5: Scores for *single<sub>only</sub>*. Values in parentheses represent standard deviations of the scores from 10 models.

Multi-symptom	id (correct or error)	Generated post
<i>Influenza and Fever</i>	1 (correct)	高熱が出て、インフルエンザにかかってしまった。(I got a high <b>fever</b> and caught <b>flu</b> )
	2 (correct)	昨日、インフルかかったから今日は休むわ。(I'll take a break today because I've got a <b>flu</b> yesterday.)
	3 (error)	高熱で仕事できないから今日は休むわ。(I can't work because of the high <b>fever</b> , so I'm taking a day off.)
	4 (error)	兄もインフルで寝込んでる。(My brother is also in bed with the <b>flu</b> .)
	5 (error)	風邪ひいた。(I've got a <b>cold</b> )
<i>Hay fever and Runny nose</i>	6 (correct)	花粉症で鼻水止まらない。(I can't stop <b>running nose</b> with <b>hay fever</b> )
	7 (correct)	まさか自分が花粉症になるとは思ってもなかったわ。(I never thought I would get <b>hay fever</b> .)
	8 (error)	鼻水止まん。(I can't stop my <b>runny nose</b> .)
	9 (error)	兄が花粉症で、今日も休むわ。(My brother has <b>hay fever</b> , so I will be absent today.)
	10 (error)	そんなにひどい頭痛は久しぶりだなあ。(I haven't had a <b>headache</b> that bad in a long time.)
<i>Cold and Runny nose</i>	11 (correct)	風邪で鼻水止まらない。(I have a <b>cold</b> and can't stop my <b>runny nose</b> .)
	12 (correct)	鼻風邪ひいた。薬飲んだら眠くなってきた。(I caught a <b>nose cold</b> . I fell asleep after taking medicine)
	13 (error)	風邪ひいた。(I caught a <b>cold</b> .)
	14 (error)	兄が鼻風邪で寝込んでる。(My brother is in bed with a <b>nose cold</b> .)
	15 (error)	花粉症のせいか、鼻風邪が治らない。
Single-symptom	id (correct or error)	Generated post
<i>Fever</i>	16 (correct)	今年一番の熱。今日は仕事休むわ。(The most horrible <b>fever</b> of the year. I'm off work today)
	17 (error)	弟が熱でて、仕事休むわ。(My brother has a <b>fever</b> , so I will be absent from work.)
	18 (error)	これって風邪？(Is this a <b>cold</b> ?)
<i>Runny nose</i>	19 (correct)	今日は鼻水止まらない。(My <b>nose</b> won't stop <b>running</b> today.)
	20 (error)	兄が鼻水でぐったりしてる。(My brother is exhausted from a <b>runny nose</b> .)
	21 (error)	日本には花粉症の人が多くいるんだってね。(There are many people with <b>hay fever</b> in Japan.)
<i>Cold</i>	22 (correct)	また、風邪ひいたかも。(I might have caught a <b>cold</b> , again.)
	23 (error)	中国で大流行した風邪が流行ってるらしいね。(It seems that there is an epidemic of <b>cold</b> in China.)
	24 (error)	日本の夏は本当に寒いんだけど・・・？(Summer in Japan is really cold, but...?)

Table 6: Examples of generated posts. The upper half is the examples of multi-symptom posts, and the lower half is examples of single-symptom posts

for single-symptom post generation).

We observed three types of typical errors.

**Shortage error:** The first type of error is a shortage of symptom expressions. In this type of error, even when given multiple conditions, such as “*Influenza and Fever*”, generated posts only contain a part of symptom expressions. Examples are posts in ids 3, 8, and 13 in Table 6.

**Out-of-user error:** The second type of error is non-user symptom expressions. In this type of error, the posts are referring to a symptom of a non-user, rather than that of the X (previously Twitter) user who wrote the post. Examples are posts in ids 4, 9, and 14 in Table 6.

**Label inconsistency error:** The third type of error is those that include other symptoms. Examples are posts in ids 5, 10, and 15 in Table 6. Note that no symptom at all is also classified into this type of error.

We observed the same types of errors as multi-symptom post generation in single-symptom post generation. Since single-symptom post generation is supposed to satisfy only one condition, the first type of error observed in multi-symptom was not

observed. Examples of the second type of error are shown in ids 17, 20, and 23 in Table 6, and the third type of error is shown in ids 18, 21, and 24.

## 7.2. Scores for Individual Symptom Combination

Since multi-symptom post generation has more complex conditions, more types of error can occur compared to single-symptom post generation. A possible reason for the lower scores for multi-symptoms is that the number of types of error in the multi-symptoms condition is larger than that in the single-symptom condition. We analyze the scores for individual symptom combinations.

The scores improved after the removal of multi-symptom labels. This suggests that the scores for multi-symptom labels are lower than those of single-symptom labels. Also, the difference in evaluation scores between *single<sub>all</sub>* and *single<sub>without</sub>* suggests that using generated multiple-symptoms posts in training had a negative influence even on classification of posts labeled with single and no symptom. We investigate those two assumptions by analyzing the scores for multi-symptom labels and single-symptom labels.

As mentioned in Section 3, multi-symptom la-

Model	Label	Accuracy	F1 (micro avg.)	Recall (micro avg.)	Precision (micro avg.)
$M_{gen(st)}$	<i>Influenza and Fever</i>	0.561 (0.0349)	0.701 (0.0055)	0.639 (0.0284)	0.776 (0.0218)
	<i>Hay fever and Runny nose</i>	0.874 (0.0126)	0.930 (0.0046)	0.912 (0.0111)	0.948 (0.0060)
	<i>Cold and Runny nose</i>	0.886 (0.0179)	0.952 (0.0020)	0.945 (0.0089)	0.960 (0.0047)
$M_{org}$	<i>Influenza and Fever</i>	0.754 (0.0346)	0.824 (0.0271)	0.779 (0.0325)	0.876 (0.0202)
	<i>Hay fever and Runny nose</i>	0.874 (0.0197)	0.904 (0.0171)	0.876 (0.0214)	0.934 (0.0122)
	<i>Cold and Runny nose</i>	<b>0.928</b> (0.0431)	<b>0.961</b> (0.0064)	0.956 (0.0307)	0.966 (0.0176)

Table 7: Metrics score for each multi-symptom label. Values in parentheses represent standard deviations of the scores from 10 models. The highest scores for accuracy and F1 are shown in bold.

Model	Training data	single-symptom labels				"No symptom"
		Accuracy	F1 (micro avg.)	Recall (micro avg.)	Precision (micro avg.)	Accuracy
$M_{gen(st)}$	MIX	0.900 (0.0208)	0.939 (0.0126)	0.931 (0.0137)	0.948 (0.0108)	0.562 (0.0161)
	SINGLE	0.907 (0.0179)	0.928 (0.0021)	0.907 (0.0179)	0.949 (0.0109)	0.723 (0.0161)
$M_{org}$	MIX	0.929 (0.0252)	0.953 (0.0148)	0.944 (0.0235)	0.963 (0.0133)	0.750 (0.0312)
	SINGLE	<b>0.950</b> (0.0207)	<b>0.962</b> (0.0160)	0.950 (0.0207)	0.974 (0.0110)	<b>0.802</b> (0.0245)

Table 8: Average metrics score for single-symptom labels and for "No symptom". Values in parentheses represent standard deviations of the scores from 10 models. The highest scores for accuracy and F1 are shown in bold.

bels with more than 30 posts are "Cold and Runny nose", "Hay fever and Runny nose", and "Influenza and Fever". We present the scores for those three multi-symptom labels in Table 7. As shown in Table 7, only the multi-symptom combination "Influenza and Fever" has apparently different results between  $M_{gen(st)}$  and  $M_{org}$ . This suggests that the other two combinations ("Hay fever and Runny nose" and "Cold and Runny nose") have less influence on the gap between  $M_{gen(st)}$  and  $M_{org}$  in overall scores, and improvement after removal of posts with multi-symptom labels.

Scores for single-symptom labels, we found that most of them have similar results. The scores of  $M_{org}$  and  $M_{gen(st)}$  from *single<sub>all</sub>* to *single<sub>without</sub>* tend to slightly increase compared to those of with multiple-symptom. We present the average scores of  $M_{gen(st)}$  and  $M_{org}$  in *single<sub>all</sub>* and *single<sub>without</sub>* for "Fever", "Runny nose", "Cold", "Diarrhea", "Headache" and "Cough" in Table 8.

"MIX" represents that model is trained on mixed data of multi, single, and no symptom posts, and "SINGLE" represents that model is trained on only single and no symptom posts. As shown in Table 8, the scores of the four models do not differ much.

Only the label "No symptom" had different results from others. Due to this, we present the results for "No symptom" in Table 8. As shown in Table 8, the exact match accuracy for "No symptom" improved after the removal of posts with multi-symptom labels.

To summarize, the scores for multi-symptoms are lower than those of single-symptoms in general. Especially, the scores for the label "Influenza and Fever" was the lowest among three symptom combinations.

### 7.3. Prompting Methods in Post Generation

The results showed that adding the sentinel token to the prompts effectively improves the classification performance. The improvement implies that the quality of the data generated by *Gen* can be improved with proper instructions.

We explore the soft-prompting method in our experiment. Despite the findings in the previous work, we did not see an improvement from the baseline method. Although we did not analyze the reason for the underperformance of the soft prompting method, it would be interesting to investigate how we can apply the prompting methods usually used in natural language understanding tasks to generation tasks (such as experimenting with the different numbers of soft prompts). As mentioned in (Schick and Schütze, 2021), methods to avoid overfitting are necessary for prompting methods in future work.

## 8. Conclusions and Future Work

This study experimented with a method for generating a distributable surrogate corpus and investigated its viability. We experimented with different prompting methods in fine-tuning the pre-trained language generation model and evaluated the quality of generated corpora by the classification task. The results showed that when generating posts that contain multiple symptoms, the generated corpora suffer from the problem of semantic inconsistency between the labels and the generated content. Still, if the surrogate corpus was used in simpler settings, the generated data would be comparatively viable as a training corpus, as demonstrated in a

single-symptom classification without using multi-symptom posts for training data.

In further pursuit of the research in this direction, we plan to 1) generate corpora in different languages than Japanese, namely *MedWeb*'s English and Chinese datasets, 2) measure the downstream performance in generating a larger amount of surrogate corpora than the original corpus, and 3) compare different models (other than T5) to investigate the impact of the choice of the model architecture on generation quality.

## 9. Limitations

Although we considered posts as a corpus in the medical domain, some clinical texts, such as discharge summaries, consist of much longer sentences. Since the pre-trained model used in this experiment accepts only less than 512 tokens, the low scalability to long texts, especially those with more than 512 tokens, is the limitation of this work.

## 10. Ethics Statement

The data used in this study, *MedWeb*, is deemed ethically sound. However, in the context of generating training data for medical NLP tasks, it is crucial to acknowledge the potential presence of errors in the generated data. Consequently, it is strongly advised against employing this data for tasks that have a direct impact on human life, such as automated diagnosis. Additionally, the study recognizes the possibility of the generated model memorizing and reproducing training data, emphasizing the importance of continuously integrating improvements based on relevant research findings.

## 11. Bibliographical References

Ali Amin-Nejad, Julia Ive, and Sumithra Velupillai. 2020. [Exploring Transformer Text Generation for Medical Dataset Augmentation](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4699–4708.

Eiji Aramaki, Shoko Wakamiya, Shuntaro Yada, and Yuta Nakamura. 2022. [Natural Language Processing: from Bedside to Everywhere](#). *Yearbook of Medical Informatics*.

Vincent Claveau, Antoine Chaffin, and Ewa Kijak. 2021. [Generating artificial texts as substitution or complement of training data](#). *arXiv preprint*, arXiv:2110.13016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of](#)

[Deep Bidirectional Transformers for Language Understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Udo Hahn and Michel Oleynik. 2020. [Medical Information Extraction in the Age of Deep Learning](#). *Yearbook of Medical Informatics*, 29(1):208–220.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, So-main Verma, Stephen Puntis, Rudolf N Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *NPJ digital medicine*, 3(1):69.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. [How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering](#). *Transactions of the Association for Computational Linguistics*, 9:962–977.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific data*, 3(160035):1–9.

Bosung Kim and Ndapa Nakashole. 2022. [Data augmentation for rare symptoms in vaccine side-effect detection](#). In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 310–315.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. [ALBERT: A Lite BERT for Self-supervised Learning of Language Representations](#). *arXiv preprint*, arXiv:1909.11942.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. [The Power of Scale for Parameter-Efficient Prompt Tuning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. [Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing](#). *arXiv preprint*, arXiv:2107.13586.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. [GPT Understands, Too](#). *arXiv preprint*, arXiv:2103.10385.



- Dimitris Pappas, Prodromos Malakasiotis, and Ion Androutopoulos. 2022. [Data Augmentation for Biomedical Factoid Question Answering](#). *arXiv preprint*, arXiv:2204.04711.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. [Language models are unsupervised multitask learners](#). *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. [Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Timo Schick and Hinrich Schütze. 2021. [Few-Shot Text Generation with Natural Language Instructions](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 390–402.
- Tawanda Sibanda and Ozlem Uzuner. 2006. [Role of local context in automatic deidentification of ungrammatical, fragmented text](#). In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 65–73.
- Özlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. [Evaluating the State-of-the-Art in Automatic De-identification](#). *Journal of the American Medical Informatics Association*, 14(5):550–563.
- Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. [PromDA: Prompt-based Data Augmentation for Low-Resource NLU Tasks](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4255.
- Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. [A survey of controllable text generation using transformer-based pre-trained language models](#). *arXiv preprint*, arXiv:2201.05337.
- Zheming Zuo, Matthew Watson, David Budgen, Robert Hall, Chris Kennelly, and Noura Al Moubayed. 2021. [Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study](#). *JMIR Med Inform*, 9(10):e29871.

## 12. Language Resource References

- Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma, and Eiji Aramaki. 2019.

# CliniRes: Publicly Available Mapping of Clinical Lexical Resources

Elena Zotova\*<sup>†</sup>, Montse Cuadros\*, German Rigau<sup>†‡</sup>

\*SNLT group at Vicomtech Foundation, Basque Research and Technology Alliance (BRTA)  
Mikeletegi Pasealekua 57, 20009, Donostia/San-Sebastián, Spain  
{ezotova, mcuadros}@vicomtech.org

<sup>†</sup>Department of Languages and Computer Systems, University of the Basque Country (UPV-EHU)  
Paseo Manuel de Lardizábal, 1, 20018, Donostia/San-Sebastián, Spain

<sup>‡</sup> HiTZ Basque Center for Language Technologies  
german.rigau@ehu.eus

## Abstract

This paper presents a human-readable resource for mapping identifiers from various clinical knowledge bases. This resource is a version of UMLS Metathesaurus enriched with WordNet 3.0 and 3.1 synsets, Wikidata items with their clinical identifiers, SNOMED CT to ICD-10 mapping and Spanish ICD-10 codes description. The main goal of the presented resource is to provide semantic interoperability across the clinical concepts from various knowledge bases and facilitate its integration into mapping tools. As a side effect, the mapping enriches already annotated medical corpora for entity recognition or entity linking tasks with new labels. We experiment with entity linking task, using a corpus annotated both manually and with the mapping method and demonstrate that a semi-automatic way of annotation may be used to create new labels. The resource is available in English and Spanish, although all languages of UMLS may be extracted. The new lexical resource is publicly available.

**Keywords:** clinical coding, entity linking, data interoperability, lexical resource, clinical NLP

## 1. Introduction

Annotation of training corpora for clinical coding, clinical concepts detection, entity disambiguation and entity linking tasks is very expensive in expertise and time. Considering that most clinical concepts are transferable across various knowledge bases, terminologies, lexicons and languages, we hypothesise that we can transfer one type of annotated code to another. For this purpose, we create CliniRes—a mapping human-readable resource to get related synonyms in various clinical lexicons so that target entities or concepts can be annotated in different clinical notations. This resource permits to align different types of clinical identifiers (IDs, codes) from different knowledge bases (KB) such as UMLS (Bodenreider, 2004), ICD-10 (World Health Organization (WHO), 2004), SNOMED CT (Donnelly et al., 2006) and others. Also, we enrich the resource with lexical resources, such as Wikidata items (Vrandečić and Krötzsch, 2014) and Wordnet synsets (Fellbaum, 2005). This allows to make clinical codes inter-operable, to use it in data annotation or other applications where clinical codes are involved. Moreover, it allows us to enrich manually annotated corpora with extra clinical codes and to obtain multilingual inter-operable corpora annotated with various coding notations. For instance, if we have a corpus annotated in UMLS codes we can map each code to SNOMED CT codes in order to derive automatically a new version of the corpus with SNOMED CT annotations. And vice versa, corpus annotated with SNOMED CT codes can be used to derive automatically new cor-

pora annotated with UMLS codes, semantic types or groups.

This research is an extension of previously published works (Zotova et al., 2022, 2023a) where we described ClinIDMap<sup>1</sup>, a clinical IDs mapping tool with the presented database integration. The functionality of the mapping application includes mapping of a source code (may be UMLS CUI, SNOMED CT, ICD-10-CM and ICD-10-PCS) to the clinical IDs and lexical resources such as Wikidata, Wikipedia and WordNet, including WordNet domains. It also allows updating of the database, as new versions of the ontologies are released yearly and Wikidata annotations are added regularly. The application is developed as REST API, accepts queries in JSON format, the database is indexed in Elasticsearch (Lucene). The source code is Dockerized, so it can be easily deployed.

The main contribution of this work is CliniRes— an alignment resource for mapping of clinical identifiers based on UMLS Metathesaurus, enriched with Spanish version of ICD-10, WordNet, and Wikidata items and annotations. It is ready to be integrated into a mapping application or be processed for synonym and annotation generation. This database is available under the licence of UMLS and SNOMED CT<sup>2</sup>. Also, this paper contributes to the methods of semi-automatic corpus annotation in clinical cor-

<sup>1</sup><https://github.com/Vicomtech/ClinIDMap>

<sup>2</sup>[https://www.nlm.nih.gov/research/umls/knowledge\\_sources/metathesaurus/release/license\\_agreement\\_snomed.html](https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/license_agreement_snomed.html)

pora showing that the entity linking systems trained with the corpora annotated with this method, perform with the same accuracy as the systems trained with gold-standard corpora.

This paper is organized as follows. In Section 2 we briefly describe the background of clinical IDs mapping. Section 3 is dedicated to the knowledge bases and the mapping method. In Section 4 we give the details about the experiment with entity linking task done with the code mapping method. Finally, Section 5 concludes the work and discusses the future work in this topic.

## 2. Related Work

Two main parts of clinical codes mapping exist: (1) concept alignment, or ontology alignment (also known as ontology matching); (2) applications which use the concept mapping to enrich biomedical text or extract these concepts.

**Ontology matching.** The aim of ontology matching is to find semantically related entities in knowledge bases of different notations. For instance, the OAEI Campaign (Ontology Alignment Evaluation Initiative)<sup>3</sup> organizes every year an ontology matching evaluation shared task. The applied methods combine multiple strategies such as lexical matching, structural matching, logical reasoning, using background knowledge such as general purpose lexical resources, automatic translation and pretrained language models (Portisch et al., 2022; Wang et al., 2021). For instance, WordNet graphs were broadly used to clinical ontology matching (Lin and Sandkuhl, 2008) for measuring semantic similarity between the concepts (Pedersen et al., 2007). Some attempts to integrate WordNet to the clinical knowledge bases (Smith and Fellbaum, 2004) were made. Nevertheless, we should admit that the most of the studies are done with the resources in English. Novel machine learning and deep learning methods, such as generative adversarial networks, are also applied to ontology alignment (Chen et al., 2021; Kim et al., 2017).

**Concept mapping applications.** To our knowledge, there are not many open-source applications for concept mapping, especially for languages different from English. One of them is I-MAGIC, an application, implemented by US National Library of Medicine, that visualises clinical ID mappings. A demo version of the application is also available<sup>4</sup>. Using the rule-based SNOMED-CT to ICD-10-CM Mapping (Fung and Xu, 2012), the algorithm determines whether a valid ICD-10-CM code can be found based on the SNOMED-CT term and patient

context information (age and gender). The application allows one to search a term in SNOMED-CT vocabulary, however, it is limited to a lexical match. The tool does not consider synonyms, nor other languages other than English, and its code is not open-source.

Most applications for clinical coding are designed to enrich clinical text with clinical concepts and relations. MetaMap<sup>5</sup> (Aronson and Lang, 2010; Aronson, 2001) is an application for mapping biomedical text to the UMLS Metathesaurus or, equivalently, to discover UMLS concepts referred in the text. MetaMap uses a knowledge-intensive approach based on symbolic, NLP and computational-linguistic techniques to provide a link between the text of biomedical literature and the KB, including synonymy relationships, embedded in the Metathesaurus. The input of the application is English text. It is based on a lexical lookup of input words. Another example is CLAMP (Soysal et al., 2017), which takes two approaches: a machine learning using Conditional Random Field and a dictionary-based approach, which maps mentions to standardised ontologies. Apache cTAKES (Bodenreider, 2004) uses a dictionary look-up in unstructured clinical text, detects named entities and each mention is mapped to a UMLS concept.

Some applications are also private, as they are developed by big tech companies. Spark NLP<sup>6</sup> and Amazon Comprehend Medical<sup>7</sup> offer service for mapping clinical findings to ICD-10-CM, SNOMED CT and other codes, in addition to entities and relations extraction.

There are also studies in topic of UMLS and Wikipedia connection, for instance, Rahimi et al. (2020) proposes to match UMLS concepts to Wikidata using a cross-lingual neural re-ranking model which is fine-tuned as a pair binary classification model aimed to categorize if a pair of texts is similar or not. As the UMLS descriptions are brief and the medical entity pages in Wikipedia provide detailed descriptions (also enriched with the Wikidata knowledge graph), they use the UMLS concept description to query the Wikidata entity aliases to retrieve the best matching Wikipedia pages.

## 3. Method

This section describes the knowledge bases and lexical resources used to create CliniRes and the method of mapping of clinical identifiers.

<sup>5</sup>[https://www.nlm.nih.gov/research/umls/implementation\\_resources/metamap.html](https://www.nlm.nih.gov/research/umls/implementation_resources/metamap.html)

<sup>6</sup>[https://demo.johnsnowlabs.com/healthcare/ER\\_ICD10\\_CM/](https://demo.johnsnowlabs.com/healthcare/ER_ICD10_CM/)

<sup>7</sup><https://aws.amazon.com/es/comprehend/medical/>

<sup>3</sup><http://oaei.ontologymatching.org/2023/>

<sup>4</sup><https://imagic.nlm.nih.gov/imagic/code/map>

### 3.1. Knowledge Bases

To interconnect the different identifiers from the knowledge bases of interest, we use the following existing KBs and mappings created by clinical experts.

**UMLS Metathesaurus**<sup>8</sup>. This database has been derived from the 2023AB UMLS Metathesaurus Files which contains approximately 3.15 million concepts from 220 source vocabularies, including ICD-10, MeSH, and SNOMED-CT, Hierarchies, definitions, and other relationships and attributes. The Metathesaurus is the biggest component of the UMLS. It is organised as a set of Concept Unique Identifiers (CUI), which links all the names from the source vocabularies with the same meaning (synonyms) in various languages. The Metathesaurus assigns several types of unique, permanent identifiers to the concepts and concept names it contains, in addition to retaining all identifiers present in the source vocabularies. The Metathesaurus concept structure includes concept names, their identifiers, and key characteristics of these concept names (e.g., language, vocabulary source, name type). The majority of the concept descriptions are short, less than one sentence. The entire concept structure appears in a single file in the Rich Release Format (MRCONSO.RRF). The distribution across the non-English languages is not proportional, as we can see in Table 1 there are significantly less concepts and synonyms in Spanish than in English.

**The Semantic Network and Semantic Groups** from UMLS is used to map semantic groups of each CUI. The Semantic Network consists of a set of broad subject categories, or Semantic Types, that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus. The concepts are also grouped according to the semantic types assigned to them. For certain purposes, however, an even smaller and coarser-grained set of semantic type groupings may be desirable. The following principles were used to design the groupings: semantic validity, parsimony, completeness, exclusivity, naturalness, and utility. The semantic groups provide a partition of the UMLS Metathesaurus for 99.5% of the concepts. Examples of semantic groups are Organisms, Anatomical structures, Biological functions, Chemicals, Events, Physical objects, Concepts or Ideas. These types are suitable for corpus annotation and training sequence labelling models and further linking to UMLS.

**SNOMED-CT to ICD-10-CM Mapping**<sup>9</sup>. The main

purpose of the SNOMED-CT to ICD-10-CM mapping is to support semi-automated generation of ICD-10-CM codes from clinical data encoded in SNOMED-CT for reimbursement and statistical purposes. It is designed as a directed set of relationships from SNOMED-CT source concepts to ICD-10-CM target classification codes. This mapping is curated by trained terminology specialists, and it is more comprehensive than the Metathesaurus CUI linking. About a third of all active SNOMED-CT concepts are within the scope of the mapping, about 125,000 SNOMED-CT codes from the international version are mapped to ICD-10-CM codes. About 57,000 codes from the Spanish SNOMED-CT are included in the mapping (around 30% of all Spanish SNOMED-CT codes). Due to the differences in granularity, emphasis and organising principles between SNOMED-CT and ICD-10-CM, it is not always possible to have one-to-one mappings between a SNOMED-CT concept and an ICD-10-CM code, moreover, not all ICD-10-CM codes will appear as targets.

**ICD-10-CM** (International Statistical Classification of Diseases and Related Health Problems) establishes a standardized coding that allows the statistical analysis of mortality and morbidity of patients in healthcare services. The corresponding Spanish version is called CIE-10-ES and it consists of 100,158 codes, which are organised hierarchically. We use the official Spanish version of the CIE-10 from January 2022.

**ICD-10-PCS** (Procedure Coding System) is an international system of medical classification used for procedural coding, it consists of 80,266 codes, organised hierarchically. We use the official Spanish version of the ICD-10-PCS from March, 2022<sup>10</sup>.

**Wikidata**<sup>11</sup> (Vrandečić and Krötzsch, 2014) is a free and open knowledge base that can be consulted and edited by both humans and machines. Wikidata is a central repository for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wiktionary, Wikisource, and others. The Wikidata repository consists mainly of items, each with a label, a description and several aliases. Wikidata items related to clinical concepts are manually annotated with UMLS ID (CUI), Medical Subject Headings (MeSH) (Rogers, 1963), NCBI<sup>12</sup> (biomedical and genomic database) and other clinical taxonomies, so we can search items in Wikidata by these identifiers and extract the corresponding articles in all available languages.

**WordNet 3.1**<sup>13</sup> (Fellbaum, 2005) is the latest ver-

<sup>8</sup>[https://www.nlm.nih.gov/research/umls/knowledge\\_sources/metathesaurus/index.html](https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html)

<sup>9</sup>[https://www.nlm.nih.gov/research/umls/mapping\\_projects/snomedct\\_to\\_icd10cm.html](https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html)

<sup>10</sup><https://www.sanidad.gob.es/fr/estadEstudios/estadisticas/normalizacion/home.htm>

<sup>11</sup><https://www.wikidata.org>

<sup>12</sup><https://www.ncbi.nlm.nih.gov/>

<sup>13</sup><https://wordnet.princeton.edu/>

Code	Num
Total rows	13,501,908
Unique CUIs	3,145,136
ENG CUIs	8,510,801
ENG Unique	3,144,365
SPA CUIs	1,371,376
SPA unique	491,713
CUIs with SNOMED mapping	359,757
SNOMED codes with CUI mapping	367,700

Table 1: Number of concepts in UMLS

sion of a lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked using conceptual-semantic and lexical relations. The WordNet also contains senses which are discrete representations of each aspect of the meaning of words. In the database, each sense has its unique sense key index (SKI) which provides a method for accessing synsets and word senses in the WordNet database. This version contains 155,327 words organised in 175,979 synsets for 207,016 word-sense pairs.

**WordNet 3.0**<sup>14</sup> (Fellbaum, 2005) is the previous release of the lexical database. The WordNet 3.0 release has 117,798 nouns, 11,529 verbs, 22,479 adjectives, and 4,481 adverbs. The average noun has 1.23 senses, and the average verb has 2.16 senses. In total, there are 206,941 sense keys. As far as we know, no direct mapping between WN 3.0 and WN 3.1 exists, so we map the WordNet 3.1 to the WordNet 3.0 on the base sense key index. There are about 1,000 senses in all WordNet graph that cannot be transferred.

### 3.2. Code Mapping

To generate the enriched version of UMLS Thesaurus, we extracted all the Wikidata items annotated with UMLS CUI, NCBI, Wordnet 3.1, ICD-10 and SNOMED CT identifiers (updated on October 5, 2023). The Wikidata items are manually annotated by Wikidata experts. As shown in Table 2, there are about 860,000 items labelled with clinical IDs, the largest number is for UMLS CUI (about 86%), followed by NCBI IDs, and a smaller proportion, about 4% of items is annotated with WordNet synsets. Some of the Wikidata items are annotated with multiple WordNet synsets, up to six per item, in the table they are separated with a blank space. The less present identifiers in Wikidata items are SNOMED CT and ICD-10-PCS.

We merge all the tables databases described in Subsection 3.1 to the UMLS Thesaurus based on the CUI, SNOMED CT and ICD-10 codes. As a

Source	Num
UMLS CUI	742,537
NCBI	623,397
Wordnet 3.1	31,897
Wordnet 3.0	31,884
MeSH	46,023
ICD10	7,650
ICD10CM	15,618
ICD10PCS	74
SNOMED_CT	1,579
Total Wikidata items	860,245

Table 2: Number of Wikidata items annotated with clinical codes extracted from Wikidata database.

result, we obtain a large matrix of 37 columns and 15,945,228 rows where the first 18 columns are from the original UMLS table, and the rest of the columns are added through the SNOMED CT to ICD-10 mapping and through the Wikidata annotations. Figure 1 schematically depicts the method of mapping, where we can see how the identifiers are connected. Wikidata annotations and Semantic groups and types are connected through the UMLS CUI, the SNOMED CT to ICD-10 mapping is based on SNOMED CT codes, and ICD-10 codes presented in UMLS are extended to their Spanish definitions.

SNOMED CT to ICD-10 mapping adds more mappings between CUI and ICD-10, because the Spanish version of ICD-10 is not presented in UMLS, while SNOMED CT is presented. Spanish descriptions of ICD-10 codes are added, too. The codes extracted from Wikidata are marked as `_WIKI`, and this code may be different to the UMLS mapping, because of the manual expertise of the Wikidata editors. As the table is large, the definitions of the columns are detailed in Appendix A, Table 6. We also encourage the reader to see the sample of the resulting table in the GitHub repository<sup>15</sup>.

This large matrix allows us to extract all related information based on any ID, WordNet sense or Wikidata item and then extend to more details. Wiki-

<sup>14</sup><https://wordnetcode.princeton.edu/3.0>

<sup>15</sup><https://github.com/Vicomtech/ClinIDMap/tree/master/LREC2024/samples>

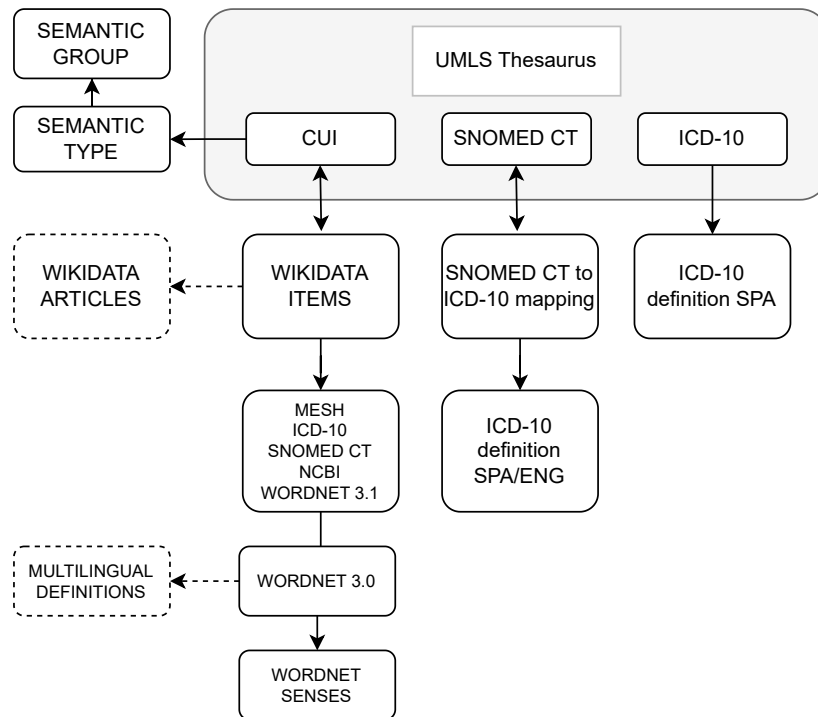


Figure 1: Scheme of clinical resources mapping, where they are connected by CUI, SNOMED CT or ICD-10 IDs. Dotted lines show a possible connection to more information about each concept.

data annotations allow us to derive further details, such as extracting more information about the item: its Wikipedia articles in all available languages, aliases, synonyms and other annotations. With that, short KB descriptions are extended to lexical resources, encyclopedia definitions and contexts. Moreover, Wikidata items exist, annotated both with CUI and ICD-10, so that we can consider it to be new code mapping. This provides 8,698 ICD-10 and 5,746 codes present in Wikidata annotations but not in UMLS mapping.

#### 4. Entity Linking Experiment

To show that the codes in the clinical KBs are interoperable, we have already experimented with named entity recognition task to detect diagnosis and procedures or semantic types from UMLS notation in previous study, the work is described in our previous publication (Zotova et al., 2022). Now, we experiment with the entity linking task. Entity linking, or entity normalisation, is the key technology enabling semantic applications and informatics pipelines in the biomedical domain. This task aims to assign an identifier from clinical KB to the text span from clinical text written in natural language. In our case, both texts and KBs are in Spanish.

For the experiment, we use two annotated datasets of similar nature. These datasets consist of clinical case reports—a type of textual genre in medicine that describes a patient’s medical history, symp-

toms, diagnosis, and treatment in detail. Both datasets are prepared for the entity linking task. Short descriptions of each corpus are below.

- MedProcNER (Lima-Lopez et al., 2023) is a collection of 1,000 clinical case reports written in Spanish, from which 750 documents are prepared for system training and 250 are for testing. In the train set, 4,857 text spans are manually annotated with SNOMED CT codes; 1,829 are unique, and some of the codes are composite, where two or more codes overlap. All codes are also annotated as procedures.
- CodiEsp 2020 (Miranda-Escalada et al., 2020) is a collection of 1,000 clinical case reports written in Spanish, where 750 documents are prepared for training purposes and 250 documents are reserved for testing. All documents were manually annotated by professional clinical coders with codes from the Spanish version of ICD-10 (procedure and diagnosis), and contain 3,427 unique codes, 2,557 of them are diagnoses and 870 codes are procedures. There are overlapping codes, too. The train set consists of 13,658 annotated text spans.

We evaluate unsupervised systems and use the whole training subset, without splitting it to the development subset. With the mapping tool, we transfer the gold-standard annotations (SNOMED CT

and ICD-10) to UMLS CUI and obtain corpora annotated with new codes. There are limitations of exact mapping because of the granularity of the ontologies and the annotations guides; some of SNOMED CT or ICD-10 codes have no direct mapping to another vocabulary. In MedProcNER corpus, from 4,857 entities there are 176 codes which cannot be transferred to CUI, in CodiEsp corpus 100% of ICD-10 codes are transferable to CUIs. At the same time, one SNOMED CT or ICD-10 code may be mapped to various CUIs, to simplify the experiment and make it comparable with the single SNOMED CT annotations we take only one CUI, the first in the database.

The example from MedProcNER corpus below shows a case of codes mapping from SNOMED CT to UMLS CUI. The term "Serologías específicas para Brucella" (*Specific serologies for Brucella*) is annotated with SNOMED CT code 104279004 "prueba de anticuerpos anti-Brucella" (*anti-Brucella antibody test*) and mapped to the CUI C0523269.

Durante el ingreso se solicitan Hemocultivos: positivo para Brucella y **Serologías específicas para Brucella**: Rosa de Bengala +++; <...>

*During admission, blood cultures are requested: positive for Brucella and **specific serologies for Brucella**: Rosa de Bengala +++; <...>*

Our approach to entity linking task is based on Semantic Text Similarity (STS) techniques. STS determines how similar two textual documents are by measuring their degree of semantic closeness. Semantic search is based on STS, allowing retrieval of relevant text results beyond mere lexical matching. The main concepts of semantic search are query, collection of documents (database), and degree of relevance between a query and retrieved documents. There are different methods of measuring the degree of relevance and relatedness of two pieces of text—cosine distance, inner product, etc. We implement the following two unsupervised approaches.

- Statistical method with BM25 algorithm (Robertson et al., 1998). This function ranks a set of documents based on the query terms appearing in each document, regardless of their proximity, and it works on the concept of bag-of-words and TF-IDF. In a search time all the documents and a query are tokenized by white space and lower-cased.
- Transformer-based Semantic Search based on pre-trained Transformer models (Vaswani et al., 2017) to obtain the corresponding embeddings (multidimensional vectors) and compute the score using a similarity metric, in this

case it is normalised inner product. This type of approach is implemented with HuggingFace (Wolf et al., 2020) and FAISS framework (Johnson et al., 2021).

The semantic search involves embedding all entries (sentences, documents, or, in this case, KB code descriptions) into a single vector space. At search time, the query, represented by a texts span from a clinical narrative, is also embedded into the same vector space. This allows a direct comparison of vectors using cosine distance between the vectors. The closest document, in our case, CUI, ICD-10 or SNOMED CT description, is linked to our query and the code assigned to this document is returned as a prediction of linked identifier. In this case we encode the texts with SapBERT-XLM-R-large model (Liu et al., 2021), as it is a XLM-RoBERTa-large model (Conneau et al., 2019) trained on the descriptions from UMLS Thesaurus and brings the domain knowledge to the entity linking system. An embedding dimension of 1024 is enough to encode all the terminology and corpus entities without truncation. [CLS] token of the transformer's architecture is used for the vector representation of a text.

To reduce the search space and make it comparable to the original task, developed with clinical experts (search in SNOMED CT and ICD-10 databases in Spanish), we filter UMLS Thesaurus to Spanish terms only, lowercase the descriptions, and obtain about 1.28 million Spanish synonyms from different vocabularies related to approximately 490,000 unique CUIs, still being the largest of three collection of documents to search. Table 3 shows the exact number of the KBs for entity linking, where we can see that the size of SNOMED CT and ICD-10 is comparable, but UMLS is much larger.

KB	Num
UMLS CUI	1,283,535
SNOMED_CT	242,228
ICD10CM+ICD10PCS	180,424

Table 3: Size of knowledge bases for entity linking.

Annotation	Method	Accuracy
SNOMED CT (gold)	SapBERT	43.44
	BM25	19.96
UMLS CUI (map)	SapBERT	34.78
	BM25	24.27

Table 4: Performance of semantic search approach on MedProcNER corpus.

The results of the search methods, as depicted in Tables 4 for MedProcNer corpus and 5 for CodiEsp corpus, are comparable across the gold-standard and mapped corpus, regardless of different size

Annotation	Method	Accuracy
ICD-10 (gold)	SapBERT	29.62
	BM25	10.57
UMLS CUI (map)	SapBERT	28.76
	BM25	25.75

Table 5: Performance of semantic search approach on CodiEsp-2020 corpus.

of the vector space and different coding systems. Transformer based system in case of MedProcNer performs 9 points better on the gold-standard corpus, but we should admit that the database to search in is much bigger— 1,3 millions entries in UMLS versus 242,000 in SNOMED CT. The BM25 model perform 4 point better in case of mapped labels (CUI). It can be explained with the fact that UMLS contains various vocabularies which could be closer lexically to the corpus. As the STS method based on distance measure, it highly depends on the number of documents in the collection to search in, the less is the collection, the easier is the retrieving task.

Entity linking on CodiEsp also shows very similar result in both systems, we observe less than one point difference in transformer-based semantic search with SapBERT model. The BM25 method, which performs in the gold-standard corpus worse. It can be explained by the broader variety of synonyms in UMLS, which represents better the lexical content of the corpus.

These STS models do not perform with the state of the art accuracy scores, and we do not compete with these scores, moreover the original task designed for these corpora, based first on the named entity recognition task, and then, the recognised entities should be linked to the KBs. We skip the named entity recognition step and experiment with entity linking only, where the exact text span is already known and manually annotated. That is why our results are nor comparable to the previous studies. In summary, we can conclude that the new models and corpora are quite inter-operable with respect the different coding systems.

## 5. Conclusions and Future Wok

In this paper we described a human-readable database for interoperability between clinical concepts of various knowledge bases. For this, we explained how we enriched the UMLS Thesaurus with Wikidata items, WordNet senses and SNOMED CT to ICD-10 mappings, we also added Spanish definitions of ICD-10 codes present in this resource. This resource is ready to be integrated into any application or be used for clinical synonyms generated. We demonstrate the use of the resulting resource in the mapping tool which is publicly available as opens-source, both the code and the demo-version

of the API.

We experimented with entity linking task on the corpora annotated with different coding systems, showing that the labels obtained with the mapping method can be used to build new entity linking or information retrieval systems, as the results of the entity linking systems are comparable.

As future work we see the experimentation on deep learning methods and large language models for mapping between English and multilingual concepts, paying special attention to underrepresented in UMLS languages. As we mentioned in Subsection 3.1, the distribution of the UMLS concepts and vocabularies are disproportional for non-English languages. We see the opportunity to use the novel approaches to contribute to creation of clinical terminologies and create background knowledge for concept matching in multilingual setting. We also plan to experiment with more methods for entity linking in clinical databases and lexical resources.

## 6. Bibliographical References

- Alan Aronson. 2001. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 2001:17–21.
- Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.
- Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.
- Jiaoyan Chen, Ernesto Jiménez-Ruiz, Ian Horrocks, Denvar Antonyrajah, Ali Hadian, and Jaehun Lee. 2021. Augmenting ontology alignment by semantic embedding and distant supervision. In *European Semantic Web Conference*, pages 392–408. Springer.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.
- Kevin Donnelly et al. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279.
- Christiane Fellbaum. 2005. Wordnet and wordnets. In *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford.



- Kin Wah Fung and Junchuan Xu. 2012. Synergism between the Mapping Projects from SNOMED CT to ICD-10 and ICD-10-CM. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2012:218–227.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Sun Kim, Nicolas Fiorini, W. John Wilbur, and Zhiyong Lu. 2017. Bridging the Gap: Incorporating a Semantic Similarity Measure for Effectively Mapping PubMed Queries to Documents. *Journal of Biomedical Informatics*, 75:122–127.
- Salvador Lima-Lopez, Eulaia Farre-Maduell, Luis Gasco, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2023. Overview of MedProcNER task on medical procedure detection and entity linking at BioASQ 2023. In *Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum*.
- Feiyu Lin and Kurt Sandkuhl. 2008. [A survey of exploiting wordnet in ontology matching](#). *International Federation for Information Processing Digital Library; ARTIFICIAL INTELLIGENCE IN THEORY AND PRACTICE II* ;, 276:341–350.
- Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Learning Domain-Specialised Representations for Cross-Lingual Biomedical Entity Linking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 565–574. Association for Computational Linguistics.
- George A Miller. 1998. *WordNet: An Electronic Lexical Database*. MIT press.
- Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. 2020. Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.
- Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. 2007. Measures of Semantic Similarity and Relatedness in The Biomedical FDomain. *Journal of biomedical informatics*, 40(3):288–299.
- Jan Portisch, Michael Hladik, and Heiko Paulheim. 2022. Background knowledge in ontology matching: A survey. *Semantic Web*, pages 1–55.
- Afshin Rahimi, Timothy Baldwin, and Karin Verspoor. 2020. WikiUMLS: Aligning UMLS to Wikipedia via cross-lingual neural ranking. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5957–5962, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Stephen E. Robertson, Steve Walker, and Micheline Hancock-Beaulieu. 1998. Okapi at TREC-7: Automatic Ad Hoc, Filtering, VLC and Interactive. In *Proceedings of The Seventh Text REtrieval Conference, TREC 1998, Gaithersburg, Maryland, USA, November 9-11, 1998*, volume 500-242 of *NIST Special Publication*, pages 199–210. National Institute of Standards and Technology (NIST).
- Frank Rogers. 1963. Medical subject headings. *Bulletin of the Medical Library Association*, 51:114–116.
- Barry Smith and Christiane Fellbaum. 2004. Medical WordNet: A New Methodology for the Construction and Validation of Information Resources for Consumer Health. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING '04*, page 371–es, USA. Association for Computational Linguistics.
- Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei V. S. Pakhomov, Hongfang Liu, and Hua Xu. 2017. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association : JAMIA*, 25:331 – 336.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM*, 57(10):78–85.
- Peng Wang, Yun Zhong Hu, Shaochen Bai, and Shiyi Zou. 2021. [Matching biomedical ontologies: Construction of matching clues and systematic evaluation of different combinations of matchers](#). *JMIR Medical Informatics*, 9.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger,

Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. *Transformers: State-of-the-art natural language processing*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

World Health Organization (WHO). 2004. *ICD-10 : international statistical classification of diseases and related health problems : tenth revision*, 2nd ed edition. World Health Organization.

Elena Zotova, Montse Cuadros, and German Rigau. 2022. ClinIDMap: Towards a clinical IDs mapping for data interoperability. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3661–3669, Marseille, France. European Language Resources Association.

Elena Zotova, Montse Cuadros, and German Rigau. 2023a. Towards the integration of wordnet into clinidmap. In *Proceedings of the 12th Global Wordnet Conference*, pages 352–362.

Elena Zotova, Aitor Garcia-Pablos, Montse Cuadros, and German Rigau. 2023b. VICOMTECH at MedProcNER 2023: Transformers-based Sequence-labelling and Cross-encoding for Entity Detection and Normalisation in Spanish Clinical Texts. In *Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum*.

## A. Appendix A

Column Name	Description
CUI	Concept unique identifier
LAT	Language of terms
TS	Term status
LUI	Lexical (term) Unique Identifiers
STT	String Type
SUI	String Unique Identifiers
ISPREF	Atom status - preferred (Y) or not (N)
AUI	Atom identifier
SAUI	Source asserted atom identifier
SCUI	Source asserted concept identifier
SDUI	Source asserted descriptor identifier
SAB	Abbreviated source name, for example, SNOMEDCT_US or ICD10CM
TTY	Abbreviation for term type in source vocabulary
CODE	Most useful source asserted identifier
STR	String
SRL	Source restriction level
SUPPRESS	Suppressible flag
CVF	Content View Flag
ICD10CM_SPA	ICD-10-CM definition in Spanish
ICD10PCS_SPA	ICD-10-PCS definition in Spanish
SNOMEDCT2ICD10	ICD-10 identifier, mapped to SNOMED CT identifier
SNOMEDCT2ICD10_ENG	ICD-10 definition in English from SNOMED CT to ICD-10 mapping
WIKIDATA	Wikidata item identifier
MESH_WIKI	MeSH identifier extracted from Wikidata
SNOMED_CT_WIKI	SNOMED CT identifier extracted from Wikidata
ICD10_WIKI	ICD-10 identifier extracted from Wikidata
ICD10CM_WIKI	ICD-10-CM identifier extracted from Wikidata
ICD10PCS_WIKI	ICD-10-PCS identifier extracted from Wikidata
NCBI_WIKI	NCBI identifier extracted from Wikidata
WN31	WordNet 3.1 identifiers, blank space separated
WN30	WordNet 3.0 identifiers, blank space separated
WN_SENSE	WordNet synsets, blank space separated
TUI	Semantic Type Unique Identifier
SEMTYPE	Name of the Semantic Type
SEMGROUP	Semantic group abbreviation
DEF	Definition of the Semantic Group

Table 6: Description of the columns in the lexical resource.

# MedDialog-FR: a French Version of the MedDialog Corpus for Multi-label Classification and Response Generation related to Women’s Intimate Health

Xingyu Liu,<sup>1,3</sup> Vincent Segonne,<sup>2</sup> Aidan Mannion,<sup>1,4</sup>  
Didier Schwab,<sup>1</sup> Lorraine Goeuriot,<sup>1</sup> François Portet<sup>1</sup>

<sup>1</sup>Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

first.last@univ-grenoble-alpes.fr

<sup>2</sup>Université Bretagne Sud, UMR CNRS 6074, IRISA, F-56000 Vannes, France

first.last@univ-ubs.fr

<sup>3</sup>Shesmet <sup>4</sup>EPOS SAS

## Abstract

This article presents MedDialog-FR, a large publicly available corpus of French medical conversations for the medical domain. Motivated by the lack of French dialogue corpora for data-driven dialogue systems and the paucity of available information related to women’s intimate health, we introduce an annotated corpus of question-and-answer dialogues between a real patient and a real doctor concerning women’s intimate health. The corpus is composed of about 20,000 dialogues automatically translated from the English version of MedDialog-EN. The corpus test set is composed of 1,400 dialogues that have been manually post-edited and annotated with 22 categories from the UMLS ontology. We also fine-tuned state-of-the-art reference models to automatically perform multi-label classification and response generation to give an initial performance benchmark and highlight the difficulty of the tasks.

**Keywords:** Medical Corpus, Women’s Intimate Health, Multi-label Question Classification, Response Generation

## 1. Introduction

Medical conversation data is an essential resource for advancing healthcare research on dialogue systems. However, freely-available medical conversation data in certain languages is often limited, which poses a significant challenge for researchers working in those languages. This is particularly true for certain specialised domains, such as women’s intimate health, where the data sources are scarce and data collection is challenging due to ethical and privacy concerns.

There are large-scale dialogue corpora available in the field of healthcare in both English and Chinese, such as *MedDialog* (Zeng et al., 2020) and *medical\_conversation* (Song et al., 2020), two extensive medical dialogue datasets covering various medical specialities. Regarding French corpora, we are aware of only two examples of dialogue datasets. In the context of the *PVDial* project, Campillos-Llanos et al. (2020) created a virtual patient for medical education purposes. Interacting with both clinicians and non-clinicians, they released *PG-logs-eval*, a dataset comprising 115 dialogues. The dialogues simulate medical consultations. While this dataset is clearly useful for studying lexical choices and dialogue, it is based on a virtual agent and not real human patients. More recently, Laleye et al. (2020) introduced a medical conversation corpus of 41 dialogues, *Labforsims*, as part of the development

of a dialogue system between virtual patients and physicians. It is also worth mentioning the *Px-Corpus* (Kocabiyikoglu et al., 2023)(Kocabiyikoglu et al., 2022) composed of spoken dialogues between a smartphone and 55 participants including clinicians for drug prescription in French. Although these datasets clearly serve to enrich the French-speaking community, they are too small in size to train data-driven systems. Furthermore, none of them include topics related to women’s intimate health, while it is known that in many societies, the latter are disadvantaged by discrimination rooted in sociocultural factors (Aleksanyan and Weinman, 2022; Mehta et al., 2022).

In this study, we address the challenge of limited healthcare dialogue data in the French language by building a corpus of 20,000 dialogues on general medicine and women’s intimate health<sup>1</sup>. Due to the lack of available data in French, we leveraged the *MedDialog-EN* English dataset Zeng et al. (2020) and translated it into French. We then post-edited and annotated the translated corpus to support experiments of a multi-label classification task and a response generation task in French<sup>2</sup>.

The paper presents (1) the construction of a new medical dialogue dataset in French related to women’s intimate health and general medicine

<sup>1</sup>The corpus is available at our Zenodo repository: <https://doi.org/10.5281/zenodo.10889881>.

<sup>2</sup>The code for the experiments can be found at: <https://github.com/getalp/FRMedDialog>.

(in section 2), (2) the implementation and comparison of state-of-the-art approaches for two tasks: multi-label classification and response generation (in section 3).

## 2. The French MedDialog dataset

In this section, we present the method used in constructing the corpus, including the post-editing of the machine translation of dialogues selected and the manual annotation of questions. Additionally, we provide statistics on the corpus, such as the number of dialogues selected, the analysis of post-editing results, as well as the distribution of topics covered.

### 2.1. Method

#### 2.1.1. Data Selection and Translation

The source of our corpus, *MedDialog-EN* dataset, contains 257,454 English consultations between patients and doctors. Each consultation is composed of a textual single-turn dialogue: a patient describing their medical condition and asking a question which is answered by a physician.

**MedDialog-FR-women** We extracted dialogues concerning women's intimate health through the use of specific keywords provided by 2 women's intimate health practitioners. Each keyword corresponds to a distinct entry in the UMLS Metathesaurus (Bodenreider, 2004), an ontological medical data resource that combines multiple terminology systems. The corresponding keywords and UMLS entries are presented in the appendix [Appendix A](#).

Using a set of 17 keywords, we extracted a total of 16,149 dialogues. However, it should be noted that some of the dialogues extracted were not related to women's intimate health. For instance, the patient identified menopause as her health condition in her question, however, the query pertains to her liver issues, thus falling outside our study's focus.

**MedDialog-FR-general** In order to provide a more general-domain set of dialogues alongside the focus on women's intimate health, we extracted and translated an additional set of 7,120 dialogues based on a broad set of medical keywords, of which 500 were subsequently post-edited. We consulted with a French physician involved in the field of medical informatics research to put together a list of health conditions judged to be particularly interesting and/or important for general medical practice. The keyword list used to select the general-domain documents from *MedDialog-EN* is detailed in the appendix [Appendix A](#).

The subsequent step involved the use of neural machine translation to automatically translate the chosen dialogues. We employed DeepL's API<sup>3</sup> to automatically translate the extracted dialogues.

#### 2.1.2. Post-Editing

Machine translation has greater productivity and reduced costs compared with human translation. Current state-of-the-art automated translation from English to French appears to yield comprehensible and fluent translations. However, when it comes to medical text, accuracy and clarity are critical.

In order to align the translated text more closely with human standards of accuracy, fluency, and naturalness, and to ensure the accurate translation of medical terminology, we performed post-editing with native French speakers on a portion of the translated text.

**PE Platform** The post-editing was conducted with *doccano* (Nakayama et al., 2018) (see Figure 1), an open-source web-based text annotation tool. It provides annotation features for text classification, sequence labelling, and sequence-to-sequence tasks. *Doccano* met our requirements in this regard.

**Participants** In Figure 1, the presented example illustrates a machine translation that translated the English phrase "on birth control" as "*sous contrôle de naissance*" in French, a literal translation that was not natural. A more natural and accurate French expression should be "*sous contraception*". We decided that, for a dataset on women's health, female post-editors would be a more suitable choice. In our case, we recruited 2 undergraduates and 4 master's students in the fields of linguistics or natural language processing with a proficient level (B2) of English, all female. Before the PE task, we provided a one-hour training session to the post-editors with our PE guidelines. At the end of this session, we provided them with five common translations to post-edit, in order to ensure their understanding of the relevant principles.

**Guidelines** To ensure the effectiveness of the post-editing process and make the target text as consistent as possible, we prepared guidelines for the post-editing task. The basic PE rules in our guidelines rely upon the guidelines established by TAUS (Translation Automation User Society) (TAUS, 2016), such as "Ensure that no information has been accidentally added or omitted" and "Use as much of the raw MT output as possible". Our guidelines also included procedural steps relating to the use of our PE platform and a list of

<sup>3</sup><https://www.deepl.com/api>

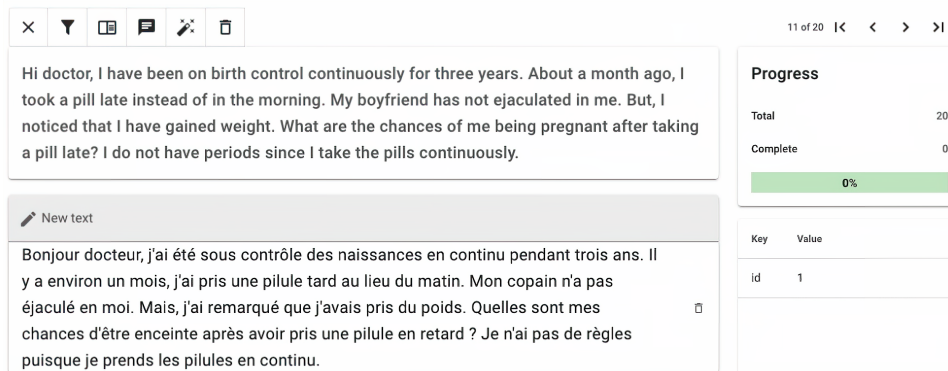


Figure 1: Post-Editing Interface with *Doccano* with an example from *MedDialog-EN*.

	Ed	TER	BLEU
Women’s health	37.81	0.079	0.91
General medicine	31.73	0.065	0.92

Table 1: Post-editing effort indicators

corresponding English to French translations for the specialized medical acronyms found in texts. The PE guidelines will be given as supplementary material in the final version of the paper.

**PE Technical Effort Indicators** Technical effort refers to the alterations made by the translator, which typically include insertions and deletions (Krings, 2001). Once post-editors finished post-editing, following previous works dealing with post-editing (Koponen, 2016; Snover et al., 2006; Alvarez et al., 2020), we calculated the following metrics as PE technical effort indicators (shown in Table 1) : Edit Distance (Ed) calculates the smallest number of edits needed to match the machine translation output with its post-edited version, Translation Edit Rate (TER) quantifies the edit operations required on the word level, and BLEU assesses the coherence of the machine-translated text with the post-edited content. A detailed explanation and examples are provided in the appendix [Appendix B](#).

Lower TER scores indicate better machine translation quality, and higher BLEU scores are generally associated with better machine translation quality, which can lead to reduced post-editing effort. In our task, TER scores below 0.1 and BLEU scores above 0.9 indicate that the results of machine translation were acceptable in general. In terms of the qualitative analysis, the main edits involved modifying medical acronyms and addressing incomplete translations.

**Anonymization** During the post-editing process, it came to our attention that certain user first names and doctors’ names had not been prop-

erly anonymized. In order to enhance data privacy, we carried out some additional anonymization steps. Names in questions were replaced with #Person1#, and names in answers with #Person2#. Additionally, URLs, email addresses, telephone numbers and other digits present in the original dataset were identified using regex and replaced by specific strings (e.g. #URL#, #EMAIL#). Manual verification was then carried out on a randomly-sampled subset of question-response pairs.

### 2.1.3. Annotation of MedDialog-FR-women

We initiated the process of multi-label annotation with questions related to women’s intimate health. As for the general medicine data, we plan to conduct the annotation in the future based on labels provided by doctors as per their requirements.

With the goal of categorizing user questions into different themes related to women’s intimate health, we leveraged the post-edited data by annotating it with predefined labels provided by 2 women’s intimate health practitioners.

**Annotation Platform** The multi-label annotation task was conducted with *doccano* (See Figure 2).

**Participants** The question labelling annotators were the same 6 annotators as for PE task. Another training session for the annotation was provided following the PE session.

**Labels** Following domain-expert recommendations, we selected 27 labels for annotation. This includes 25 topics related to women’s intimate health, such as *endometriosis*, *menopause*, and *contraception*, alongside two additional labels: *hors-sujet* (out of scope), indicating questions unrelated to women’s intimate health topics, and *autres* (other), signifying women’s intimate health subjects not covered by the other labels.

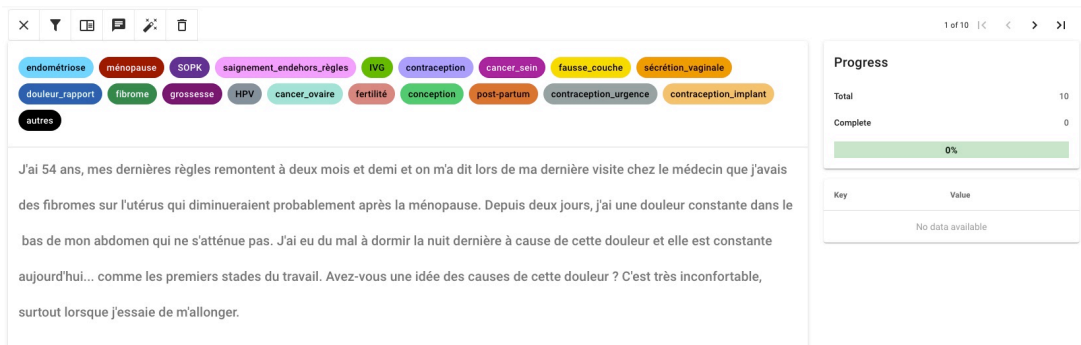


Figure 2: Multi-label Annotation Interface with *doccano*

**Guidelines** Multi-label annotation refers to annotation schemas where each question can be assigned to multiple labels or categories. In our guidelines, we presented the annotation procedure, provided explanations for each label, and offered illustrative examples to clarify situations involving multiple labels. We also asked annotators to specify the additional categories in the comments when selecting the *autres* (other) label for cases not covered by predefined categories.

#### 2.1.4. Post-Processing of Labels

After instructing annotators to specify the category of the question in comments when selecting the *autres* (other) label, we subsequently processed these comments to consolidate the list of labels and introduce new ones into our dataset.

Additionally, due to the limited occurrence of certain labels, we merged them to create more broadly defined categories, such as combining *contraception\_implant* (contraceptive implants), *contraception\_urgence* (emergency contraception), and *pillule* (contraceptive pill) into *contraception* (contraception), a single, higher-granularity label.

The post-processing of labels was validated by an expert in women’s intimate health. In the end, our multi-label dataset contained 22 labels (showed in Table 2). In the final dataset, we retain both the initial labels and the post-processed labels.

#### 2.1.5. Data Partitioning

We split the *MedDialog-FR-women* multi-label dataset into a training set of 500 instances, a validation set of 100 instances and a test set of 300 instances. The ratio was chosen to balance the need for maximizing the amount of fine-tuning data available while also ensuring that the test set is large enough for the results to be statistically significant, given the scarcity of some categories. To maintain consistent label distribution, we leveraged the iterative stratification algorithm (Sechidis

labels-en	labels-fr
endometriosis	endométriose
menopause	ménopause
PCOS	SOPK
conception	conception
painful sex	douleur_rapport
contraception	contraception
disorders of breast	affection_sein
ovarian cancer	cancer_ovaire
vaginal discharge	sécrétion_vaginale
abortion/VTP	IVG
uterine fibroid	fibrome
fertility/infertility	fertilité
cervical cancer	cancer_col_utérus
abdominal pain	douleur_abdominale
menstruation disorders	menstruation
swelling	gonflement
hot flushes	bouffée_chaleur
emotional disorder	troubles_humeur
out of scope	hors_sujet
sexually transmitted infections	IST
pelvic inflammatory disease	affection_appareil_génital
pregnancy, childbirth or the puerperium	g_a_p

Table 2: 22 post-processed French labels with English translation

et al., 2011) during the data splitting process. The label distribution is shown in Figure 3.

## 2.2. Statistics

To construct our French *MedDialog* Dataset (*MedDialog-FR*), we initially extracted from *MedDialog-EN* and automatically translated a total of 16,149 dialogues related to women’s intimate health and an additional 7,120 dialogues related to general medicine. From this dataset, we randomly selected 900 dialogues on women’s intimate health and 500 dialogues concerning general medicine for the PE task. Table 3 shows the statistics of the post-edited data. Subsequently, we performed multi-label annotation on the 900 questions extracted from these same dialogues focused on women’s intimate health. In total, 1,286 labels were distributed over the 900 dialogues, averaging 1.43 labels per instance.

The 6 annotators were thus tasked with post-editing 900 dialogues related to women’s intimate health and 500 dialogues concerning general medicine. Additionally, they annotated in to-

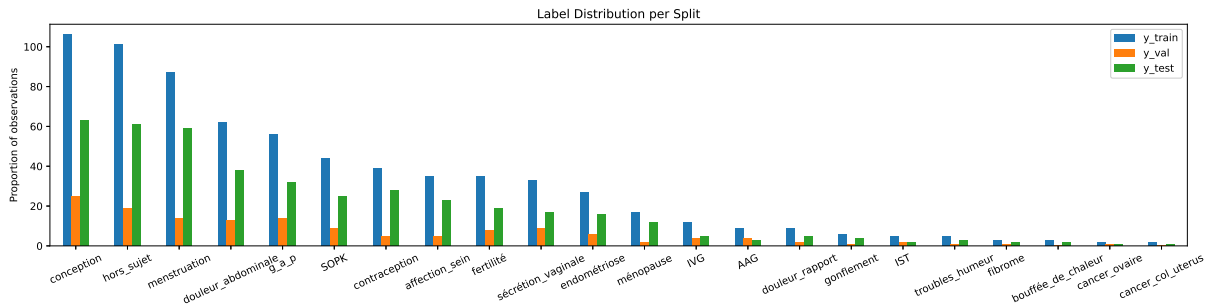


Figure 3: Label distribution per split

	Women	General	Total
# Dialogues	900	500	1400
# Tokens	199,574	81,958	281,532
Avg. # of tokens/Dialogue	225	163	203
Max. # of tokens/Dialogue	897	320	897
Min. # of tokens/Dialogue	59	40	40
Med. # of tokens/Dialogue	162	151	159

Table 3: Size of post-edited data

total 900 questions from the post-edited dialogues on women’s intimate health with multi-labels. The summary of the dataset is shown in Table 4.

Task	Women	General
Machine translation (# dialogs)	16,149	7,120
Post-editing (# dialogs)	900	500
Multi-label annotation (# questions)	900	-

Table 4: Statistics of MedDialog-FR dataset

### 3. Experiments

In this section, we introduce two tasks carried out using the MedDialog-FR-women dataset: multi-label question classification and response generation. We undertook these tasks with future applications in mind. These processes could be beneficial for determining the topic of a user’s question and proposing responses within a dialogue system focused on women’s intimate health.

#### 3.1. Multi-label Question Classification

##### 3.1.1. Method

Multi-label classification is concerned with categorizing instances into multiple classes at the same time. Each class associated with a given instance is referred to as a label. Following the previous work (Nam et al., 2014), we adopted the Binary Relevance method to convert the multi-label classification challenge into multiple single-label classification tasks. This approach involves treating each label as an independent binary classification problem.

#### 3.1.2. Models

Our classification architecture comprised a pre-trained BERT type model and a linear layer to convert the BERT representation to a classification task. The [CLS] representation is fed into a linear classification layer. We furthermore utilized a binary cross-entropy loss over sigmoid output (BCELosswithlogits with PyTorch) to measure the error for each label. Given the label imbalance in the dataset, we also experiment with weighted versions of the loss function, which aims to balance the precision-recall tradeoff by multiplicatively weighting positively-labelled examples in proportion to their prevalence in the training data.

As for the baseline models, we ran experiments on state-of-the-art large language models for the French language: *FlauBERT* (Le et al., 2020) and *CamemBERT* (Martin et al., 2020); and specialized French models tailored for the biomedical domain: *CamemBERT-bio* (Touchent et al., 2023) and *DrBERT-4G* (Labrak et al., 2023).

#### 3.1.3. Results

We present the results of our experiments involving two labeling approaches: one using all 22 categories and the other restricted to the 12 most common, grouping the 11 least frequently occurring labels under the *autres* (other) category. The purpose of using the 12 labels was to assess the performance of our method on a less imbalanced dataset, where we grouped the 11 least frequently occurring labels under the *autres* (other) category.

**22 labels** Once the model training was finished, its output could be construed as a probability distribution across the labels for a given instance. Based on its probability in the output, a threshold was required to decide if a label was predicted. We conducted a correlation analysis between various thresholds and F1 scores on the validation set using the best checkpoint during training to establish an optimal threshold for each model, as shown in Figure 4 for example. For the *CamemBERT-bio-base-weighted* model,



Model	macro			weighted
	P	R	F1	F1
FlauBERT-base	0.36	0.43	<b>0.38</b>	0.59
FlauBERT-base-weighted	0.41	0.36	0.37	0.54
CamemBERT-base	0.23	0.33	0.26	0.58
CamemBERT-base-weighted	0.38	0.29	0.32	0.53
CamemBERT-bio-base	0.33	0.40	0.35	0.59
CamemBERT-bio-base-weighted	0.45	0.44	<b>0.42</b>	0.63
DrBERT-4gb	0.45	0.29	0.33	0.50
DrBERT-4gb-weighted	0.40	0.31	0.31	0.46

Table 5: Model performance with on the MedDialog-FR-women test set containing 22 labels

when the threshold was 0.41, we achieved the highest F1 scores on the validation set.

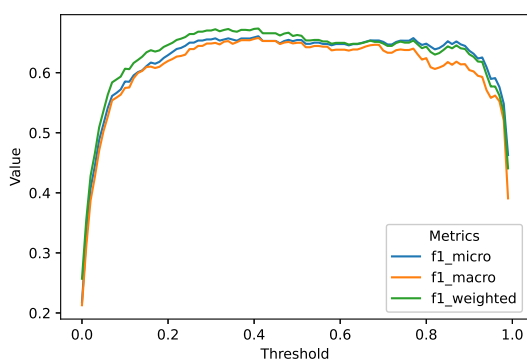


Figure 4: F1 scores of CamemBERT-bio-base-weighted on the MedDialog-FR-women validation set of 22 labels with different thresholds

Table 5 shows the performance of the models on the annotation test set of 22 labels. Models with the “weighted” suffix in their names indicate their utilization of class weighting in the loss function. The precision, recall and F1 score are calculated as the macro average across all labels. Additionally, the weighted F1 scores are calculated. The two best-performing models are CamemBERT-bio-base-weighted and FlauBERT-base, with macro-F1 scores of 0.42 and 0.38 respectively.

However, even for these two models, the F1 scores for certain labels with low occurrences, such as *cancer\_ovaire* (ovarian cancer) and *troubles\_humeur* (mood disorders), are exceedingly low and even reached 0. In an effort to mitigate the impact of data imbalance, we subsequently aggregated the 11 least frequently occurring labels under the *autres* (other) label and evaluated our approach on the remaining 12 labels with CamemBERT-bio-base-weighted and FlauBERT-base.

**12 labels** Table 6 shows the performance of CamemBERT-base-weighted and FlauBERT-base on the test set of 12 labels.

Model	macro			weighted
	P	R	F1	F1
FlauBERT-base	0.54	0.60	0.56	0.58
CamemBERT-bio-base-weighted	0.56	0.67	0.60	0.61

Table 6: Model performance on the MedDialog-FR-women test set containing 12 labels

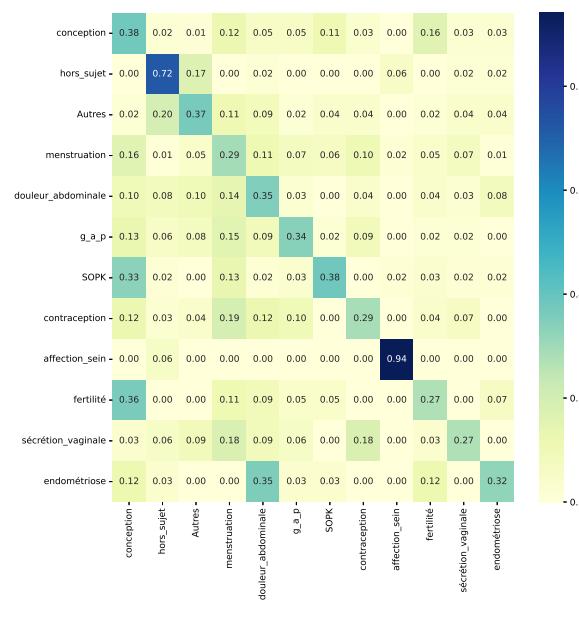


Figure 5: Confusion matrix for CamemBERT-bio-base-weighted on MedDialog-FR-women test set of 12 labels

In contrast to the 22 labels, the macro-averaged metrics for the 12 labels show significant improvement, which is understandable given the reduction in label imbalance. Moreover, the weighted-average scores are similar to macro-average scores, which suggests that the models are not significantly biased towards the larger labels and perform consistently across our dataset.

To assess the model’s performance for each label individually, we used the confusion matrix on the predictive accuracy of CamemBERT-base-weighted (see Figure 5). The matrix layout consists of rows representing true labels and columns representing predicted labels. High values along the diagonal indicate the model’s proficiency in making accurate predictions.

By examining the matrix, we can observe that the *affection\_sein* (breast disorders) category has yielded the highest number of accurate predictions. This can be attributed to the relatively independent nature of breast disorders within the domain of women’s intimate health, where questions of this category are less related to other topics. Additionally, there is some confusion between *fertilité* (fertility) and *conception* (conception), two closely related concepts. It can be difficult even for humans

Model	PPL ↓	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	Meteor ↑	BertScore ↑
Barthez	3.5	30.6%	18.6%	16.1%	20.9%	70.1
mBarthez	2.4	27.3%	6.8%	13.2%	18%	68.9
LLama2 (FT)	1.1	23.9%	4.5%	13.2%	15.8%	62.0

Table 7: Results of the response generation task using Barthez and LLama2 (w/t Fine-Tuning) on the test set of the corpus.

to clearly distinguish them apart.

## 3.2. Response Generation

### 3.2.1. Method

We also evaluated our corpus in a response generation task in which the goal is to automatically generate a response to a given user’s question. We followed the experimental protocol described in Zeng et al. (2020) where language models are used to generate answers.

We evaluated several language models from two families. On the one hand, we used the pretrained seq2seq models Barthez (French-only) and mBarthez (multilingual) (Kamal Eddine et al., 2021). We also included LLaMa2<sup>4</sup> (Touvron et al., 2023), an autoregressive model pretrained on a multilingual corpus. For the response generation task, we fine-tuned each of the pre-trained models on the training post-edited and non-post-edited data and evaluated them on the same validation and test post-edited data splits as used in the multi-label question classification task. For inference, we generated responses to questions through a beam search with (n=5) and a top-k random sampling (Fan et al., 2018) set to 50. To measure the quality of our generated answers, we compared them to the gold answers (i.e provided by real doctors) applying standard automatic methods used in generation and automatic translation: perplexity, ROUGE score (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (Zhang et al., 2020). These metrics capture different aspect of the quality of the generated answers: perplexity measures the quality of language modeling, ROUGE and METEOR are used in machine translation to evaluate the similarity between the hypothesis and the reference through n-gram matching. BERTScore measures the similarity between sentences using the BERT language models’ representations.

### 3.2.2. Results

The results of our experiments on the response generation task are presented in Table 7. We also provide examples of generation by different models in Table 8. Overall, the models fine-tuned on our corpus are able to generate well-formed and

coherent responses, which is reflected in good perplexity scores (the lower the better). The best model in this regard is LLama2, with an average perplexity of 1.1 on the entire test set, while Barthez obtains the highest perplexity with 3.5.

When examining the evaluation metrics obtained from automatic translation, they consistently behave and reveal two trends: Firstly, the overall results are relatively low, with the best ROUGE-1 score reaching approximately 30%, underscoring the challenging nature of the task with this dataset. Secondly, the Barthez model consistently outperformed other models across all metrics, while LLama2 consistently performed the poorest. Specifically, when comparing the performance of Barthez and LLama2 in terms of BERTScore, Barthez’s output appears to be more semantically aligned with human responses than LLama2’s. This observation may suggest the challenge faced by general multilingual autoregressive models in adapting to specialized domains and tasks.

## 4. Conclusion and Future Works

In this research, we present the MedDialog-FR dataset, a French version of the MedDialog-EN dataset. This dataset comprises 16,149 dialogues related to women’s intimate health topics and an additional 7,120 dialogues covering general medicine. Within this dataset, 1,400 dialogues have been post-edited, and 900 questions have been annotated with multiple labels pertaining to women’s intimate health topics. The post-editing process required several adjustments to align the data more closely with the French medical context, including changes in the names of medications and clinical procedures to their French equivalents. Additionally, the names within the dataset have been converted to French names to better adapt to the target medical system and enhance anonymization. To the best of our knowledge, this dataset represents the first dedicated resource in French for dialogues related to women’s intimate health.

Applying current state-of-the-art model to the multi-label classification task exhibits moderate performance with the best model achieving a 60% macro F1 score when dealing with 12 labels. Given the relatively small size of the dataset and the imbalance present in the data, these outcomes can be considered reasonable. They underscore the challenge that current NLP models face when tackling this task.

Automated response generation remains a challenging task, with the highest METEOR score in our experiments being 20%. Overall, the performance of the generative models in these experi-

<sup>4</sup>We used the Llama2-7b-chat version

<b>User question</b>	Bonjour docteur, j'ai beaucoup d'irritation sur l'ouverture de mon vagin et quand je le gratte, ça fait du bien mais ça devient plus irritant. Je suis allée chez le médecin récemment et j'ai eu une infection urinaire, mais je ressens toujours de fortes démangeaisons et je commence à m'inquiéter. J'ai 22 ans et pour l'instant je ne prends que des antibiotiques pour l'infection urinaire.
<b>Human response</b>	Bonjour, cela ressemble à un furoncle. Et c'est surtout lié à l'infection urinaire (pièce jointe supprimée pour protéger l'identité du patient). Continuez à prendre des antibiotiques. Il peut être en phase de guérison et c'est pourquoi les démangeaisons doivent être présentes. Mais vous pouvez prendre du sirop Cital 2 fois par jour pendant 15 jours. Et vous pouvez aussi appliquer de la pommade Candid-B qui aidera à réduire les démangeaisons. Ne vous grattez pas. Cela augmenterait l'infection. Buvez plus d'eau. Gardez la zone vaginale aussi sèche que possible. Si vous avez des pertes blanches par voie vaginale, vous pouvez prendre les comprimés Candid V6 par voie vaginale tous les soirs pendant 6 jours. Si l'infection n'est pas contrôlée par les antibiotiques donnés, il peut être nécessaire de changer d'antibiotiques ou d'utiliser des antibiotiques par injection."
<b>barthez response</b>	Bonjour, d'après vos antécédents, il semblerait que vous ayez développé une candidose, une infection génitale. Faites faire votre culture d'urine pour détecter la bactérie causant l'infection et prendre des antibiotiques en conséquence. Vous pouvez prendre des comprimés contre les infections urinaires jusqu'à deux fois par jour."
<b>mBarthez response</b>	"Bonjour, il pourrait s'agir d'une bactérie. Faites-moi savoir si je peux vous aider davantage. Pour plus d'informations, consultez un obstétricien et un gynécologue en ligne → <a href="https://icliniq.com/ask-a-doctor-online/obstetrician-and-gynaecologist">https://icliniq.com/ask-a-doctor-online/obstetrician-and-gynaecologist</a> ."
<b>Llama2</b>	"Bonjour, je suis désolé d'apprendre que vous ressentez des démangeaisons et de l'irritation à l'ouverture de votre vagin. Il est important de consulter un médecin pour obtenir un diagnostic précis et un traitement adéquat. Il est possible que votre infection urinaire soit la cause de vos démangeaisons et de l'irritation, mais il est également possible que cela soit lié à d'autres facteurs, tels que des infections vaginales ou des allergies. Il est donc important de consulter un médecin pour obtenir un diagnostic précis et un traitement adéquat. En ce qui concerne les antibiotiques que vous prenez pour l'infection urinaire, il est important de suivre les instructions du médecin et de prendre le médicament jusqu'à la fin de la prescription, même si vous vous sentez mieux. Cela aidera à éliminer l'infection et à vous aider à vous sentir mieux.

Table 8: Examples of generated responses to questions from the validation dataset

ments fell short of providing a solid basis for meaningful human evaluation of the generated outputs. Nevertheless, a brief examination of some of the generated responses revealed that existing automatic metrics may not be suitable for assessing response generation accurately. As a result, future work will involve implementing more refined fine-tuning techniques to attain outputs that are usable, and subsequently, we intend to conduct thorough human evaluations to more comprehensively assess the models' performance.

To further enable improved model performance, a key component of our upcoming efforts will be to expand our annotations, paying special attention to the less common categories that have low occurrences. We also aim to introduce annotations for medical entities within our datasets, with the aim of making significant contributions to the advancement of task-oriented medical dialogue systems.

## 5. Acknowledgements

This research was supported by Association nationale de la recherche et de la technologie (ANRT) under grant CIFRE 2021/0693 and was partially supported by MIAI@Grenoble-Alpes (ANR-19-P3IA-0003). We would also like to thank the participants to the annotation process.

## 6. Ethics Statement and Limitations

Access to actual medical data is heavily restricted in France. We thus used an already publicly available corpus in English. In addition to translation, a number of additional steps were taken to ensure that the MedDialog-FR dataset is fully anonymized and properly adapted for French-language applications. We first made sure that no personal informa-

tion could be found in the data. This is why we replaced all names that could have been kept in the original data. We also performed post-edition after automatic translation to adapt the phrasing and medical terminology to more natural French. In addition, our annotation project adhered to strict ethical guidelines, which include, but are not limited to, fair compensation for annotators. We do not foresee any direct social consequences or ethical issues.

The primary focus of this study centers on the dataset, with the conducted experiments serving as an initial benchmark to assess the task's complexity. Our preliminary goal is to reach a decent value on automatic metrics such as BLEU and METEOR prior to allocating valuable human resources for output evaluation. These experiments are designed to highlight the tasks' challenges, and we plan to undertake human evaluations once our models attain more robust metrics.

Since the original corpus is derived from dialogues in the U.S.A., there might be some cultural differences with French-speaking countries in the way people interact with doctors and which treatments and medical advises can be provided.

## 7. Bibliographical References

- Yeva Aleksanyan and Jason P. Weinman. 2022. Women, men and covid-19. *Social Science & Medicine*, 294:114698.
- Sergi Alvarez, Antoni Oliver, and Toni Badia. 2020. Quantitative analysis of post-editing effort indicators for NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 411–420,

- Lisboa, Portugal. European Association for Machine Translation.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Olivier Bodenreider. 2004. [The unified medical language system \(UMLS\): integrating biomedical terminology](#). PubMed PMID: 14681409; PubMed Central PMCID: PMC308795.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Moussa Kamal Eddine, Antoine Tixier, and Michalis Vazirgiannis. 2021. [BARThez: a skilled pretrained French sequence-to-sequence model](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9369–9390, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ali Can Kocabiyikoglu, François Portet, Prudence Gibert, Hervé Blanchon, Jean-Marc Babouchkine, and Gaëtan Gavazzi. 2022. A spoken drug prescription dataset in french for spoken language understanding. In *LREC 2022*.
- Maarit Koponen. 2016. Is machine translation post-editing worth the effort? a survey of research into post-editing and effort. *The Journal of Specialised Translation*, 25(2).
- Hans P Krings. 2001. *Repairing texts: Empirical investigations of machine translation post-editing processes*, volume 5. Kent State University Press.
- Yanis Labrak, Adrien Bazoge, Richard Dufour, Mickael Rouvier, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Drbert: A robust pre-trained model in french for biomedical and clinical domains. *bioRxiv*, pages 2023–04.
- Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. [FlauBERT: Unsupervised language model pre-training for French](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamel Seddah, and Benoît Sagot. 2020. [CamemBERT: a tasty French language model](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.
- Kajal Mehta, Hana Arega, Natalie L. Smith, Kathleen Li, Emma Gause, Joohee Lee, and Barclay Stewart. 2022. [Gender-based disparities in burn injuries, care and outcomes: A world health organization \(WHO\) global burn registry cohort study](#). *The American Journal of Surgery*, 223(1):157–163.
- Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. [doccano: Text annotation tool for human](#). Software available from <https://github.com/doccano/doccano>.
- Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. 2014. Large-scale multi-label text classification—revisiting neural networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14*, pages 437–452. Springer.
- Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proceedings, Part III 22*, pages 145–158. Springer.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. [A Study of Translation Edit Rate with Targeted Human Annotation](#). In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- TAUS. 2016. [Taus post-editing guidelines](#).
- Rian Touchent, Laurent Romary, and Eric de La Clergerie. 2023. [CamemBERT-bio: a](#)

tasty french language model better for your health. *arXiv preprint arXiv:2306.15550*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Rannan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating text generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

## 8. Language Resource References

Leonardo Campillos-Llanos, Catherine Thomas, Eric Bilinski, Pierre Zweigenbaum, and Sophie Rosset. 2020. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation. *Natural Language Engineering*, 26(2):183–220.

Ali Can Kocabiyikoglu, François Portet, Prudence Gibert, Hervé Blanchon, Jean-Marc Babouchkine, and Gaëtan Gavazzi. 2023. [Px-Corpus : A Spoken Drug Prescription Dataset in French for Spoken Language Understanding and Dialogue](#). 10.5281/zenodo.10080490.

Fréjus AA Laleye, Gaël de Chalendar, Antonia Blanié, Antoine Brouquet, and Dan Behnamou. 2020. A french medical conversations corpus annotated for a virtual patient dialogue system. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 574–580.

Yan Song, Yuanhe Tian, Nan Wang, and Fei Xia. 2020. Summarizing medical conversations via identifying important utterances. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 717–729.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. [MedDialog: Large-scale medical dialogue datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

## Appendix A. Matching UMLS Terms for Keywords

Keyword	UMLS Concept
endometriosis	Endometriosis (C0014175)
menstruation/period	Menstruation(C0025344)
menopause	Menopause (C0025320)
PCOS	Polycystic Ovary Syndrome (C0032460)
spotting	Metrorrhagia (C0728993)
conception	Fertilization (C0015914)
pregnancy	Pregnancy (C0032961)
pain&sex/pain&intercourse	Dyspareunia (C1384606)
contraception	Contraceptive methods (C0700589)
breast cancer	Malignant neoplasm of breast (C0006142)
ovarian cancer	Ovarian neoplasm (C0919267)
white discharge/vaginal discharge	Vaginal Discharge (C0227791)
miscarriage	Spontaneous abortion (C0000786)
abortion/VTP	Induced abortion (procedure) (C0392535)
postpartum	Postpartum Period (C0086839)
uterine fibroid	Uterine Fibroids (C0042133)
fertility/infertility	Female infertility (C0015895)
papillomavirus/HPV	Human Papillomavirus (C0021344)

Table 9: Keywords for women’s health dialogue selection and corresponding UMLS entries

Keyword(s)	UMLS Concept
angioedema	Angioedema (C0002994)
hypertension, high blood pressure	Hypertension or high blood pressure (C3843080)
hypoglyc(a)emia	Hypoglycemia (C0020615)
ACS, acute coronary syndrome	Acute Coronary Syndrome (C0948089)
pulmonary (o)edema	Pulmonary Edema (C0034063)
cardiac arrhythmia	Cardiac Arrhythmia (C0003811)
diabet(es)ic	Diabetes (C0011847)
ketoacidosis	Ketoacidosis (C0220982)
meningitis	Meningitis (C0025289)
cholecystitis	Cholecystitis (C0008325)
pyelonephritis	Pyelonephritis (C0034186)

Table 10: Keywords for general-domain dialogue selection and corresponding UMLS entries

## Appendix B. Post-Editing Example

The example below show the difference between a translated sentence and a post-edited one:

Machine translation: *Mes règles ont été retardées de 5 jours. D’habitude, j’ai des cycles réguliers. Je prends de la metformine 1000 mg depuis 6 mois, après avoir appris que j’ai un problème de PCOS. J’ai fait un test de grossesse aujourd’hui matin, mais le résultat était négatif...*

Post-edited: **J’ai un retard de règles de 5 jours.** *D’habitude, j’ai des cycles réguliers. Je prends de la metformine 1000 mg depuis 6 mois, après avoir appris que j’ai un problème de **SOPK. Ce matin,** j’ai fait un test de grossesse, mais le résultat était négatif...*

- Ed: Levenshtein distance calculates the minimum number of single-character edits (insertions, deletions or substitutions). For this example, 58 single-character edits are needed to transform the machine translation text into the reference text. This measure provides a granular view of the textual differences, reflecting the extent of similarity or divergence at the character level.

- TER: the Translation Edit Rate (TER) measures human edits on the machine translation, including the insertion, deletion, and substitution of single words. Using the post-edited text as reference, the TER score is calculated as the number of edits needed to change the machine translation into the reference, divided by the total number of words in the reference. There are 13 edits in the example above involving substitution (like “PCOS” (English acronym of Polycystic Ovary Syndrome → “SOPK” (French version)), deletion (such as “aujourd’hui” (today) to say “this morning” in a natural way in French), and insertion (including “J’ai”, “un”, “retard”, to express “I have a late period” in French). The reference text contains 45 tokens, thus, for this single example, the TER score is 13/45 (~0.289). We also calculate the average TER across all instances.

- BLEU: BLEU measures n-gram correspondence between the machine translated and the reference text. BLEU typically considers n-grams from 1 (unigrams) to 4 (4-grams). For example, the phrase “Je prends de la” (a 4-gram) appears in both texts and will contribute to the 4-gram precision. For each n-gram, the score represents the number of matching n-grams in the machine translation and reference, divided by the total number of n-grams in the machine translation. In this example, the BLEU score was 0.688.

# Exploring the Suitability of Transformer Models to Analyse Mental Health Peer Support Forum Data for a Realist Evaluation

Matthew Coole<sup>1</sup>, Paul Rayson<sup>1</sup>, Zoe Glossop<sup>2</sup>, Fiona Lobban<sup>2</sup>, Paul Marshall<sup>2</sup>, John Vidler<sup>1</sup>

<sup>1</sup>UCREL Research Centre, Computing and Communications, Lancaster University, UK

<sup>2</sup>Spectrum Centre, Health and Medicine, Lancaster University, UK

Lancaster University

{m.coole, p.rayson, z.glossop, f.lobban, p.marshall4, j.vidler}@lancaster.ac.uk

## Abstract

Mental health peer support forums have become widely used in recent years. The emerging mental health crisis and the COVID-19 pandemic have meant that finding a place online for support and advice when dealing with mental health issues is more critical than ever. The need to examine, understand and find ways to improve the support provided by mental health forums is vital in the current climate. As part of this, we present our initial explorations in using modern transformer models to detect four key concepts (connectedness, lived experience, empathy and gratitude), which we believe are essential to understanding how people use mental health forums and will serve as a basis for testing more expansive realist theories about mental health forums in the future. As part of this work, we also replicate previously published results on empathy utilising an existing annotated dataset and test the other concepts on our manually annotated mental health forum posts dataset. These results serve as a basis for future research examining peer support forums.

**Keywords:** mental health, peer support, transformers, machine learning

## 1. Introduction

Our project, related to improving the understanding and analysis of mental health peer online forums (iPOF<sup>1</sup>), aims to understand how mental health peer online support forums work and how we can improve them. The project is undertaking a realist evaluation (Pawson and Tilley, 1997), drawing together existing knowledge in a realist synthesis (Pawson et al., 2005), generating programme theories in the form of a series of Context, Mechanism, and Outcome (CMO) configurations. CMO configurations are short explanatory statements that articulate how health and social care programmes achieve their impacts. An example of one of the project's many CMOs is:

*When forums bring together people with shared mental health experiences (context), forum users will have access to mental health narratives that resonate with their own (mechanism), leading to improved sense of social connectedness (outcome).*

The project's CMOs will be evaluated using a mixed methods approach including surveys, qualitative interviews, linguistic analysis and NLP techniques. These different approaches will be triangulated together to create insights and recommendations that can be used by forum hosts and commissioners to build and manage more effective online peer support communities. This paper presents a set of NLP models built as an initial means of evaluation of some of our CMO configurations. Some of the features we need to investigate have been studied before in NLP. However, some are com-

pletely novel, and no comparative evaluations are possible.

The data collected in the iPOF project stems from several forum partners who agreed to participate in the project. There are eight forum partners in total, each participating at different levels. Language data has been gathered from seven forum partners; this includes any posts and replies made by users to the forums. There are several ethical considerations when gathering data, as some forums are openly accessible, some are closed but free to sign up to, and some require application or referral. This led to different procedures for different forum users to give informed consent for their data to be collected:

- opt-out: The project was advertised on open forums and users who did not wish their data to be collected could opt-out.
- implicit opt-in: Some closed forums collect consent to research from users at sign-up.
- explicit opt-in: Some closed forums (without a research consent option) required the collection of explicit consent from individual users or where forum use is dependent on giving research consent.

Each forum was given a bird name code by which it was referred to as part of our anonymisation procedures. In this paper, we will utilise the Starling corpus. The Starling corpus was selected to build the dataset and models as it was one of the larger and open forums participating in the project.

Based on the initial set of CMO configurations, we identified a set of common features or concepts which were crosscutting across multiple pro-

<sup>1</sup><https://www.lancaster.ac.uk/ipof/>

gramme theories. Thus, we annotated data and built models to predict when users shared a lived experience narrative in their posts and when a post showed connectedness, empathy and gratitude. These language models and experiments provide one avenue to test our programme theories and allow us to show that NLP methods can be used to scale up qualitative analysis to a much larger dataset than would otherwise be possible. We show that it is possible to develop NLP models for these concepts, and this will facilitate further work on other concepts and features that are important to a wider set of CMOs and programme theories. This will subsequently feed into guidelines for the description and improvement of peer online support forums, particularly in the areas of mental health, but potentially more widely in the future.

## 2. Related Work

Empathy and its detection has been explored extensively in recent NLP literature. Empathy in news articles through shared tasks has resulted in several approaches based on BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) models and its variants (RoBERTa, ALBERT) (Tafreshi et al., 2021) as well as exploring differences between demographics (Guda et al., 2021). BERT-based models, as well as LSTMs (Long short-term memory), have also been applied in the context of medical texts (Dey and Girju, 2023) as well as in online cancer survivor forum posts (Hosseini and Caragea, 2021) which also provides a dataset<sup>2</sup> with labelled texts for those seeking and giving empathy.

Detecting and using gratitude for extrinsic analysis has been explored for various reasons. Recently, many techniques have been applied to the detection of gratitude. BiLSTMs (Bidirectional LSTM) have been utilised for detection in obituaries (Sabbatino et al., 2020), RNNs (Recurrent Neural Networks) in online question and answer dialogues (Noseworthy et al., 2017), SVMs (Support Vector Machines) in online tweets (Danescu-Niculescu-Mizil et al., 2013) as well as BERT based models in online therapy texts (Burkhardt et al., 2022). In health forums specifically, KNN (K-Nearest Neighbours) and Naive Bayes have been applied (Sokolova and Bobicev, 2013).

Unlike empathy and gratitude, little previous work explores the automatic detection of lived experience narratives from online texts. The closest existing approach that could be comparable is the detection of personal medical disclosures (Valizadeh et al., 2021; Arseniev-Koehler et al., 2018). However, personal medical disclosures do not fall into the

same category of interest as discussion around lived experience, which go far beyond such disclosures and may include many things such as coping strategies, treatment experience and emotional support received from family and friends. Connectedness, as it is defined within the CHIME (Connectedness, Hope & optimism, Identity, Meaning, Empowerment) framework (Leamy et al., 2011), has yet to be explored through automatic detection.

Our review of existing literature around the detection of our four key concepts (connectedness, empathy, experience and gratitude) shows the varying degrees to which these concepts have been explored, from richly with available labelled datasets (empathy) to poorly defined vague concepts that need refinement and investigation (connectedness).

## 3. Methodology

### 3.1. Data

The dataset we chose to annotate for the concepts of connectedness, lived experience, and gratitude was the Starling corpus (for empathy, we made use of an existing dataset). This corpus was built from threads of a popular social media site with a sub-section for supporting mental health. This forum was selected from our forum partners as it was open with consent from users falling into the opt-out category (no users opted out). The dataset for this forum consists of 47k posts in 10k threads from 6k users. The tokenised dataset contains roughly 5.5M words.

### 3.2. Annotation

The Starling dataset was manually annotated for the key concepts connectedness, lived experience and gratitude. The annotated dataset was generated by sampling a random 2,000 posts and annotating each post for the three key concepts. A primary annotator tagged all 2,000 posts, and a secondary annotator also annotated a 5% subset to ensure the annotation guidelines were clear and to check for inter-annotator agreement (94.6%). For the empathy annotations, we made use of an available dataset annotated on cancer forum data (Hosseini and Caragea, 2021) as the medium of an online support forum is comparable to our own data, we attempt to replicate their results.

#### 3.2.1. Connectedness

Connectedness is a very difficult concept to identify and annotate in online text. For our purposes, we relied on a practical definition from the CHIME framework (Jagfeld et al., 2021). This meant we

<sup>2</sup><https://github.com/Mahhos/Empathy>



were looking for people discussing connecting with others. This may include:

- Discussion of peer support groups (their availability or experience of)
- Supportive relationships with family and friends, as well as intimate relationships
- Support from professionals, including helpful (and unhelpful) interactions with therapists and healthcare professionals.
- Discussion of actively participating in the community or online group.

### 3.2.2. Lived Experience (Narratives)

Lived experience narratives were annotated on the basis that the post included a personal experience of the user relating to their mental health or their experience caring for another with a mental health problem. This may include:

- Personal encounters or challenges with mental health services or treatments.
- Use of medications and their effects.
- Coping mechanisms employed by the individual.
- Interactions, both helpful and unhelpful, with friends and family.
- Personal accounts of emotional, professional, or personal issues stemming from their mental health condition.

### 3.2.3. Gratitude

Gratitude was annotated with particular emphasis on when users were showing gratitude to others in the forum. This was often shown by replies with very simple terms (“thanks :), “cheers!”), but the annotators also attempted to pick out more subtle indicators where only looking for specific wording may miss e.g. “That’s really helpful! :)”. Posts where specific terms of gratitude were used that appeared to merely be expressions of politeness were also not annotated as gratitude. This was very common in the data after initial posts on threads where terms of gratitude were used simply as a sign off to the post.

These concepts were annotated in this way as an initial exploration into some of the concepts that appear in our programme theories. These concepts will need to be developed further based on further refinement of what they mean within the context of different CMO configurations and based upon how their automatic detection can be used to test our theories. The concept of connectedness, in particular, is highly likely to require refinement and further annotation before it can be applied to theory testing.

Model	Precision	Recall	F1-Score	Support
<b>Connectedness</b>				
Naive Bayes	0.57	0.57	0.57	64
SVM	0.47	0.49	0.35	64
Random Forest	0.68	0.65	0.61	64
Distilbert	0.84	0.83	0.82	64
RoBerta	0.82	0.81	0.81	64
<b>Empathy</b>				
Naive Bayes	0.75	0.44	0.44	1001
SVM	0.71	0.49	0.50	1001
Random Forest	0.66	0.53	0.55	1001
Distilbert	0.79	0.79	0.79	1001
RoBerta	0.82	0.81	0.82	1001
<b>Experience</b>				
Naive Bayes	0.75	0.64	0.58	243
SVM	0.72	0.72	0.71	243
Random Forest	0.76	0.76	0.76	243
Distilbert	0.86	0.86	0.86	243
RoBerta	0.93	0.93	0.93	243
<b>Gratitude</b>				
Naive Bayes	0.81	0.81	0.81	91
SVM	0.84	0.80	0.79	91
Random Forest	0.92	0.91	0.91	91
Distilbert	0.98	0.98	0.98	91
RoBerta	0.98	0.98	0.98	91

Table 1: Model metrics (macro averages)

### 3.3. Models

Our annotated dataset was undersampled on the most common class to create three balanced datasets for connectedness, lived experience and gratitude. An existing dataset for empathy was used. A 60/20/20 train/test/validate split was used to ensure there was sufficient test data for some of the less frequently annotated concepts.

To compare our results, we also created baselines using Naive Bayes, SVM and Random Forest methods. For these baseline methods, we used word level TF-IDF (Term Frequency Inverse Document Frequency) vectors to train and used the popular SKLearn library<sup>3</sup>.

The BERT-based models we chose to use were DistilBERT (Sanh et al., 2019) and RoBERTa using the transformers library<sup>4</sup>. These were hyperparameter tuned using Optuna<sup>5</sup> with 50 runs varying learning rate, training and evaluation batch sizes. The final models were then trained over 30 epochs to assess their training loss and learning rate (see Figure 1).

## 4. Results

The results of our best model runs (with tuned hyperparameters) on our validation set are shown in Table 1. These results were taken when the best metric (F1 score) was found on our final training run (up to 30 epochs). In some cases, F1 score began to deteriorate (possibly as the result of overfitting).

<sup>3</sup><https://scikit-learn.org/stable/index.html>

<sup>4</sup><https://huggingface.co/docs/transformers/index>

<sup>5</sup><https://optuna.org/>

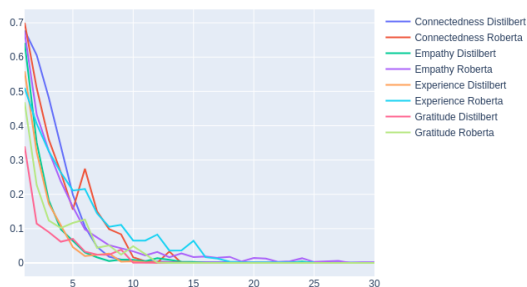


Figure 1: Model loss during training

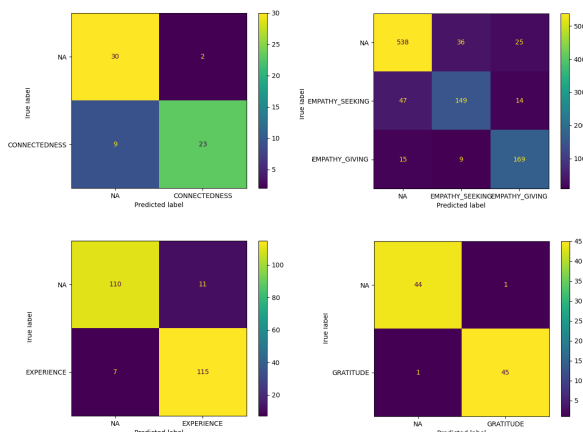


Figure 2: Confusion Matrices

We find that across all four concepts, BERT-based models outperform all baseline models, often by a substantial margin. Confusion matrices are shown in Figure 2.

Connectedness appeared to show little difference in performance between DistilBERT and RoBERTa, but they did both outperform Random Forest, which was the best baseline model. It could be that for connectedness (as this concept needs to be developed further and there is a low support) in the future these results may change. RoBERTa achieved the highest F1 score in the detection of experience; this concept was the only one where there was a notable difference between RoBERTa and DistilBERT. Gratitude is the only concept where one of the baseline models was comparable to the BERT-based models. Random Forest achieved excellent F1 scores for the detection of gratitude on our dataset, but DistilBERT and RoBERTa still proved to be the most effective in the detection of this concept.

Our experiments looking at the concept of Empathy using a pre-existing dataset achieved similar results to the original paper, which utilised the original BERT model. Tuned RoBERTa and DistilBERT models achieved F1 scores of 0.82 and 0.79 respectively, which are comparable to the original published score of 0.74 using BERT. The baseline models were also similarly comparable to the original publication. 0.44, 0.50, 0.55 for Naive Bayes, SVM and Random Forest respectively vs 0.39, 0.59, 0.55.

## 5. Conclusion

We have presented our initial exploration of building models for the detection of four key concepts related to mental health peer support forums: connectedness, empathy, experience and gratitude. We have demonstrated that modern transformer-based models (DistilBERT and RoBERTa) outper-

form classic baseline models (Naive Bayes, SVM, Random Forrest). We have also replicated previous work utilising BERT models on an existing empathy dataset.

Future work will take these models and apply them across seven datasets that have been collected as part of the iPOF project. We will then explore ways that they can be used to test a set of realist theories by looking for correlations between these and other annotations, e.g. Does the sharing of lived experience lead to an increase in sentiment in a forum thread? Does receiving expressions of empathy contribute to how active a forum user becomes?

## 6. Ethics Statement

There are very important ethical issues in analysing forum posts. People often share details about the things that are causing them distress, in the hope that other people who have faced similar problems can help them. It is vital that the forum feels a safe space in which to do this. We do not want this research to jeopardise this feeling of safety in any way. Therefore, we have developed a comprehensive ethical framework for this study. This has been developed with input from legal, clinical, academic and lived expertise, and approved by the Health Research Authority (IRAS 314029). As the project progresses we may need to make changes to how the study is conducted. Any changes will be approved by the study sponsor and the ethics committee and will be updated online for information<sup>6</sup>.

<sup>6</sup><https://www.lancaster.ac.uk/ipof/>

## 7. Acknowledgements

This study is funded by the NIHR Health and Social Care Delivery Research (HS&DR) (NIHR134035). The study is hosted by Berkshire Healthcare NHS Foundation Trust. The sponsor is Lancaster University.

## 8. Bibliographical References

- Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer. 2018. What type of happiness are you looking for?-a closer look at detecting mental health from language. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 1–12.
- Hannah Burkhardt, Michael Pullmann, Thomas Hull, Patricia Areán, and Trevor Cohen. 2022. Comparing emotion feature extraction approaches for predicting depression and anxiety. In *Proceedings of the eighth workshop on computational linguistics and clinical psychology*, pages 105–115.
- Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Priyanka Dey and Roxana Girju. 2023. Investigating stylistic profiles for the task of empathy classification in medical narrative essays. *arXiv preprint arXiv:2302.01839*.
- Bhanu Prakash Reddy Guda, Aparna Garimella, and Niyati Chhaya. 2021. Empathbert: A bert-based framework for demographic-aware empathy prediction. *arXiv preprint arXiv:2102.00272*.
- Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13018–13026.
- Glorianna Jagfeld, Fiona Lobban, Paul Marshall, and Steven H Jones. 2021. Personal recovery in bipolar disorder: Systematic review and “best fit” framework synthesis of qualitative evidence—a poetic adaptation of chime. *Journal of affective disorders*, 292:375–385.
- Mary Leamy, Victoria Bird, Clair Le Boutillier, Julie Williams, and Mike Slade. 2011. Conceptual framework for personal recovery in mental health: systematic review and narrative synthesis. *The British journal of psychiatry*, 199(6):445–452.
- Jasy Suet Yan Liew and Howard R Turtle. 2016. Exploring fine-grained emotion detection in tweets. In *Proceedings of the NAACL student research workshop*, pages 73–80.
- Fiona Lobban, Matthew Coole, Emma Donaldson, Zoe Glossop, Jade Haines, Rose Johnston, Steven H Jones, Christopher Lodge, Karen Machin, Paul Marshall, et al. 2023. Improving peer online forums (ipof): protocol for a realist evaluation of peer online mental health forums to inform practice and policy. *BMJ open*, 13(7):e075142.
- Michael Noseworthy, Jackie Chi Kit Cheung, and Joelle Pineau. 2017. Predicting success in goal-driven human-human dialogues. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 253–262.
- Ray Pawson, Trisha Greenhalgh, Gill Harvey, and Kieran Walshe. 2005. Realist review—a new method of systematic review designed for complex policy interventions. *Journal of health services research & policy*, 10(1\_suppl):21–34.
- Ray Pawson and Nick Tilley. 1997. *Realistic evaluation*. sage.
- Valentino Sabbatino, Laura Bostan, and Roman Klinger. 2020. Automatic section recognition in obituaries. *arXiv preprint arXiv:2002.12699*.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Marina Sokolova and Victoria Bobicev. 2013. What sentiments can be found in medical forums? RANLP.
- Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–104. Association for Computational Linguistics.
- Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying medical self-disclosure in online communities.

# Revisiting the MIMIC-IV Benchmark: Experiments Using Language Models for Electronic Health Records

Jesus Lovon-Melgarejo, Thouria Ben-Haddi, Jules Di Scala,  
Jose G. Moreno and Lynda Tamine

University of Toulouse, IRIT, 31000 Toulouse, France  
{firstname.lastname}@irit.fr

## Abstract

The lack of standardized evaluation benchmarks in the medical domain for text inputs can be a barrier to widely adopting and leveraging the potential of natural language models for health-related downstream tasks. This paper revisited an openly available MIMIC-IV benchmark for electronic health records (EHRs) to address this issue. First, we integrate the MIMIC-IV data within the Hugging Face *datasets* library to allow an easy share and use of this collection. Second, we investigate the application of templates to convert EHR tabular data to text. Experiments using fine-tuned and zero-shot LLMs on the mortality of patients task show that fine-tuned text-based models are competitive against robust tabular classifiers. In contrast, zero-shot LLMs struggle to leverage EHR representations. This study underlines the potential of text-based approaches in the medical field and highlights areas for further improvement.

**Keywords:** Large language models, MIMIC-IV benchmark, Text-based mortality classification

## 1. Introduction

Recent advancements in natural language processing (NLP) and information retrieval (IR) tasks have been significantly driven by Transformers-based models, such as BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019). These models have been trained on raw linguistic information with minimal supervision. Furthermore, the emergence of large language models (LLMs), such as ChatGPT (Achiam et al., 2023) and Llama 2 (Touvron et al., 2023), has extended these capabilities by scaling in parameters size and training data. In the medical domain, applying LLMs has emerged as a novel tool for patients and healthcare practitioners (Meskó and Topol, 2023). For example, electronic health records (EHR), composed of non-linguistic information such as laboratory measurements, procedures, and medication codes, are translated into linguistic reports using these models (Van Veen et al., 2023). However, it is still unclear how useful EHR model representations are in non-linguistic tasks. Beyond privacy concerns, the critical issue preventing the broad adoption of LLMs in for this problem is effectively transforming patient structured information from the raw EHR format to a linguistic unstructured format that can leverage the potential of LLMs' text-based representations. Existing Transformer-based models for patient data, such as TransformEHR (Yang et al., 2023) and BEHRT (Li et al., 2020), have adapted their architecture to consider tabular input data. However, this process requires a costly pre-training step that does not take advantage of the advancements in improved LLMs and free EHR benchmarks such

as MIMIC IV (Johnson et al., 2023). The latter provides large-scale intensive care unit (ICU) patient data in a tabular form related to established cohorts used in different downstream tasks (e.g., mortality patient classification). Consequently, we argue that improving the accessibility of these resources to meet the models' evolution is crucial for the field.

In this paper, we propose a simple but effective methodology to standardize the MIMIC-IV benchmark towards using state of the art (SOTA) Transformer-based architectures (BERT, DistilBERT (Sanh et al., 2019) and RoBERTa), and LLMs (Llama 2, Meditron (Chen et al., 2023)) for health-related predictive tasks. For this purpose, we identify six main groups of features on the ICU data and propose a template-based data-to-text transformation. Thus, we are able to provide a text document input that summarizes the patient's ICU entry. Additionally, and for the sake of reproducibility, we provide a Hugging Face *datasets* object<sup>1</sup> that automatically produces a clinical cohort in the desired textual format<sup>2</sup>. Our main contributions are as follows: 1) A standard MIMIC-IV benchmark, integrated into the Hugging Face *datasets* library, allowing flexible use of the EHRs representation in health-related downstream tasks; 2) A comprehensive set of experiments using eight different models for evaluating the effectiveness of our revisited MIMIC-IV benchmark on the mortality classification task.

<sup>1</sup><https://huggingface.co/docs/datasets/index>

<sup>2</sup>Publicly available at <https://huggingface.co/datasets/thbndi/Mimic4Dataset>

## 2. Background and Related Work

### 2.1. MIMIC Collections and Benchmarks

The Medical Information Mart for Intensive Care (MIMIC) collection (Johnson et al., 2023; Johnson et al.) is one of the largest and most recent EHR datasets. It includes more than 250,000 patients admitted to intensive care at Beth Israel Deaconess at Boston’s Beth Israel Deaconess Medical Center. For each patient, details of the full journey in the ICU are available in a deidentified form for privacy concerns<sup>3</sup>. The current version is the MIMIC-IV collection (Gupta et al., 2022) which collect patient data between 2008-2019 and uses ICD-9 and ICD-10 versions of the International Classification of Diseases (ICD)<sup>4</sup> to list diagnoses and to link medical procedures to diagnoses.

In recent works, multiple benchmarks were proposed for the medical domain (Harutyunyan et al., 2019; Gupta et al., 2022; Wang et al., 2020) using MIMIC collections (Johnson et al., 2023, 2016). They appear as a mainstream mean of model comparability and reproducibility. The MIMIC-IV data pipeline (Gupta et al., 2022) is proposed to preprocess data for downstream tasks. This pipeline is able to transform raw data into a ready-to-use tabular representation of the patient’s data. Additionally, it provides the mapping to ICD as well as standard techniques for dimensionality reduction. Although a first step is the proposal of the benchmarks, we aim to go for two steps forward in this work by proposing the integration of the MIMIC IV benchmark into *datasets* of Hugging Face<sup>5</sup>, one of the largest hub ready-to-use datasets, as well as the possibility of using Transformer-based models (including LLMs) for predictive tasks on EHRs.

### 2.2. Transformers for EHRs

Transformers-based models of the general domain, such as BERT, have been adapted to the clinical domain using medical-related linguistic collections such as PubMed (BioBERT (Lee et al., 2020) and ClinicalBERT (Alsentzer et al., 2019)). Recently, efforts to encode non-linguistic information of EHRs to model patient data have emerged with models such as BEHRT (Li et al., 2020), MedBERT (Rasmy et al., 2021), and TransformEHR (Yang et al., 2023). These models encode different health modalities in flexible architectures. However, they require pre-training on large-scale datasets and do not benefit from the significant progress of

<sup>3</sup>With regard to the Safe Harbor provision of the HIPAA.

<sup>4</sup><https://www.who.int/standards/classifications/classification-of-diseases>

<sup>5</sup>Our implementation respects MIMIC’s access policies by asking the user to provide the original data.

Transformer-based models in the NLP literature. Furthermore, LLMs such as ChatGPT (Achiam et al., 2023), Llama 2 (Touvron et al., 2023), and its medical variant Meditron (Chen et al., 2023) have shown outstanding performance in different clinical tasks related to adapting non-linguistic health data, such as images and EHR diagnostics, into text (Meskó and Topol, 2023; Yeo et al., 2023). However, the exploration of this linguistic EHR representation for non-linguistic tasks, referred to as EHR downstream tasks, is limited. In order to bridge this gap, we present experiments on EHR data to explore their potential.

## 3. MIMIC-IV Benchmark Revisited

Here, we detail the pipeline and EHR data used, then we describe the templates proposed for transforming tabular EHR data into textual inputs.

### 3.1. The Pipeline

We rely on the MIMIC-IV benchmark to produce the standard evaluation framework for text. Thus, first, we integrated the recommended pre-processing guidelines in the *datasets* library and implemented all the features of the MIMIC-IV-Data-Pipeline<sup>6</sup> provided in a tabular form, as shown in the left side of Figure 1. After the preprocessing steps, we obtained a tabular representation that includes the demographic, current diagnosis features and time-series features related to labs, medications, procedures, and vitals, as show in Table 1.

Note that features like CHAR/LAB are given in time intervals, thus a reduction/expansion strategy must be applied to normalize the representation size. Data imputation is commonly applied by sampling data from a fixed number of time windows or even averaging values across a sequence of time windows. As shown in Section 4.2, we did not find large differences between sampling or averaging these features.

### 3.2. Proposed Templates

Finally, feature EHR data is transformed to text using a template-based strategy as shown in right side of Figure 1 and described below:

“The patient {ethnicity} {gender}, {age} years old, covered by {insurance} was diagnosed with {cond\_text}.” With {cond\_text} corresponding to the textual description from ICD10 of the diagnoses.

“The chart events measured were: {chart\_text}.” With chart\_text the list of biological measurements of the form: {mean\_val} for {feat\_label}, mean\_val

<sup>6</sup><https://github.com/healthylaife/MIMIC-IV-Data-Pipeline>

Figure 1: Dataset generation pipeline for the tabular format and the text format.

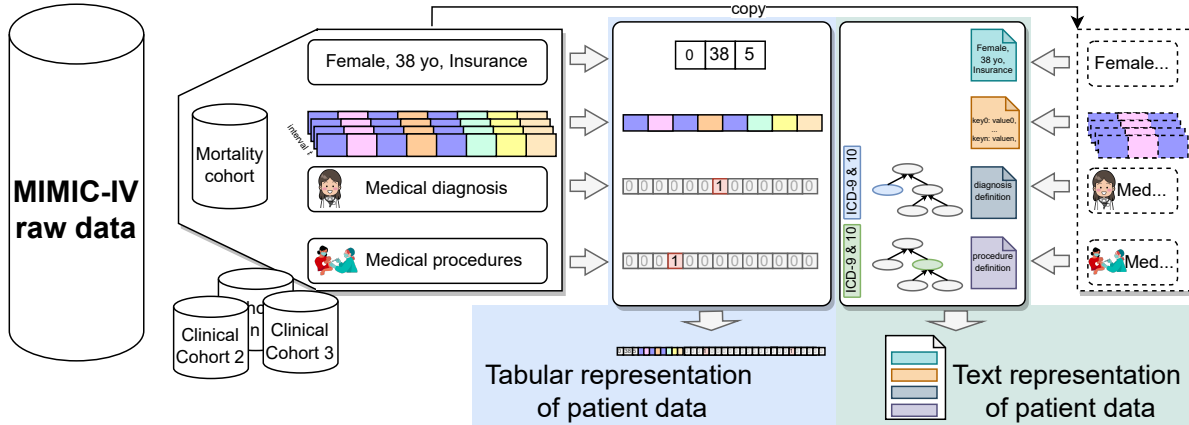


Table 1: Features list for the MIMIC-IV benchmark.

Name	Description
<i>Demo graphics (DEMO):</i>	The list of demographic data is a tiny vector corresponding to the patient’s gender, ethnicity, medical insurance, and age category. This data is encoded to obtain a numerical vector.
<i>Diagnosis (COND):</i>	The list of diagnoses established on a patient’s admission is encoded using a one-hot vector of all ICD codes including the patient’s identified diseases. Note that this vector could be large w.r.t. other features.
<i>Chart Events/Lab (CHART/LAB):</i>	Gives the value of the biological <i>item_id</i> performed in time interval <i>t</i> .
<i>Medications (MEDS):</i>	For each <i>item_id</i> corresponding to a medication the quantity administered in time interval <i>t</i> or zero if not administered.
<i>Procedures (PROC):</i>	The list of medical procedures performed is given as a form of a one-hot vector setting to 1 the <i>item_id</i> of procedures performed in time interval <i>t</i> .
<i>Output Events (OUTE):</i>	The list of biological samples taken is encoded using a one-hot vector of each <i>item_id</i> of the samples performed in time interval <i>t</i> .

is the mean value of the {feat\_label} measurement over the episode.

“The mean amounts of medications administered during the episode were: {meds\_text}.” With {meds\_text} the list of quantities of drugs administered of the form: {mean\_val} for {feat\_label}, {mean\_val} the average value over the episode of the quantity of drug {feat\_label}.

“The procedures performed were: {proc\_text}.” With {proc\_text} the list of medical procedures performed during the episode.

“The outputs collected were: {out\_text}.” With {out\_text} the list of biological prebiological samples taken during the episode.

Table 2 shows a sample of the textual input.

## 4. Experiments

### 4.1. Experimental Setup

To ensure a fair reproducibility of our experiments, we develop a *datasets* object that is able to produce tabular information as well as template-based textual data.

For **tabular data**, we create *Representation 1*, which follows the default configuration used in (Gupta et al., 2022), but other configurations are available in our implementation. Similarly, *Representation 2* is an aggregated representation of the same data. The main difference is the number of final features as the former uses 2766 features (as result of the concatenation of each window representation) and the latter 1110 features (as the values of all windows are averaged). We evaluated our revised MIMIC-IV benchmark on patient mortality classification as a pilot downstream task as provided in (Gupta et al., 2022). Evaluation focuses on both benchmark reproducibility (Cf. Section 4.2) and both feasibility and effectiveness using representative models (Cf. Section 4.3). Model parameters were selected using a 5-fold cross validation for classical machine learning algorithms available on Scikitlearn library<sup>7</sup>. We used algorithms for tabular data, such as Gradient Boosting (default parameters), XGBoost (objective=“binary:logistic”), Random Forest (n\_estimators=300, criterion=“gini”), and Logistic regression (default parameters).

For **textual data**, we fine-tuned six different Transformer-based models. We used optimal hyperparameters including learning rate of  $5e - 5$ ,

<sup>7</sup><https://scikit-learn.org/>

Table 2: Example of a text-based representation of a patient from the MIMIC-IV benchmark dataset. Values were changed to avoid leaking the example.

Feature	Example text
DEMO	The patient white male, 55 years old, covered by Other
COND	was diagnosed with Streptococcal sepsis; Acute pancreatitis; resistance to anti-microbial drugs.
CHAR/LAB	The chart events measured were: 73.655 for Heart Rate; 116.859 for Heart rate Alarm - High; ...
MEDS	The mean amounts of medications administered during the episode were: 44.778 of Albumin 5%; ...
PROC	The procedures performed were: Dialysis Catheter; 18 Gauge; EKG; ...
OUTE	The outputs collected were: OR EBL; OR Urine; Pre-Admission; ...

Table 3: Comparative evaluation of our standardized MIMIC-IV vs. original benchmark (Gupta et al., 2022) on the patient mortality classification task.

Algorithm	Representation 1		Representation 2		(Gupta et al., 2022)	
	AU-ROC	AU-PRC	AU-ROC	AU-PRC	AU-ROC	AU-PRC
<b>Gradient Boosting</b>	0.86	0.53	0.86	0.53	0.85	0.48
<b>XGBoost</b>	0.86	0.51	0.85	0.51	0.84	0.47
<b>Random Forest</b>	0.82	0.49	0.84	0.50	0.79	0.39
<b>Logistic Regression</b>	0.77	0.36	0.77	0.37	0.67	0.24

AdamW optimization and 3 epochs. For our zero-shot setup with LLMs, we explored multiple prompts. In the following, we report two of these prompts, which provide the highest number of valid responses for the task. We limited the output generation to 2 tokens.

We refer as *P1* for the prompt:

Prompt P1: "You are an extremely helpful health-care assistant. You answer the question using only 'yes' or 'no' and considering a patient hospital profile: '[textual EHR]'.  
Question: Is the patient dead?.  
Answer (only yes or no): "

Similarly, we refer as *P2* for the prompt:

Prompt P2: "Analyze the provided ICU data for a patient. The data covers the first 48 hours of the ICU stay, including vital statistics, lab test results, and treatments administered. Answer only Yes for a prediction of survival or No for a prediction of mortality. The patient ICU data is: '[textual EHR]'.  
Based on this data, answer.  
Question: Will the patient survive in the next 24 hours?.  
Answer (use only yes or no): "

We set a limit of 512 tokens for input length for

fine-tuned models and 1024 tokens for zero-shot models. It should be noted that this truncation only affected the fine-tuned models, and at times, it removed relevant information related to MEDS, PROC, and OUTE features. In Section 4.3, we discuss an ablation study that looks into the impact of these features.

## 4.2. Evaluation with Tabular EHR Data

Our results on tabular data and the reference values from the original benchmark (Gupta et al., 2022) are presented in Table 3. Note that our results are presented for two different aggregation strategies, *Representation 1* and *Representation 2*. In both cases, our results are slightly higher than those of the approach proposed in (Gupta et al., 2022) and used as a starting point. This is mainly due to our careful pre-processing of the data. As an important result, note that the *Representation 2* column performs similarly to *Representation 1* but uses significantly fewer features. Additionally, 1,034 values among 1,110 from the vector representation are sparse as they are dedicated to the diagnosis representation. These results lead us to pursue the text-based representation as only 66 values from biological signals combined with textual data (diagnosis) are enough to achieve state-of-the-art results on tabular data.

### 4.3. Evaluation of Using Template-based Text Inputs

Our main results on using text-based models for patient mortality classification tasks are presented in Table 4. For the fine-tuned models, we used the three general purpose trained models, namely DistilBERT (distilbert-base-uncased (Sanh et al., 2019)), BERT (bert-base-uncased (Devlin et al., 2018)), and RoBERTa (roberta-base (Liu et al., 2019)) (top three), and three others from the medical domain, namely BioClinicalBERT (Bio\_ClinicalBERT (Alsentzer et al., 2019)), BioBERT (dmis-lab/biobert-v1.1 (Lee et al., 2020)), and BiomedNLP (microsoft/BiomedNLP (Gu et al., 2021)) (bottom three). We reported only results with oversampling<sup>8</sup>. Our results show that the general purpose and domain-specific models behave similarly regarding AU-ROC, with all models getting close values (between 0.87 and 0.88). However, AU-PRC values differ as models from the medical domain outperform the general-purpose ones. Although a slight improvement was observed for general-purpose models in terms of AU-PRC, this is not enough to achieve the performance of the domain-specific models. Unsurprisingly, there is a clear interest in fine-tuning medical texts. However, general-purpose models, such as RoBERTa, closely follow top performances.

Furthermore, we explored using two LLMs, Llama2 (13b) (meta-llama/Llama-2-7b-hf (Touvron et al., 2023)) and its medical variant Meditron (7b) (epfl-llm/meditron-7b (Chen et al., 2023)) in a zero-shot setup considering two different prompts named *P1* and *P2*. We generally observed a lower performance from the Zero-shot section (as shown in Table 4) compared to Fine-tuned models. After analyzing the Zero-shot section, we found that prompt *P1* received better scores than *P2*. These results indicate that models are sensitive to the query format for this task. In addition, we noticed that domain-specific models, such as Meditron, performed better than general ones like Llama 2, using both prompts, similar to the fine-tuned setup. These findings suggest that SOTA LLMs struggle to encode and transfer EHR representations to downstream tasks within the explored prompts. A possible development towards using LLMs with tabular data is to define better translation methods to integrate this structured knowledge into the language models. Also, these findings motivate further research and experimentation by applying alternative techniques such as in-context learning (Dong et al., 2022) or prompt-tuning (Lester et al., 2021).

Moreover, in this setup, in addition to right or wrong answers, we also consider unanswered questions. Such questions occur when the LLM

<sup>8</sup>We found similar results without oversampling.

Table 4: Results of the general purpose and medical domain models on the patient mortality task using text representations of patient data.

Models	AU-ROC	AU-PRC
<b>Fine-tuned</b>		
DistilBERT	0.87	0.42
BERT	0.87	0.43
RoBERTa	0.88	0.47
BioClinicalBERT	0.87	0.43
BioBERT	0.88	0.45
BiomedNLP	0.88	0.46
<b>Zero-shot with prompt P1</b>		
Llama 2 (13b)	0.50	0.38
Meditron (7b)	0.61	0.39
<b>Zero-shot with prompt P2</b>		
Llama 2 (13b)	0.50	0.13
Meditron (7b)	0.51	0.23

Table 5: Number of answered and unanswered samples by the LLMs for the zero-shot setup.

Model	# answered	# unanswered
<b>With prompt P1</b>		
Llama 2 (13b)	5952 (96.70%)	203 (3.30%)
Meditron (7b)	6152 (99.96%)	3 (0.04%)
<b>With prompt P2</b>		
Llama 2 (13b)	1885 (30.63%)	4270 (69.37%)
Meditron (7b)	6155 (100.0%)	0 (0.00%)

fails to provide an output from the expected tokens, which are “Yes” or “No” in our case. For our experiments, we consider “No” the default answer for the results reported in Table 4. To provide further details, we display the number of answered and unanswered questions per model in Table 5. Upon analysis, we found that the Llama 2 model left 3.30% of the dataset unanswered, while the Meditron model left only 0.04% unanswered using prompt *P1*. In contrast, the prompt *P2* obtained 69.37% of unanswered questions with Llama 2 and no unanswered questions by Meditron.

By comparing the different prompts used to describe the task, we can observe that Llama 2 (general domain model) struggles to understand the task while making some modifications. In contrast, Meditron (domain-specific models) is more stable when using different reformulations of the task.

We further our analysis by performing an ablation study with two representative models, BERT and BiomedNLP, to study the accumulative effect of the features. Results are presented in Table 6. As a main feature, we can easily identify COND as a clear buster in performance. This feature alone





- finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. 2022. [An Extensive Data Processing Pipeline for MIMIC-IV](#). In *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pages 311–325. PMLR.
- Hrayr Harutyunyan, Hrant Khachatryan, David C. Kale, Greg Ver Steeg, and Aram Galstyan. 2019. [Multitask learning and benchmarking with clinical time series data](#). *Scientific Data*, 6(1):96.
- Alistair EW Johnson, Lucas Bulgarelli, Tom J Pollard, Steven Horng, L A Celi, and R Mark. [MIMIC-IV version 2.2](#).
- Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1):1–9.
- Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. *arXiv preprint arXiv:2005.10433*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.
- Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Ghulamreza Salimi-Khorshidi. 2020. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. 2021. Simplified data wrangling with `ir_datasets`. In *SIGIR*.
- Bertalan Meskó and Eric J Topol. 2023. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120.
- Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.
- Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askill, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.
- Huan Song, Deepta Rajan, Jayaraman J. Thiagarajan, and Andreas Spanias. 2017. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the 2018 AAAI Association for the Advancement of Artificial Intelligence*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava,

Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Dave Van Veen, Cara Van Uden, Louis Blanke-meier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. 2023. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. 2020. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. 2023. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 14(1):7857.

Yee Hui Yeo, Jamil S Samaan, Wee Han Ng, Peng-Sheng Ting, Hirsh Trivedi, Aarshi Vipani, Walid Ayoub, Ju Dong Yang, Omer Liran, Brennan Spiegel, et al. 2023. Assessing the performance of chatgpt in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clinical and molecular hepatology*, 29(3):721.

# Unraveling Clinical Insights: A Lightweight and Interpretable Approach for Multimodal and Multilingual Knowledge Integration

Kanimozhi Uma, Marie-Francine Moens

KU Leuven, Belgium

{kanimozhi.uma, sien.moens}@kuleuven.be

## Abstract

In recent years, the analysis of clinical texts has evolved significantly, driven by the emergence of language models like BERT such as PubMedBERT, and ClinicalBERT, which have been tailored for the (bio)medical domain that rely on extensive archives of medical documents. While they boast high accuracy, their lack of interpretability and language transfer limitations restrict their clinical utility. To address this, we propose a new, lightweight graph-based embedding method designed specifically for radiology reports. This approach considers the report's structure and content, connecting medical terms through the multilingual SNOMED Clinical Terms knowledge base. The resulting graph embedding reveals intricate relationships among clinical terms, enhancing both clinician comprehension and clinical accuracy without the need for large pre-training datasets. Demonstrating the versatility of our method, we apply this embedding to two tasks: disease and image classification in X-ray reports. In disease classification, our model competes effectively with BERT-based approaches, yet it is significantly smaller and requires less training data. Additionally, in image classification, we illustrate the efficacy of the graph embedding by leveraging cross-modal knowledge transfer, highlighting its applicability across diverse languages.

**Keywords:** Multimodal graph learning, Knowledge integration, Clinical reports, Language models

## 1. Background and Introduction

The advent of transformer-based architectures has revolutionized the field of medical text and image processing. Fine-tuned versions of the Bidirectional Encoder Representation from Transformers (BERT) model, such as ClinicalBERT (Alsentzer et al., 2019) and BioBERT (Lee et al., 2020), have demonstrated remarkable performance (Gu et al., 2021). ClinicalBERT, in particular, excels in tasks related to radiology reports of X-ray scans, including text-based disease classification and report generation. However, the direct application of general natural language processing (NLP) methods to the medical domain presents significant challenges, advocating for the development of domain-specific solutions for processing medical text. Due to the fact that the BERT models primarily focus on the English language, adapting these large language models for multilingual use poses significant challenges, leading to reduced performance. Also, Spanish counterparts of BERT, such as BETO (Cañete et al., 2023) and bio-cli-52k (Carrino et al., 2021) are trained on approximately ten times less data, resulting in inferior performance.

Rather than relying on self-supervision, we can employ structured medical knowledge. The Unified Medical Language System (UMLS) comprises standardized definitions and relationships within medical terminologies and vocabularies across 25 languages (Bodenreider, 2004). The UMLS can be implemented across various national hospitals and even transnationally, and it is particularly beneficial for countries that lack access to large med-

ical datasets due to their smaller population size or limited financial resources. The specific ontologies within UMLS offer additional advantages, as SNOMED CT provides connections between concepts within its respective ontology. The additional information from this knowledge base can be valuable, as expert-level annotation is rare in the medical domain. A particularly useful application of SNOMED CT is in clinical reports, which are widely available in public datasets but are largely unannotated. The structure of UMLS and SNOMED CT makes them suitable for representation with knowledge graphs, which can efficiently represent structured sets of entities (Chang et al., 2020). In specialized domains like medicine, language models must learn directly from domain-specific terminologies to enhance reliability, rather solely relying on corpus-based learning.

This paper introduces a novel self-attention graph embedding method for structuring clinical reports, integrating information from existing medical knowledge by leveraging both the report's structure and its linguistic content. Our experimental results demonstrate that the proposed self-attention graph embedding achieves competitive performance, and the text embeddings from the clinical report offer a more computationally efficient, more interpretable, and more intuitive alternative to existing embedding methods. We demonstrate that the utilization of UMLS and SNOMED CT facilitates effortless translation across languages, and finally, the proposed report graph can be integrated into a multimodal framework for knowledge transfer to images, enabling improved classification accuracy.

## 2. Related Work

Two main approaches are currently utilized for embedding medical text: pre-trained models on biomedical datasets and fine-tuned versions of those models. The pre-trained models include BioWordVec (Zhang et al., 2019), and the fine-tuned models include BERT variants such as BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), PubMedBERT (Gu et al., 2021) and many more, which are fine-tuned on (bio)medical and clinical datasets. SapBERT (Liu et al., 2021a), which introduced a self-alignment strategy for learning from UMLS synonym pairs via a multi-similarity (MS) loss function to force related concepts closer to one another in BERT’s representation space. These embeddings form the basis of the most recent state-of-the-art methods for radiology report generation and outperform previous methods by a large margin. Knowledge graphs have been employed to improve patient record-based diagnosis (Heilig et al., 2022), enhance entity extraction from radiology reports (Jain et al., 2021), and supplement image diagnosis (Prabhakar et al., 2022). The graphs used to generate a clinical report are typically small (approximately containing 15 nodes) consisting of disease labels, proving to be an effective method for capturing the global context (Yan, 2022; Liu et al., 2021c,b). Our approach suggests embedding from a clinical radiology report, rather than for generating one. No prior research has explored structured representations encompassing full clinical reports with knowledge graphs and medical ontologies.

## 3. Methodology

Our proposed method consists of three steps: extracting entities from the clinical report, constructing a knowledge graph, and encoding the graph, as shown in Figure 1. The graph’s nodes represent words in the report that match terms from the clinical database, and the edges represent relationships between these terms and their locations in the report.

### 3.1. Named Entity Recognition

Clinical concepts  $C$  embedded within the plain text of the clinical reports are extracted using Named Entity Recognition (NER) through MetaMap (Aronson and Lang, 2010), on English UMLS concepts; and for Spanish, we utilize UMLSMapper (Perez et al., 2020). From  $C$ , we extract UMLS Concept Unique Identifiers  $CUIs$ :  $\{i_0, i_1, \dots\} = Concepts(C)$ . For each CUI, the corresponding SNOMED-CT concept is extracted, resulting in the final set of clinical concepts from report  $C$ :  $\{c_0, c_1, \dots\}$ .

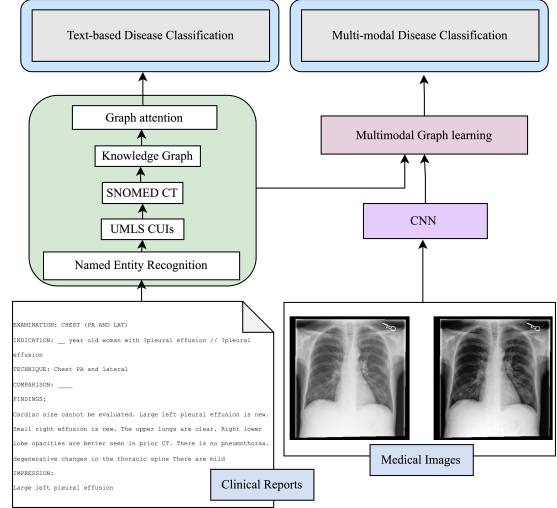


Figure 1: Overview of the architecture showcasing the process of constructing and assessing knowledge graph embeddings and encoding the graph with both textual and image data for disease classification.

### 3.2. Knowledge Graph Construction

An undirected graph is considered in our graph construction phase where the Graph  $G_C = (N_C, E_C)$  defined by a set of nodes  $N_C = \{n_0, n_1, \dots\}$  and edges  $E_C = \{e^{l \leftrightarrow m}, \dots\}$ , with  $e^{l \leftrightarrow m} = (n_l, n_m)$ . The structure of the clinical report is captured by considering each sentence separately with sentence node  $S^i$  of sentence  $i$ . The SNOMED-CT concepts extracted from the clinical report  $C$  gives us a set of concepts per sentence:  $\{c^{i,0}, c^{i,1}, \dots\}$ . To capture the context between nodes we have the global connect node  $g_n$  with the following set of nodes  $N_C = \{s^i, g_n, c\}$ . There are three types of edges: edges between the concept nodes, edges from the sentence nodes, and the global connect node, which connects to every concept node  $E_C = \{e^{c \leftrightarrow c}, e^{s \leftrightarrow c}, e^{g_n \leftrightarrow s}, e^{g_n \leftrightarrow c}\}$ . The types of edges are as follows: (a) Concept nodes  $c$  are linked by edges  $e^{c \leftrightarrow c}$  if they have a contextual relation based on the SNOMED CT ontology. (b) Sentence nodes  $s^i$  are linked to all the  $c$  nodes in their sentence by edges  $e^{s \leftrightarrow c}$ , representing the local structure of the report. (c) A global node  $g_n$  connects to every concept node and sentence node through edges  $e^{g_n \leftrightarrow c}$  and  $e^{g_n \leftrightarrow s}$ , allowing for communication across the whole report and thus,  $E_C = \{(c^l, c^m)\}, \{(c^l, s^i)\}, \{(s^i, g_n)\}, \{(c^l, g_n)\}$ .

### 3.3. Self-Attention Knowledge Graph Encoding

Graph attention networks use the self-attention mechanism to allow nodes in  $G_C$  to focus on their neighbors effectively (Velickovic et al., 2017) and are the preferred method for encoding knowledge graphs. The node  $N_C^u$  and neighbor nodes  $N_C^v$

with their weight matrices  $W \in R^{L_f * L_f}$ , where  $L_f$  is the length of the node feature. The normalized attention score between these nodes can be written as:  $att_{uv} = softmax(LRelu(W_{att^x}[W_{N_C^u} || W_{N_C^v}]))$ . The node  $N_C^u$  is encoded as:  $N_C^{u'} = \sigma(\sum att_{uv} \cdot W_{N_C^v})$ , where  $\sigma$  is a nonlinear function. The whole graph with  $n$  attention layers stacked on top of each other is encoded as:  $G'_C = f_{attention}^n(N_C, E_C)$ .

### 3.4. Analysis and Inferences

To determine whether the constructed graph embeddings capture information representative of the content of the clinical report, we evaluate them on two classification tasks, as illustrated in Figure 1. The effectiveness of our knowledge graph embedding is assessed for a diagnosis classification based on the clinical report, and we compare our method against biomedical variants of BERT. Disease classification is performed on the entire report through a max pooling operation on encoded node representations  $N'_C$ , followed by a classification.

The ability of our graph embedding to transfer the information it contains across modalities is tested by integrating the embeddings into a knowledge distillation with variational inference (KDVI) (Ahn et al., 2019), where graph embeddings are employed to improve image-based disease classifications. A conditional latent variable model was introduced to distill the information from clinical report  $C$  to chest X-ray scan through variational inference and we draw inspiration from conditional variational inference through KG reasoning (CVIR) (Chen et al., 2018).

The evidence lower bound objective (ELBO) of CVIR consists of a term for reconstructing the data and a term for measuring the Kullback-Leibler (KL) divergence:  $L = E[\log p(y|I, z_I)] - D_{KL}[q(z)||p(z_I|I)]$  where  $z_I$  is a hidden representation of  $I$ ,  $y$  are the class labels, and  $p(z_I|I)$  is the prior distribution over  $z$ .  $p(z_I|I) \cdot q(z)$  is the posterior distribution over  $z$ , which is usually an isotropic Gaussian distribution  $\mathcal{N}(0, I)$ . In KDVI, the posterior is  $q(z_C|C)$ , where  $z_C$  is a hidden representation of  $C$ . This new posterior allows us to extract information from  $C$  to  $I$  by minimizing this KL term:  $D_{KL}[q(z_C|C)||p(z_I|I)]$ . This way, we transfer knowledge and information from text to imaging. During the training phase, we need both as input, but in the testing phase, we only need the imaging data. With this approach, we use our text-based knowledge graph embeddings of the clinical report to enhance the image representations.

## 4. Experiments and Results

Three datasets were used for training and evaluation: MIMIC-CXR (Johnson et al., 2019), OpenI

(Demner-Fushman et al., 2016), and PadChest (Bustos et al., 2020). MIMIC-CXR comprises 377,110 chest X-rays and 227,827 anonymized radiology reports, with disease labels generated using a rule-based labeler. OpenI consists of 7,470 chest X-rays and 3,955 anonymized reports with similar disease labels. PadChest contains 160,000 radiology images and Spanish clinical reports with 174 disease labels. The disease labels from all three datasets can be consolidated into a unified label space. There are no limitations on the number of entities that can be extracted from the clinical report using NER with MetaMap or on the number of edges within a graph. We initialize the nodes using vectorized representations of individual UMLS concepts, which are obtained by pre-training on datasets containing (bio)medical data (Beam et al., 2019). These 200-dimensional non-contextual embeddings are inspired by word2vec and can be directly integrated into our method without requiring additional processing steps. Embedding initializations for  $s_i$  and  $g_n$  are calculated by averaging the node embeddings  $n$  over the sentence and the entire graph, respectively.

Metric	BioBERT	PubMedBERT	ClinicalBERT	Ours
AUC	0.920	0.946	0.974	0.948
Recall	0.840	0.846	0.868	0.850
Precision	0.593	0.600	0.611	0.632
F1	0.688	0.713	0.726	0.726

Table 1: Comparison of our approach with baseline models on MIMIC-CXR dataset

Metric	BioBERT	PubMedBERT	ClinicalBERT	Ours
AUC	0.929	0.929	0.965	0.966
Recall	0.768	0.818	0.818	0.864
Precision	0.560	0.571	0.602	0.625
F1	0.680	0.717	0.700	0.717

Table 2: Comparison of our approach with baseline models on OpenI dataset

Metric	BETO	Bio-cli-52k	Ours
AUC	0.881	0.918	0.966
Recall	0.509	0.678	0.841
Precision	0.219	0.490	0.545
F1	0.555	0.574	0.682

Table 3: Comparison of our approach with baseline models

The graph attention encoder consists of 1, 3, 6, or 12 consecutive graph attention layers with hidden sizes of 512, 1024, 2048, or 4094, and the graph classification is performed using a multilayer perceptron (MLP) with dimensions 512, 256, 14, 8, employing cross-entropy loss. The results are reported using the AUC metric, in line with existing

benchmarks, and the evaluation of VKD is conducted with a latent space size of 2048 and 12 sequential graph attention layers for encoding. Other hyperparameter settings are directly adopted from (Chen et al., 2018), and a dropout rate of 0.5 is applied to all layers of the architecture.

Training is performed on two NVIDIA GeForce RTX 3090 GPUs, utilizing Adam optimizer with early stopping and a tolerance of 1%. Tables 1, 2, and 3 present the disease classification results obtained from our graph embedding approach, and the results are shown for an encoder with a hidden size of 1024 and 3 graph attention layers, which demonstrated the best overall performance. Our method achieves comparable performance to BioBERT and PubMedBERT, and slightly lower performance compared to ClinicalBERT while utilizing significantly fewer parameters (Table 4). On the Spanish PadChest dataset, our method outperforms BERT-based methods. This can be attributed to the smaller training datasets of these models, which are ten times smaller than their English counterparts. Furthermore, our method exhibits faster inference rates (samples per second) on GPU platforms and performs relatively better on the smaller OpenI dataset, highlighting the effectiveness of our embedding for report representation without the need for large datasets.

Dataset	ClinicalBERT	Ours
MIMIC-CXR (Recall)	0.564	0.528
MIMIC-CXR (Precision)	0.537	0.537
MIMIC-CXR (F1)	0.508	0.469
MIMIC-CXR (AUC)	0.838	0.825
OpenI (Recall)	0.581	0.579
OpenI (Precision)	0.533	0.528
OpenI (F1)	0.509	0.497
OpenI (AUC)	0.884	0.864
PadChest (Recall)	0.528	0.534
PadChest (Precision)	0.510	0.532
PadChest (F1)	0.480	0.474
PadChest (AUC)	0.814	0.821

Table 4: Classification Accuracy

Employing graphs within a multi-modal framework can enhance our understanding of how effectively the graph captures intricate information structures that can span across modalities. Table 4 presents the results of our method for multi-modal knowledge transfer compared to existing methods that utilize ClinicalBERT as a clinical report embedding. Training this framework revealed that the convergence of this model is complex for graphs with shallow encoders and smaller hidden layers. As a result, Table 4 displays results obtained with an encoder comprising 12 graph attention layers and a hidden size of 2048. While not outperforming the existing ClinicalBERT method, graph embeddings demonstrate applicability on both MIMIC-

CXR and OpenI. Performance improvements are observed compared to image-only classification, indicating successful multi-modal knowledge transfer with very limited pre-training data.

	MIMIC-CXR	OpenI	PadChest
Full graph	0.930	0.946	0.950
w/o graph	0.924	0.939	0.940
w/o $g_n$	0.917	0.931	0.933
w/o $C$	0.917	0.935	0.925
w/o $g_n$ & $S$ & $C$	0.919	0.915	0.920

Table 5: Graph ablations on the classification tasks

Also, we delved into the application of graph encoders for disease classification, investigating the influence of encoder count and hidden size on performance. Parameter count emerges as a crucial factor, with ClinicalBERT delivering superior performance but requiring greater computational resources. The performance discrepancy between the smallest, say, 0.4M parameters and the largest, say, 62M parameters models is relatively minor, indicating that graph construction effectively captures medical knowledge irrespective of encoder size. Furthermore, we conducted an ablation study on the graph components in Table 5. The importance of node types is highlighted by removing them from the graph. The global node and edges connecting SNOMED CT concepts stand out as key elements in the graph structure. This underscores how the integration of report composition and the medical knowledge base (SNOMED CT) yields a rich representation of the report and shows how our graph handles repeated terms more efficiently than ClinicalBERT: ours consists of 34 nodes, while tokenization with ClinicalBERT requires as many as 124 tokens. For instance, our knowledge graph captures the entire word such as 'dencities', 'opacities' as an entity, but ClinicalBERT, tokenizes into 'den', 'cities', 'o', 'pa', and 'cities', respectively. The final token in this sequence clearly carries a different contextualized meaning within the general-purpose language BERT model that underlies ClinicalBERT, which demonstrates that graphs can capture medical terminology in a more intuitive and interpretable way.

## 5. Conclusion

This paper presents a novel knowledge graph-based method for creating structured representations of clinical reports. The knowledge graph embeddings explicitly encode medical knowledge from clinical knowledge bases, facilitating transfer across domains and languages without relying on large datasets. Concurrently, the proposed method maintains a significantly smaller model size compared to existing BERT-based models. By captur-

ing both structural and content relationships embedded within existing knowledge bases, the proposed self-attention-based graph representations achieve comparable performance to current state-of-the-art transformer-based models in English and Spanish, resulting in more informative representations of clinical reports. Our future scope is to extend this architecture to Dutch, German, Estonian, and more languages.

## 6. Acknowledgements

We thank the anonymous reviewers for the valuable comments. The authors acknowledge the AIDAVA project financed by Horizon Europe (EU HORIZON-HLTH-2021-TOOL-06-03) and the CHIST-ERA ANTIDOTE project financed by FWO G0H6820N for supporting the research.

## 7. Bibliographical References

- Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. 2019. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9163–9171.
- Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.
- Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.
- Andrew L Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. 2019. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing 2020*, pages 295–306. World Scientific.
- Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl\_1):D267–D270.
- Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data. *arXiv preprint arXiv:2308.02976*.
- Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *arXiv preprint arXiv:2109.03570*.
- Arlene Casey, Emma Davidson, Michael Poon, Hang Dong, Daniel Duma, Andreas Grivas, Claire Grover, Víctor Suárez-Paniagua, Richard Tobin, William Whiteley, et al. 2021. A systematic review of natural language processing applied to radiology reports. *BMC medical informatics and decision making*, 21(1):179.
- David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt, and Richard Andrew Taylor. 2020. Benchmark and best practices for biomedical knowledge graph embeddings. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 167. NIH Public Access.
- Wenhu Chen, Wenhan Xiong, Xifeng Yan, and William Wang. 2018. Variational knowledge graph reasoning. *arXiv preprint arXiv:1803.06581*.
- Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Niclas Heilig, Jan Kirchhoff, Florian Stumpe, Joan Plepi, Lucie Flek, and Heiko Paulheim. 2022. Refining diagnosis paths for medical diagnosis based on an augmented knowledge graph. *arXiv preprint arXiv:2204.13329*.



- Jinpeng Hu, Jianling Li, Zhihong Chen, Yaling Shen, Yan Song, Xiang Wan, and Tsung-Hui Chang. 2021. Word graph guided summarization for radiology findings. *arXiv preprint arXiv:2112.09925*.
- Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022. Graph enhanced contrastive learning for radiology findings summarization. *arXiv preprint arXiv:2204.00203*.
- Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.
- Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514.
- Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.
- Kaveri Kale, Pushpak Bhattacharyya, Aditya Shetty, Milind Gune, Kush Shrivastava, Rustom Lawyer, and Spriha Biswas. 2022. Knowledge graph construction and its application in automatic radiology report generation from radiologist's dictation. *arXiv preprint arXiv:2206.06308*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciampi, Mohsen Ghahfoorian, Jeroen AWM Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.
- Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. [Self-alignment pretraining for biomedical entity representations](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.
- Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762.
- Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. 2021c. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, 34:16266–16279.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Naiara Perez, Pablo Accuosto, Àlex Bravo, Montse Cuadros, Eva Martínez-García, Horacio Saggion, and German Rigau. 2020. Cross-lingual semantic annotation of biomedical literature: experiments in spanish and english. *Bioinformatics*, 36(6):1872–1880.
- Chinmay Prabhakar, Anjany Sekuboyina, Hongwei Bran Li, Johannes C Paetzold, Suprosanna Shit, Tamaz Amiranashvili, Jens Kleesiek, and Bjoern Menze. 2022. Structured knowledge graphs for classifying unseen patterns in radiographs. In *Geometric Deep Learning in Medical Image Analysis*, pages 45–60. PMLR.
- Ardavan Saeedi, Yuria Utsumi, Li Sun, Kayhan Batmanghelich, and Li-wei Lehman. 2022. Knowledge distillation via constrained variational inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8132–8140.
- Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.
- Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. 2017. Graph attention networks. *stat*, 1050(20):10–48550.
- Sixing Yan. 2022. Memory-aligned knowledge graph for clinically accurate radiology image report generation. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 116–122.
- Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510.

- Dehai Zhang, Anquan Ren, Jiashu Liang, Qing Liu, Haoxing Wang, and Yu Ma. 2022. Improving medical x-ray report generation by using knowledge graph. *Applied Sciences*, 12(21):11111.
- Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):52.
- Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12910–12917.
- Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81.

# Automated Question-Answer Generation for Evaluating RAG-based Chatbots

Juan José González Torres, Mihai-Bogdan Bîndilă, Sebastiaan Hofstee, Daniel Szondy, Quang-Hung Nguyen, Shenghui Wang, Gwenn Englebienne

University of Twente

Drienerlolaan 5, 7522 NB Enschede

{j.j.gonzaleztorres, m.bindila, s.b.h.c.hofstee, d.g.szondy, nguyenquanghung}@student.utwente.nl

{shenghui.wang, g.englebienne}@utwente.nl

## Abstract

In this research, we propose a framework to generate human-like question-answer pairs with long or factoid answers automatically and, based on them, automatically evaluate the quality of Retrieval-Augmented Generation (RAG). Our framework can also create datasets that assess hallucination levels of Large Language Models (LLMs) by simulating unanswerable questions. We then apply the framework to create a dataset of question-answer (QA) pairs based on more than 1,000 leaflets about the medical and administrative procedures of a hospital. The dataset was evaluated by hospital specialists, who confirmed that more than 50% of the QA pairs are applicable. Finally, we show that our framework can be used to evaluate LLM performance by using Llama-2-13B fine-tuned in Dutch (Vanroy, 2023) with the generated dataset, and show the method's use in testing models with regard to answering unanswerable and factoid questions appears promising.

**Keywords:** LLMs, Retrieval Augmented Generation, Chatbot Evaluation, Hallucination Detection

## 1. Introduction

Chatbots' performance has been greatly enhanced with recent advancements in RAG-based LLMs, where questions are supported by verified sources of information so that LLMs can answer consistently and accurately. However, evaluating these chatbots requires an enormous amount of labelled data that is often costly to produce in terms of human and financial resources. Moreover, evaluation datasets need to satisfy a few different criteria:

- **Covers large knowledge base:** A RAG pipeline often includes thousands of documents in various topics and language levels (scientific, conversational, etc.)
- **Includes different types of answers:** Answers can be factoid or long-form, depending on the type and format of questions. There might also be unanswerable questions, whether due to failure in the retriever or lack of pre-trained knowledge.

Therefore, a proper evaluation workflow should assess the chatbot's knowledge across all the topics covered in these documents and include all types of questions users could ask. In this research, we focus on automating the creation of a comprehensive QA dataset that satisfies the criteria above, which leads to the following research questions:

1. How to cover all topics when creating the questions?
2. How to account for questions that cannot be answered?
3. How to automatically generate and filter

question-answer pairs starting from a set of documents?

4. To what extent can we compare LLMs' performance using generated data?

## 2. Related Work

The Question Generation (QG) branch of Natural Language Processing (NLP) has been of great interest recently due to the rising need for datasets for chatbot evaluation.

Usually, these datasets are produced based on a set of documents. In (Cohen et al., 2023), the authors use the QA pairs to form a knowledge base using the dataset Probably-Asked Questions (Lewis et al., 2021). Thus, QG can be perceived as a way of augmenting data for QA systems. The dataset mentioned before was also automatically generated for the task of Open-Domain Question Answering. It comprises 65 million questions in a four-step process composed of passage selection, span answer extraction based on named entities, question generation, and filtering (Lewis et al., 2021). The drawback of this method was that the generated answers were very brief. The solution implemented by the authors for this issue was to model the problem as a Long-Form Question-Answering one that creates open-ended questions that require explanation (Fan et al., 2019). The authors proposed a query-based multi-document summarization approach with sequence-to-sequence models.

Regarding the different types of questions, it is also possible to create them considering more sophisticated processes than reading comprehension.

As presented in (Wang, 2022), questions may be generated employing reasoning processes such as common sense, finding the most logical continuation of a sentence, or using deduction and induction given some premises to reach the correct conclusion. On the other hand, various techniques have also been proposed to remove irrelevant questions, such as n-gram similarity between question and context or scores given by another LLM to the quality of produced data (Yuan et al., 2022).

Detecting topics in texts is also one of the tasks researchers have covered the most in NLP. Some of them use Self-Organizing maps along with the k-means algorithm (de Miranda et al., 2020), others have focused on using graph nets for analyzing text embeddings (Romanova, 2021), and also the exploitation of temporal correlation in social media posts has been utilized to detect topics (Comito et al., 2019). Existing research tends to concentrate on clustering, but there are additional steps needed to transform raw text into meaningful topics.

### 3. Data

We perform experiments on a dataset of 1,320 leaflets (3,958 pages). These leaflets contain information in Dutch about different medical and administrative procedures to help patients navigate hospital services and find medical information. Figure 1 illustrates an example of one leaflet's page containing multiple sections. A section has a heading and one or more paragraphs that can include bullet points, tables, and images and can span across multiple pages or columns.

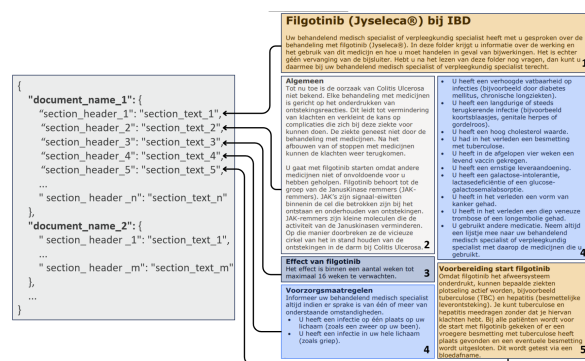


Figure 1: Format of a leaflet page

### 4. Method

We propose a multi-step framework to automate the creation process of the QA dataset, which is shown in Figure 2. In the following sections, we will discuss in detail each step.

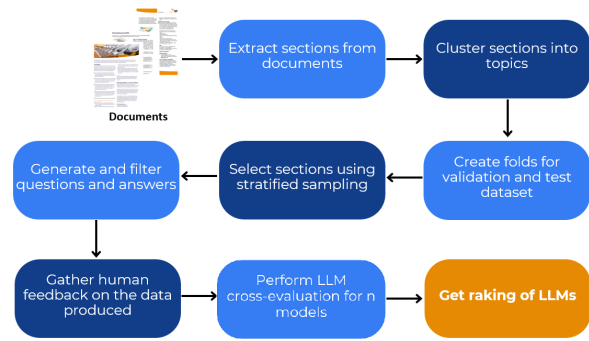


Figure 2: Steps of the proposed framework

#### 4.1. Section Extraction and Grouping

To ensure the uniqueness of the QA pairs, we create targeted questions based on sections. Questions generated on each section instead of each page or document are easier to evaluate automatically (Yuan et al., 2022). To extract these sections, we developed a PDF parser that considers font characteristics such as size and style together with the text structure and spacing between paragraphs. The sections extracted incorrectly are labeled as anomalies by BERTopic in a downstream topic detection phase and are not considered afterward. With our parser, we extracted 13,216 sections out of 1,320 digital leaflets, which are machine-readable. The algorithm is presented in Appendix 8.

The size of chunks retrieved during RAG directly influences the output quality. Rather than using sections during RAG, we introduce the concept of "emulated pages", obtained by sequentially grouping document sections. This approximates the size of the original pages while avoiding the division of sections, resulting in a median of 842 tokens and a slight deviation from the general recommended 1000-token chunk size (Rameel Ahmad, 2024).

#### 4.2. Formation of Topics

After that, we create questions based on the extracted sections. The leaflets are clustered into groups based on the topics they covered, using BERTopic (Grootendorst, 2022) - a modelling framework that extracts interpretable and concise topics. To ensure BERTopic's performance, two main hyperparameters must be considered: embedding's dimensionality and minimum cluster size.

To evaluate the quality of the clusters for each dimensionality of the embeddings, three groups of metrics are used:

- **Geometric:** Silhouette score (Davies and Bouldin, 1979a), Calinski-Harabasz (Caliński and Harabasz, 1974), and Davies-Bouldin (Davies and Bouldin, 1979b) indices.
- **Robustness:** Another interesting way to as-

ess the quality of clusters is adding noise to the data (Davidson et al., 2001), in this case, to the embeddings. Then, the higher the number of embeddings clustered into the same group with and without noise, the more robust the clusters are.

- **Document-cluster evaluation:** If we assume that only one topic is treated per document, a reasonable metric for the clusters is that all sections of a given document should ideally belong to the same cluster (except for certain exceptions, such as contact sections). If it is known that the documents contain information about different topics, this metric should be ignored. If that is not the case, the higher the matching, the better the clustering. The expression 2 yields this metric, with  $d_i$  as the metric for each document,  $s_j$  the number of sections belonging to topic  $j$ ,  $N$  the number of sections in the document, and  $M$  the number of documents in the cluster.

$$d_i = \max_j \left( \frac{s_j}{N} \right) * 100 \quad (1)$$

$$DC = \frac{\sum_{i=1}^n d_i}{M} \quad (2)$$

### 4.3. Cross-evaluation

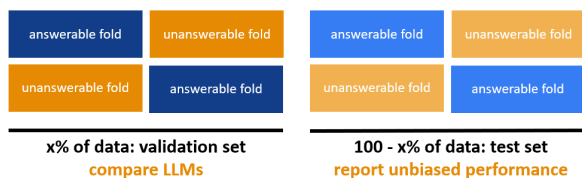


Figure 3: Cross-evaluation setup

We designed a Cross-Evaluation (CE) method for RAG-based LLMs inspired by classical machine learning techniques. This method accounts for unanswerable questions without external knowledge bases by grouping the emulated pages into disjoint information groups ("folds"). When faced with an unanswerable question, the model responds with "I don't know" instead of inventing an answer, allowing us to evaluate its hallucination level.

Figure 3 illustrates the CE setup that contains the validation and test sets. Each has two folds with a modifiable ratio of test-validation test. We use the first set to compare the performance of LLMs based on the metrics described in section 4.7, and then the second set to report the unbiased performance of the best model.

Our CE setup involves two iterations for validation and two for testing. The folds are utilized in each iteration as depicted in Figure 4. Specifically,

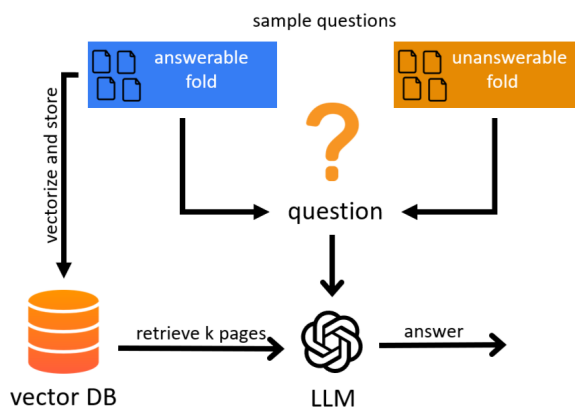


Figure 4: Use of folds while answering questions

all emulated pages from the answerable fold are stored in the vector database (e.g., Qdrant (Qdrant, 2024)), and questions are posed based on both answerable and unanswerable folds. Subsequently, the metrics presented in section 4.7 are measured for each iteration and averaged for each set, providing a comprehensive evaluation of the model's performance.

### 4.4. Creation of Folds

As presented in the section 4.2, the folds are created based on the discovered topics. We use the sections from these topics in a vectorized form represented by low-dimensionality embeddings created by BERTopic. We designed a bottom-up hierarchical approach to group the topics into folds. This grouping enables us to use stratified sampling based on topics within each fold to make the selected sections as diverse as possible. Our algorithm minimizes the probability of having overlapping information in any pair of folds by maximizing the folds' distance in the embedding space and removing the sections from common pages or documents that are in different folds.

Even though folds are represented in the CE setup by groups of emulated chunks stored in the vector database, we create the folds with sections as the atomic unit. Hence, these folds can be viewed from two perspectives:

- groups of sections from which we sample and create questions
- groups of emulated chunks used during CE that contain the sections

We choose to have two folds since we are only interested in minimizing the probability of them having overlapping information. Moreover, adding more folds implicitly reduces the distance between them, and two folds are sufficient to simulate unanswerable questions.

#### 4.4.1. Algorithm to Create the Folds

The algorithm is highly customizable, having the following parameters: ratio of folds, test set percentage, number of sampled sections for creating questions, and whether the folds should contain sections from different pages or documents. As shown in Figure 5, the procedure includes a series of steps that will be described below.

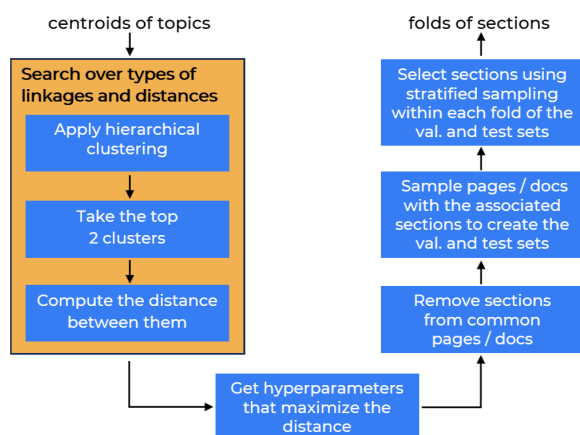


Figure 5: Algorithm for creating folds

Since topics are represented by compact clusters without outliers, as presented in section 5.2, we clustered their centroids using the agglomerative approach of the hierarchical clustering (Müllner, 2011). This approach is optimal since we can use various linkage criteria between clusters and working only with the centroids makes the computation extremely efficient. The distance between folds is maximized by searching over the space defined by the distance type (Euclidean or cosine) and linkage criteria (single, complete, average, or ward).

After clustering with each set of hyperparameters, the newly created clusters are evaluated by computing the average of the single and average linkage using the Euclidean distance. We chose this distance to account for the space between all points and, simultaneously, to weigh the distance between the closest points more.

In the next step, sections that appear in pages or documents with sections in both folds are removed to avoid having common information in both folds during the cross-evaluation. Next, we sample pages or documents and their sections at the fold level to create validation and test datasets of sections following the set ratio. This step returns two folds for validation and two for test sets.

Finally, we employ stratified sampling within each fold to select a set of diverse sections covering all topics based on which questions are created. In our setup, strata are groups of sections from the same topic.

#### 4.5. Types of Questions

For our purpose, we categorized the questions based on two different criteria: if the question is answerable based on the leaflets, and if the question is factoid or long-form.

Long-form, open-ended questions assess the machine’s ability to provide helpful advice based on its database since the chatbot is expected to interpret and explain information relevant to patient inquiries. (e.g., "What is actigraphy and how can it diagnose sleep problems in children?"). Meanwhile, factoid questions test the chatbot’s ability to accurately retrieve facts (e.g., phone numbers or email addresses). They verify if the LLM can locate precise information without hallucinating and are evaluated by a pass/fail metric. (e.g., "What is the telephone number of the radiology department?")

#### 4.6. Q&A Generation and Filtering

A two-stage approach is utilized to create the question set for long-form and factoid questions: First, a larger, diverse set of questions is generated in Dutch using GPT-3.5-turbo-instruct (OpenAI, 2022). Then, we filter the questions using embedding similarity, ROUGE score (Lin, 2004) between the answer and the section, and sorted by a score assigned by the model based on examples. Top questions are selected per source section to retain the distribution of QAs concerning the data in the folds. We present the details of this algorithm in Appendix 8.

For phone numbers and email addresses from leaflets, we use the same two-stage mentioned above to generate a question for each section and the entity extracted from it, with a modification: rather than assessing section-answer similarity, we directly verify the presence and accuracy of the email/phone number in both the answer and the section.

In the first stage, 5000 QA pairs are generated. The filtering steps reduce this to 500 pairs, with each final question corresponding to one sample section to avoid distorting the previous distribution of samples.

#### 4.7. Measured Metrics

Initially, hospital specialists will perform a qualitative evaluation of the generated QA dataset. Only the QA pairs labelled as correct are used in the cross-evaluation procedure to test various LLMs.

Secondly, answers should be analyzed quantitatively. We compute the hallucination rate for every answer - the percentage of unanswerable questions that would have been answered without the proper information. However, these events could not only be due to hallucination but also to a poor

division in the folds. Therefore, we measure the percentage of answers where the correct information has been given to evaluate factoid questions. For long-form questions, we use the standard metric for the long-answer questions: BLEU, ROUGE-1, ROUGE-2, ROUGE-L, and BLEURT (Celikyilmaz et al., 2020). Finally, humans evaluate the correctness of an answer.

#### 4.8. Using Folds to Evaluate Different Models

A Dutch language model (given the data is in Dutch), Llama-2-13B-Dutch with 8-bit quantization (Vanroy, 2023) is human-evaluated to determine how the method, and specifically the different folds, can be used in order to determine how well certain LLMs perform. These are separated into three distinct classes:

- The performance on long-form answerable questions (folds 1 and 2). This performance is the percentage of correctly answered questions evaluated by a human.
- The performance on long-form unanswerable questions (with disjoint folds  $A$  and  $B$ , the context of fold  $A$  is used for questions from fold  $B$  and vice versa). In this way, our method can be used to determine to what extent a model can indicate that there is no answer to the question based on the given context. This metric is defined as the percentage of questions answered by the model with an answer that makes it clear that the model does not have enough information to answer the question evaluated by a human. Seventy-eight annotations are made for both combinations of disjoint folds.
- The performance on answerable factoid questions. A fold  $F$  containing only factoid questions can be used to determine whether a model can extract said factoids well. This performance is defined as the percentage of questions that are answered by the model with an answer that correctly extracts the factoid information evaluated by a human. Five annotations were made, given that five factoid questions were present in the dataset.

## 5. Results

### 5.1. Formation of Topics

The performance of BERTopic with different values for the hyperparameters "min\_cluster\_size" of the HDBSCAN and "n\_components" of the UMAP algorithm is compared: 5, 10, 15, 20, 25, 30, 35, 40, 45, 50.

We set the number of topics to 200 a priori, considering that the number of leaflets and too many clusters would result in too numerous topics. A

higher embedding dimensionality implies a larger accuracy of the clustering algorithm, but less information covering the section's content will be kept. Therefore, dimensional values under 10 are hardly acceptable, so a trade-off solution should be chosen. All this a priori knowledge goes along with the results of the metrics, which are enumerated in the following list:

- **Geometry metrics:** Depends on the minimum cluster size: better results for small values (up to 35). There is no clear dependence on the embedding dimensionality.
- **Robustness metrics:** Better for intermediate values of minimum cluster size (between 20 and 30) and highly depend on the dimensionality of the embedding, where high dimensions (over 15) result in lower robustness. Here, results for high values of both variables are deceptively good only due to the lower quantity of clusters. The same behavior is observed in the document-cluster matching score.
- **Document-cluster metrics:** Highly dependent on the minimum cluster size (which was predictable). It is better to have a more increased value, up to the extreme case of too few clusters—no apparent dependence on embedding dimensionality.

The interest lies in exploring different values for hyperparameters, leading to the need for a trade-off solution. While results on semantic coherence suggest favoring fewer topics, geometry-based calculations indicate that more topics would better represent document information. Ultimately, we decided to increase the embedding dimensions from the default 5 to 15. Similarly, the "min\_cluster\_size" will be adjusted from the default 10 to 30 elements. The number of clusters typically hovers around 70 topics, which will be discussed in the following section. Outliers are not a significant concern since they remain within acceptable proportions, as demonstrated in the implementation example of BERTopic. Additionally, the abundance of labelled sections mitigates concerns about outliers.

These chosen hyperparameters balance geometric robustness metrics and the clustering of sections within the same document, ensuring effective representation.

### 5.2. Topic Analysis

This section analyzes two sets of topics created with BERTopic. The main difference between them is the transformer that creates each section's embeddings. The first one is built using distiluse-base-multilingual-cased-v1 (DBMC-v1), a distilled version of the model presented in (Yang et al., 2019). The second one is based on paraphrase-multilingual-MiniLM-L12-v2 (PMM-L12-

v2), a multilingual version of paraphrase-MiniLM-L12-v2 (Reimers and Gurevych, 2019).

These topics are composed of sections represented as embeddings with 15 dimensions. Their attributes are described in Table 1.

statistic	DBMC-v1	PMM-L12-v2
min topic size	32	30
median topic size	66	79
max topic size	634	1063
topics no	71	62
outliers %	40.55%	32.17%

Table 1: Statistics of the generated topic clusters

Many sections are detected as outliers and not included in any topics, leading to very compact clusters represented well in space by their centroids.

Lastly, the data does not reveal any correlation (-0.032 or -0.051, depending on the set of topics) between cluster size and the minimum distance from any cluster to the closest one. This indicates that topics are not isolated based on size; they are all positioned randomly in space.

### 5.3. Fold Analysis

As we have a parameterizable algorithm, we chose a test set percentage of 20% and specified that the two folds should contain sections from different documents. After running the procedure presented in section 4.4 with each set of topics created with BERTopic, we found folds with enough sections to build the validation and test sets only in the case of four combinations of hyperparameters.

As shown in Table 2, the best results are obtained for both sets of topics when the ward linkage criterion is used together with the Euclidean distance. The space between the folds increased between 6.7 and 8.21 times compared to the initial distance between the topics.

set of topics	DBMC-v1	DBMC-v1	DBMC-v1	PMM-L12-v2
linkage criterion	ward	complete	complete	ward
distance type	Euclidean	cosine	Euclidean	Euclidean
min dist topics	0.2783	0.2783	0.2783	0.2481
min dist folds	2.2882	1.923	1.8662	1.9548
small fold ratio	0.1175	0.3741	0.1574	0.4727
avg folds per doc	1.2104	1.3285	1.1208	1.3399
valid sections %	72.12%	55.84%	84.22%	55.28%

Table 2: Results of the hierarchical clustering

Our approach is better than forming a fold based on the most isolated clusters because we create them based on more topics. Regarding fold ratio, the folds are almost even in the case of PMM-L12-v2. The last two rows of Table 2 refer to the number of valid sections from the perspective that the folds contain sections from different documents. A larger average fold per document implies that more sections must be filtered out. The remaining sections

are used to compose the 80-20 validation-test split, followed by the stratified sampling step.

These sampled sections are the basis for creating questions rated by professionals. A large enough and agreed-upon size to assess was 500. The sections are selected concerning the folds ratio rounded to the first decimal, meaning that for PMM-L12-v2, we have a ratio of 0.5. This leads to the next sizes: validation fold 1 - 200, validation fold 2- 200, test fold 1- 50, and test fold 2 - 50.

To reduce human effort and focus on the higher-quality set of questions, we will only use the sampled sections in the case of PMM-L12-v2. In the other scenario, even though the distance between the folds is larger, the second fold is too small, resulting in similar sections that do not cover the entire scope of the information.

The final evaluation of these folds will be performed during CE. That is the final check if any overlapping information between folds is present.

### 5.4. Human Evaluation

Five hospital specialists annotated these 490 questions, the generated answers, and the reference section. For each QA pair, the annotators were told to choose one or more options: irrelevant question, too specific question, wrong answer, incomplete or ambiguous answer, correct answer, and, optionally, to write a short feedback. Out of the total questions, 85 were double-annotated for quality control, one wasn't evaluated, and the remaining were reviewed by a single random annotator from the pool of five. The annotators jointly agreed in 64.71% of cases and partially agreed in 15.29% of cases, meaning that they picked multiple options, of which at least one is the same.

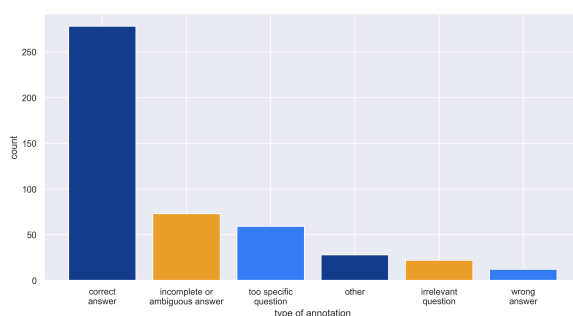


Figure 6: Distribution of annotations

Figure 6 shows the distribution of annotations for all questions except the 17 questions that produced a disagreement. A total of 278 questions were labelled as having correct answers. In 28 cases named "other", the annotators did not pick any predefined option but wrote a comment. Mostly, comments refer to the fact that the question or answer is too specific, the answer is incomplete, the



question formulation is strange, or the answer tone is offensive in only one instance. Out of the reviewed set, in only 12 cases, a "wrong answer" was generated. This low number validates the quality of the generated data.

Based on this feedback, we filtered the questions to use only the ones labelled as having correct answers in the cross-evaluation procedure.

## 5.5. Cross Evaluation Results

We used the dataset for cross-evaluation on the GPT-3.5 Turbo model using 209 question-answer pairs from two training folds, 76 from fold one and 133 from fold two. The evaluation assesses the model's information retrieval capability and response appropriateness along various metrics, which are presented in Table 3.

Metric	Fold One	Fold Two
Using Fold One RAG (%)	57.33	7.52
Using Fold Two RAG (%)	12.00	62.41
BLEU Score	0.0734	0.0777
ROUGE-1 Score	0.312	0.261
ROUGE-2 Score	0.188	0.154
ROUGE-L Score	0.259	0.222
BLEURT Score	0.592	0.517
Facts Correctly Extracted (%)	20	-

Table 3: Cross-evaluation results for GPT-3.5 Turbo

### 5.5.1. Evaluation on Llama-2-13B-Dutch

The Llama-2-13B-Dutch model has been evaluated using the folds to determine its performance on questions it cannot answer. We consider a valid reply if the model states that there is not enough information to answer. Furthermore, the factoid fold has been used to determine how well the model can extract data from a context containing specific factual information. These results can be seen in Table 4, where Q1\_CTX2 represents fold 1 with context from disjoint fold 2, Q2\_CTX1 represents fold two with context from disjoint fold one, and the Factoids Corr. % represents the percentage of factoids correctly extracted by the LLM.

Q1_CTX2 Corr. % (n=78)	Q2_CTX1 Corr. % (n=78)	Factoids Corr. % (n=5)
0	0	100

Table 4: Evaluation of unanswerable and factoid questions

On the total 278 correct determined questions, which are answerable, the model was further tested. The results of the numerical evaluation are shown in Table 5.

### 5.5.2. Answering When the Corresponding RAG Pages are Loaded

To assess the system's ability to retrieve information correctly and to decline to answer in case no information is available, we did a cross-evaluation procedure presented in section 4.3. For this, we expect that if the corresponding fold is loaded, the model should try to answer all questions, while if the unrelated documents are used, it should not answer any of the questions.

The results are found in the first two rows of Table 3. Declines to answer were either hard-wired from the failure of the retrieval or manually labeled if the model declined to answer (even though the retrieval gave some unrelated results).

## 5.6. Human Evaluation

With the human evaluation done by medical specialists, we can deduce that automatic question and answer generation is a feasible way to create relevant questions, as around 83,46% of questions were considered by them as being relevant. Many cases were flagged as irrelevant or having too specific questions, which might require adjusting, but it also gives useful feedback on the level of specificity required for this type of chatbot. Having over 56.73% fully correct QA pairs means that a reasonable portion can be used directly for evaluating LLMs.

## 5.7. Cross-evaluation

### 5.7.1. Answerable and Unanswerable Questions

The results show that the chatbot is unwilling to answer around 40% of the answerable questions. While we might need to consider that it is affected by the style of questions we had, this leaves room for improvement in the system, most likely in the retrieval.

For cases when the data is not available, the chatbot correctly declines to answer around 90% of the time, which exceeded our expectations. Considering that, seemingly, in most of the cases, it was not due to the retrieval not giving any results, the LLM decided that it did not have enough information.

### 5.7.2. Performance Metrics

All performance metrics should be treated as a baseline for comparison with other models; on their own, they might not give a clear picture of the answer quality.

The results around 0.07 for BLEU are low, probably due to the LLM's tendency to rephrase the content, resulting in low N-gram overlap.

	<b>BLEU Score</b>	<b>BERTScore Precision</b>	<b>BERTScore Recall</b>	<b>BERTScore F1</b>	<b>ROUGE-1 F-measure</b>	<b>ROUGE-L F-measure</b>
Results for tested models on verified dataset (n=278)						
Llama-2-13b-Dutch	17.9	0.761	0.833	0.793	0.466	0.417

Table 5: Llama2 results for answerable questions

The ROUGE scores vary between 0.15 and 0.3, with Rouge-2 scores being the lowest. The 0.3 score might seem acceptable, but it needs to be used as a comparative value.

In examining the BLEURT score, we recognize that it got the highest values, with its 0.5-0.6 ratings. As this metric is trained to better correlate with human judgment, having a satisfactory rating will give a better comparison later, once it can be compared with other variants of the chatbot.

### 5.7.3. Factoid Questions

In this case, the chatbot underperformed by not finding the correct address, although this is statistically insignificant since the number of factoid questions was low.

### 5.7.4. Evaluation on Llama-2-13B-Dutch

The evaluation of the Llama-2-13B-Dutch language model utilizes different folds to assess its performance on answerable, unanswerable, and factoid questions. Tables 2 and 5 demonstrate the effectiveness of this method in evaluating model performance across these aspects. Specifically, the fine-tuned Dutch model struggles to identify unanswerable questions when contextually lacking necessary information. However, it excels in extracting factoid details and performs well on answerable questions.

## 5.8. Limitations and Challenges

The proposed framework is limited; thus, the resulting folds can be unbalanced. The PDF parser that extracts the sections works only on a specific type of leaflets and should be extended to be more general. We have not found a way to ensure an approximate number of sections in each fold. Another issue is that it requires humans in the loop to evaluate the QA pairs used during cross-evaluation. Additionally, we create factoid questions only related to named entities such as phone numbers and email addresses, while many more facts are present in the leaflets. Additionally, regarding the framework’s usability in assessing the performance of models in answering answerable, unanswerable, and factoid questions, an example containing two folds of 78 questions, answers, and context triples is quite limited. Furthermore, there were only five

factoid questions in the respective fold. Despite these examples still showing the method’s potential, said sample sizes are relatively small, and fold-specific evaluation has only been done for one model. Future work expanding the evaluation of this framework is therefore welcomed.

## 6. Conclusion

This research paper presented a framework for evaluating RAG-based chatbots from a set of documents by automatically generating QA pairs and employing a cross-evaluation procedure that accounts for unanswerable questions. Our method enables the comparison of LLMs using various metrics for assessing long-form and factoid questions. The human evaluation results highlight the quality of the produced QA pairs, with 83.46% relevant questions and only 2.44% wrong answers. Moreover, although there are various limitations, we successfully demonstrated that our framework can be used to evaluate LLMs such as Llama-2-Dutch-13B or GPT-3.5 Turbo with a dataset of hospital leaflets for patients. The project’s source code and the created dataset are publicly available at this [link](#).

## 7. Acknowledgement

We acknowledge the Ziekenhuis Groep Twente (Hospital Group Twente - ZGT), Hengelo, The Netherlands for providing us with the leaflets and the computing power, as well as for manually annotating the generated QA pairs.

## 8. Bibliographical References

- Tadeusz Caliński and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.
- William W Cohen, Wenhui Chen, Michiel De Jong, Nitish Gupta, Alessandro Presta, Pat Verga, and

- John Wieting. 2023. Qa is the new kr: Question-answer pairs as knowledge bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15385–15392.
- Carmela Comito, Agostino Forestiero, and Clara Pizzuti. 2019. [Word embedding based clustering to detect topics in social media](#). In *IEEE/WIC/ACM International Conference on Web Intelligence, WI '19*, page 192–199, New York, NY, USA. Association for Computing Machinery.
- George S Davidson, Brian N Wylie, and Kevin W Boyack. 2001. Cluster stability and the use of noise in interpretation of clustering. In *Information Visualization, IEEE Symposium on*, pages 23–23. IEEE Computer Society.
- David L Davies and Donald W Bouldin. 1979a. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227.
- David L. Davies and Donald W. Bouldin. 1979b. [A cluster separation measure](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227.
- Guilherme Raiol de Miranda, Rodrigo Pasti, and Leandro Nunes de Castro. 2020. Detecting topics in documents by clustering word vectors. In *Distributed Computing and Artificial Intelligence, 16th International Conference*, pages 235–243, Cham. Springer International Publishing.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. *arXiv preprint arXiv:1907.09190*.
- Maarten Grootendorst. 2022. [Bertopic: Neural topic modeling with a class-based tf-idf procedure](#).
- LangChain, Inc. 2024. Output-fixing parser. [https://python.langchain.com/docs/modules/model\\_io/output\\_parsers/types/output\\_fixing](https://python.langchain.com/docs/modules/model_io/output_parsers/types/output_fixing). Accessed: 2024-02-02.
- Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*.
- OpenAI. 2022. GPT-3.5 Turbo: Language Model. <https://platform.openai.com/docs/models/gpt-3-5-turbo>.
- Qdrant. 2024. [Qdrant vector database](#).
- Syed Rameel Ahmad. 2024. Enhancing multilingual information retrieval in mixed human resources environments: A rag model implementation for multicultural enterprise. *arXiv e-prints*, pages arXiv–2401.
- Nils Reimers and Iryna Gurevych. 2019. [Sentencebert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Alex Romanova. 2021. Detect text topics by semantics graphs. In *Proceedings of the 2nd International Conference on Blockchain and Internet of Things (BloT 2021)*, volume 11.
- Bram Vanroy. 2023. [Language resources for dutch large language modelling](#).
- Zhen Wang. 2022. [Modern question answering datasets and benchmarks: A survey](#).
- Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.
- Xingdi Yuan, Tong Wang, Yen-Hsiang Wang, Emery Fine, Rania Abdelghani, Pauline Lucas, H el ene Sauz eon, and Pierre-Yves Oudeyer. 2022. Selecting better samples from pre-trained llms: A case study on question generation. *arXiv preprint arXiv:2209.11000*.

## Appendix A: PDF Parser for Section Extraction

The algorithm for parsing the PDF leaflets has the following steps:

1. set manually the area of interest that contains the text and excludes the header and footer with metadata such as page number and date
2. find the majority font size associated with each line of text

3. divide the PDF into groups of paragraphs, called sections, based on two delimiters: if the space between them is larger than average, or if a line of text written in a larger-than-average font is encountered
4. handles the case where entire paragraphs are written in a larger font, and each line is recognized as a separate section, merging them into a section without a header
5. merge sections that do not have a header with the previous section and header-only sections with the following sections

Our approach works even in edge cases, correctly separating sections that span multiple pages, contain bullet points, or have paragraphs delimited in various ways. The goal is to extract all sections while minimizing the number of detections as a section of a group of sections or a part of a section.

The parser was designed specifically for this dataset while testing it on 11 representative PDFs, including a comprehensive set of edge cases. It was then manually verified with a larger random sample

## Appendix B: Generation of Question-Answer Pairs

### B.1 Long-Form Questions

An instruction-tuned LLM was utilized to produce a wide range of naturally formed questions without much constraint on the type of questions created. However, the quality of the questions is highly dependent on the model's performance on these tasks. A factor that made the task more difficult was that the questions needed to be generated in Dutch. We used GPT 3.5 for this purpose, which does support Dutch.

For the language model to give us the needed text, we needed to create instructions that precisely explained the task. This was done by creating a custom prompt for the task and progressively improving it. The prompt was created in Dutch, as this seemed to cause the LLM to reliably continue using the language upon asking for completion. The prompt describes the main task and "domain" and includes the selected section.

Since we had a few examples of question-answer pairs, we used them to employ a few-shot prompting techniques. This helped the model find the right tone and length for the reference answers.

As another factor to increase the variance of the questions, we added a few random roles like "recovering patient", "elderly patient", or "parent of a sick

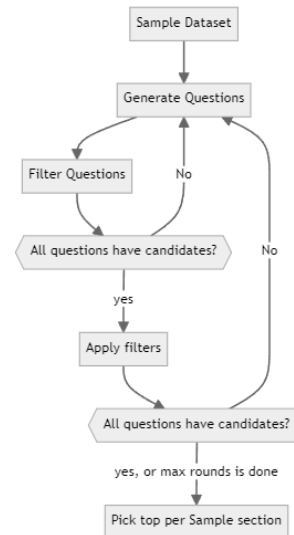


Figure 7: Long-form question generation workflow

child" so that the model creates more varied scenarios. During use, certain additions were made to the prompt to create more suitable questions.

To make the output machine parseable, we decided on a JSON scheme. The model struggled to follow these precisely enough, so the scheme description and instruction, but repeating the instructions at the beginning and end of the prompt, seemed to improve the rate of successful generation. Even with the changes, we still had several cases where the model failed to follow the scheme properly. As an additional step, "output fixing LLM"(LangChain, Inc., 2024) was employed to transform the faulty answers to the scheme. This was done since fixing the format was less costly than retrying the generation, and fixing the scheme was more reliable than the generation.

**Filtering** During the generation stage, we created more questions than we needed; however, most were not unique or high-quality enough.

We used cosine similarity on the question embeddings to filter out repeating or similar questions. Then, we executed a "drop out," where we discarded one for highly related questions until we reached the desired threshold.

We tried to ensure that the generated Q-A pair was related to the source. Unfortunately, there were quite a few cases where the model got "inspired" by the few-shot example and created content related to that over the context. We used ROUGE as a similarity metric with a low threshold on the generated answer and the context.

We needed to remove the questions that could've been considered "Short-form factoid" questions to avoid accidentally mixing the two types. To remove these, a basic rule was implemented. All questions that had a short answer and contained one of the Entity types we chose to extract were discarded.

## B.2 Factoid Questions

We wanted to ask some highly targeted questions to evaluate the chatbot's ability to recite small information sections. This was necessitated because every time the model could not answer, the user might ask for a way to contact a human, which, in the hospital's case, would be the already existing contact phone number or address related to the topic. Many of these direct contacts were already included in the flyers we used.

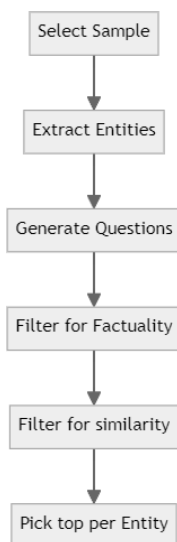


Figure 8: Factoid question generation workflow

**Entity Extraction** For extracting entities of interest, we experimented with some different options. Even though there are many Named Entity Recognition software, such as Flare, we decided to only include the most basic regex-based "phone number" and "email" extractions, as these types of data created direct questions that were suitable for our purpose.

**Generation** Similarly to the previously mentioned long-form generation, we used an LLM. The main difference was in the instruction. In addition to the context, the selected fact was provided, and all other details and examples were modified to fit the new format.

**Filtering** The questions were filtered by doing a back check on the answer, verifying that the original "fact" is extractable.

Similarly to the long-form questions, the factoid questions were filtered by their embeddings' cosine similarity.

The correctness of the extracted entities was verified at the end of this step and during the human evaluation step.

# Speech Accommodation in Health-Care Interactions: Evidence Using a Mixed-Reality Platform

Rose Baker<sup>1</sup>, Susan C. Bobb<sup>2</sup>, Dai'Sha Dowson<sup>1</sup>, Elisha Eanes<sup>1</sup>, Makyah McNeill<sup>1</sup>, Hannah Ragsdale<sup>1</sup>, Audrey Eaves<sup>1</sup>, Joseph G. Lee<sup>1</sup>, Kathrin Rothermich<sup>1</sup>

<sup>1</sup>East Carolina University, <sup>2</sup>Gordon College

<sup>1</sup>E 5th Street, Greenville, NC 27858, <sup>2</sup>255 Grapevine Road, Wenham, MA 01984  
{bakerr22, eanese23, mcneillm17, ragsdaleh18}@students.ecu.edu, susan.bobb@gordon.edu,  
{dowsond23, eavesa16, leejose14, rothermichk17}@ecu.edu

## Abstract

Many people in the US use more than one language at home, yet English remains the dominant (L1) language in US society, which can complicate medical encounters. In this study we ask in what ways effective communication can be ensured in health care settings when speakers differ in language proficiency. One strategy people use is second language (L2) speech accommodation, which is characterized by slowed speech, less complex words, and clearer enunciation. We employ a mixed-reality platform called MURSION to document how a group of Physician Assistant students use speech accommodation during a healthcare encounter. MURSION is a computer-based virtual environment where participants interact with an Avatar controlled by a human interactor in a standardized environment. We record 5-minute interactions between the student and a high or low English proficiency Avatar. Our analyses evaluate lexical choices in L1-L2 interactions with SCOPE (South Carolina Psycholinguistic Metabase) and acoustic properties with PRAAT. Results show that clinical students use slower speech and high frequency words when speaking to a low proficiency virtual patient, indicating a sensitivity for the communicative needs of L2 English users. Speech accommodation results will contribute to communication training modules for clinicians to interact efficiently with linguistically diverse populations.

**Keywords:** communicative needs, foreigner-directed speech, patient care, MURSION, PRAAT

## 1. Introduction

In the US, many people speak English as a second language, and minority language usage is on the rise. Even so, English remains the standard language in most places in US society, including healthcare environments. In these contexts, minority language users are especially vulnerable to miscommunications. While access to translators can mitigate misunderstandings, limited funding and resources restrict their availability.

One strategy that dominant English speakers (L1) apply to increase the chances of successful communication with second-language users (L2) is speech accommodation (Ferguson, 1975). L2 speech accommodation is characterized by slowed speech, use of easier words, and exaggerated enunciation. Self-reports suggest that L1 speakers often use speech accommodation with L2 speakers to ensure smooth communication. In a recent survey study, Rothermich et al. (2023) found that half of L1 respondents used accommodative speech to facilitate understanding. Though L1 speakers may intend to facilitate communication, L2 impressions suggest the effect of accommodation is not always positive. In the same study by Rothermich and colleagues (2023), almost half of L2 respondents said they do not appreciate the use of speech accommodation. More research is needed on the actual effectiveness of L2 speech accommodation. One of the few studies that has

directly investigated whether accommodation supports comprehension is a recent study by Piazza et al. (2023). Results suggested that language learners exposed to accommodative speech recognized new words faster and had better pronunciation than learners who received non-accommodated speech. Accommodated speech may be helpful for L2 speakers, but little is known about how speech accommodation helps L2 speakers successfully navigate the world outside the language learning classroom.

It is necessary to extend research on speech accommodation to the healthcare environment. In what ways do clinicians adapt their language when speaking to a person with limited proficiency? How do multifaceted features of L1 accommodation work together in conversation to sensitively meet the needs of the L2 speaker? We will investigate how L1 students with clinical training adapt their speech to be more accessible to L2 speakers at multiple level levels of language simultaneously.

## 2. Related Work

Previous research in English has shown that L1 speakers modulate their accommodation depending on perceived level of L2 English proficiency. Speech may be adjusted lexically or acoustically. Specifically, lexically accommodated speech may include lower average word frequency or fewer idiomatic expressions (Scarborough et al., 2007), and L1

speakers may repeat themselves more, clarify more often, and produce speech with a higher type/token ratio and lower mean length of utterance generally (Henzl, 1979). Relatively few studies have addressed lexical accommodation (see Rothermich et al., 2019 for a recent review), but acoustic L2 speech accommodation is well-documented: L1 speakers will speak more slowly, pause for longer (Pardo, 2006), adjust their fundamental frequency (Babel & Bulatov, 2012), duration, and intensity (Rodriguez-Cuadrado et al., 2018) and converge to the amplitude envelope of their L1 interlocutor (Lewandowski & Jilka, 2019). Additionally, the lexical and acoustic aspects of accommodation interact; vowel duration changes according to the semantic density of a word as well as its contextual predictability (Scarborough, 2010).

Accommodation is not uniform across speakers and situations. L1 speakers tend to change their speech more dramatically with L2 speakers who are less proficient (Chaudron, 1983; Dahl, 1981). L1 speakers also self-report evaluating the L2 proficiency level of their interlocutor to gauge their need for accommodation (Margić, 2017). Kühnert and Antolík (2017) found that L1 speakers used slower speech rates with less proficient L2 speakers, while they used a faster speech rate with more proficient L2 speakers. The properties of speech accommodation may therefore vary based on perceived L2 proficiency.

L1 accommodation may not always be appropriate due to the fact that spoken L2 proficiency is an unreliable indicator for comprehension ability. Previous research confirms that L2 learners frequently comprehend the target language at a higher level than their speaking ability (Litcofsky et al., 2016). Due to a discrepancy between receptive and expressive language skills, L1 speakers may use speech accommodation unnecessarily. Consequently, accommodation could be negatively perceived as condescending by the L2 speaker if it is not needed (Zuengler, 1991).

### 3. Current Study

With the advent of new technologies, such as virtual reality tools, it is now possible to capture speech accommodation dynamics and control for variables such as speaker characteristics. In this experiment, we innovate a new way of studying real-time language use by using a mixed-reality platform called MURSION. In MURSION, the study participant to interact with an avatar, controlled by a human interactor. It has mostly been used in immersive classroom simulations (Hudson et al., 2019) and to study clinical counseling (Harper, 2019). One advantage of the virtual environment is that features of the avatar, such as age, gender, race, accent, or medical history, can be standardized across sessions. We use this technology to document L2 speech accommodation in a healthcare setting. To our

knowledge, we are the first to apply it to investigate linguistic accommodations.

We invited Physician Assistant and Nurse Practitioner students (n=17) to take part in the study. We targeted clinical students for two main reasons: 1) to assess their interaction with the MURSION tool to pilot it as a method of clinical education, and 2) to test a population of emerging clinicians who might have limited clinical experience with L2 speakers. We recorded their brief interactions with the avatar patients using Zoom and analyzed their speech based on acoustic and lexical features. We expected L1 speakers to accommodate low proficiency L2 speakers by raising their mean pitch, increasing their pitch range and speaking more loudly and slowly.

**Figure 1.** Screenshots of the virtual avatar patients. Left = Avatar 1, high proficiency; Right = Avatar 2, low proficiency.



### 4. Methods

Data collection is ongoing to meet our anticipated sample size of 60. So far, we have analyzed data from 17 L1 clinical students (age range = 24 – 42; 3 male/ 14 female) who interacted with two avatars over two sessions in MURSION, (see **Figure 1**). Our MURSION paradigm simulates a patient-provider interaction using two curated scripts developed by a clinical associate professor. Students are instructed to perform an “intake” interview with their new patients, the avatars. These interviews include standardized questions to establish care, known health issues, and family medical history. The order of encounters is counterbalanced by L2 proficiency level across participants. All L1 student participants are paired with both avatars. Interactions take about 5-10 minutes per participant. The participants are not informed about the proficiency manipulation until after the study.

Both avatars had the same medical issue in each respective student encounter. Avatar 1 displays high L2 proficiency (subtle Hispanic accent, advanced vocabulary, little to no grammatical errors) and complained of heartburn. Avatar 2 displays low L2 proficiency (heavy Hispanic accent, limited vocabulary, frequent grammatical errors) and presents with a sprained ankle. The same interactor controlled Avatar 1 and 2. Both avatars present as females and appear to be in their 20s.

### Recording transcriptions

To analyze acoustic and lexical parameters of interest, the recordings need to be broadly transcribed at the word level. Each recorded interaction between the L1 speaker and the two avatars is independently transcribed by 8 transcribers (RB, DD, HR, and trained research assistants).

### Data analysis

Data were analyzed considering acoustic and lexical aspects of language. Statistical analyses will uncover significant acoustic and lexical accommodation patterns among L1 speakers that vary according to L2 proficiency.

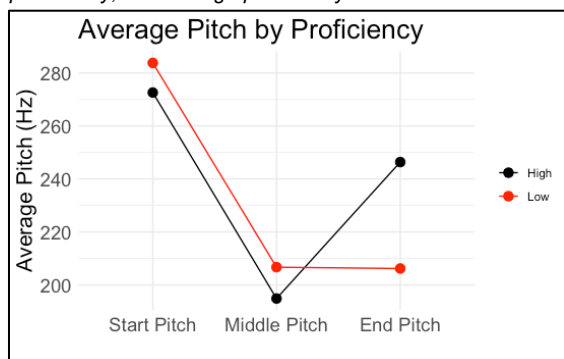
Acoustic data from 17 subjects were extracted from recordings with custom scripts in Praat, a speech analytical software (Boersma, 2002). We extracted acoustic features that are known to reflect speech accommodation, such as pitch measures (mean pitch, pitch range), intensity, and speech rate. We hypothesize that L1 speakers will accommodate low proficiency L2 speakers by maintaining their mean pitch, increasing their pitch range, and speaking more loudly and slowly.

Lexical data from 15 subjects will be extracted from transcription files with the South Carolina Psycholinguistic metabase (SCOPE; Gao et al., 2023). These data will consist of lexical frequency, contextual diversity, familiarity, and semantic diversity. We expect that L1 speakers will use words that are more common, familiar, and semantically flexible with low proficiency L2 speakers.

## 5. Preliminary Results

Preliminary findings from the high and low proficiency encounters suggest emerging L2 speech accommodation patterns in the MURSION testing environment.

**Figure 2.** Results for mean pitch over time. Red = low proficiency, black = high proficiency.



**Acoustic results.** A one-way ANOVA was conducted to examine the relationship between proficiency levels and four speech characteristics: speech rate, mean pitch, intensity, and word

fluency (see Table 1). The results indicated a significant effect of proficiency levels on speech rate ( $F(1, 28) = 4.69, p = 0.039$ ), but no significant effects on mean pitch ( $F(1, 28) = 0.635, p = 0.432$ ), pitch range ( $F(1, 28) = 0.721, p = 0.403$ ), or intensity ( $F(1, 28) = 2.396, p = 0.133$ ). Consistent with our hypotheses, speech rate was lower when participants spoke with the low proficiency avatar.

**Table 1.** Results for the acoustic analysis.

	Avatar Proficiency	
	High	Low
Mean Pitch (Hz)	201.17	212.19
Pitch Range (Hz)	575.12	579.82
Mean Intensity (dB)	60.57	56.88
Speech rate (syll/sec)	4.58	4.11

An analysis of pitch over time revealed an interesting pattern (Figure 2). While pitch patterns are similar at the beginning and the midpoint of the interaction, participants seem to speak with a higher pitch towards the end when speaking to a high proficiency avatar. An analysis of the end pitch values revealed no significant differences.

**Lexical results.** A one-way ANOVA was conducted to examine the relationship between proficiency levels and four lexical characteristics: word frequency, contextual diversity, familiarity, and semantic diversity (see Table 2). The results indicated a significant effect of proficiency levels on word frequency ( $F(1,19) = 60.07, p < .001$ ), contextual diversity ( $F(1,19) = 30.21, p < .001$ ), and semantic diversity ( $F(1,19) = 12.45, p < .001$ ) but no significant effects on familiarity ( $F(1, 19) = 0.669, p = 0.424$ ). Consistent with our hypotheses, the participants used significantly more common, contextually and semantically flexible words with the low proficiency avatar.

**Table 2.** Results for the lexical analysis.

	Avatar Proficiency	
	High	Low
Frequency (Lg10)	4.23	4.46
Contextual Diversity (Lg10)	3.48	3.6
Familiarity (z value)	2.31	2.29
Semantic Diversity	2.00	2.04

## 6. Conclusions and Ongoing Work

Our hypotheses were partially confirmed, since the data collected here demonstrate that health care providers in training adjust their speech when interacting with the avatar who displays low levels of English proficiency. Specifically, we found providers using slower speech rates in these health care encounters with the low



proficiency avatar than with the high, ostensibly to improve comprehension in this medical context. No other acoustic measures (i.e., pitch and intensity) suggested differences. Similar to our results, previous research has also shown reduced speech rates in accommodative speech (Scarborough et al., 2007). Although not significant, the data patterns for mean pitch over time suggest that speakers adjust in real-time to their speech partner. High pitch is typically associated with friendliness and conversational engagement (Chen et al., 2001), so it is interesting that we note an overall decrease in pitch toward the lower proficiency avatar across the interaction compared to the higher proficiency avatar. It is too early to tell how this change interacts with other measures of accommodation, but one possibility is that interactions across proficiency levels are effortful, taxing working memory by increasing cognitive load as the more proficient speaker monitors comprehension and speech of their speech partner while adjusting what they are saying to be comprehensible. The observed decrease in pitch could indicate disengagement with the conversation due to the required effort of the interaction. As data collection continues, we expect that our acoustic findings will grow.

Results from our lexical measures further support our hypotheses. We found that student word choice became more frequent, contextually diverse, and semantically diverse with a low proficiency English speaker. These adjustments reflect sensitivity to the vocabulary range of the patient. Previous inquiry into lexical accommodation among language teachers measured word frequency, neighborhood density, and simplification supports our findings that L1 speakers use structurally more simple language with their L2 interlocutors. Lexical choices can also be intentional with the aim of supporting effective communication (Quené, 2008). Lexical adjustments are also highly related to discourse management in L1-L2 interactions, which may contribute to changes of the structure and content of L1 language (Henzl, 1979). The intake task assigned to our participants required them to ask questions of their patient and provide clinical advice, so they tended to use clarifying questions and check for understanding often. These discourse tasks may be similar to the social role of a language teacher who tends to use longer wait times for responses and more frequent conversational repair (Weizheng, 2019). Our study is one of the first to extend lexical accommodation results to the healthcare environment, showing that clinicians use accommodation strategies to communicate effectively and manage discourse with low-proficiency L2 speakers of English.

Our results are consistent with the previous literature, including studies with more naturalistic settings, capturing some of the modulations of speech patterns that occur when more proficient

speakers of English direct language toward less proficient speakers. They suggest that virtual reality environments such as MURSION could create potentially effective training environments to document emerging cross-linguistic and cross-cultural interactional behaviors for healthcare providers who are still in training. A full sample size is needed to confirm the patterns of our initial analyses. Future work will assess whether educational interventions for clinicians can shift conversational patterns toward accommodation styles that are more effective in supporting conversations across different levels of language proficiency.

## 7. Acknowledgments

This research was supported by a National Science Foundation (NSF) Grant (#2141281) awarded to KR, SCB, JGL, and AE. Thank you to the University of West Georgia for hosting MURSION. Additional thanks to Terrie Ponder and Nina for their support in data collection.

## 8. Bibliographical References

- Babel, M., & Bulatov, D. (2012). The Role of Fundamental Frequency in Phonetic Accommodation. *Language and Speech*, 55(2), 231–248.  
<https://doi.org/10.1177/0023830911417695>
- Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glott International* 5.9/10, 341–345.  
<https://doi.org/10.1097/AUD.0b013e31821473f7>
- Chaudron, C. (1983). Simplification of Input: Topic Reinstatements and Their Effects on L2 Learners' Recognition and Recall. *TESOL Quarterly*, 17(3), 437.  
<https://doi.org/10.2307/3586257>
- Chen, A., Rietveld, T., & Gussenhoven, C. (2001). *Language-specific effects of pitch range on the perception of universal intonational meaning*. 1403–1406.  
<https://doi.org/10.21437/Eurospeech.2001-362>
- Dahl, D. A. (1981). The role of experience in speech modifications for second language learners. *Minnesota Papers in Linguistics and Philosophy of Language*, 7(2), 78–93.
- Ferguson, C. A. (1975). Toward a Characterization of English Foreigner Talk. *Anthropological Linguistics*, 17(1), 1–14.
- Gao, C., Shinkareva, S. V., & Desai, R. H. (2023). SCOPE: The South Carolina psycholinguistic metabase. *Behavior Research Methods*, 55(6), 2853–2884. <https://doi.org/10.3758/s13428-022-01934-0>
- Harper, L. R. (2019). *Simulation of Patient Caregiver Counseling in Speech Language Pathology*.

- Henzl, V. M. (1979). Foreigner Talk in the Classroom. *IRAL: International Review of Applied Linguistics in Language Teaching*, 17(2), 159–167.
- Hudson, M. E., Voytecki, K. S., Owens, T. L., & Zhang, G. (2019). Preservice teacher experiences implementing classroom management practices through mixed-reality simulations. *Rural Special Education Quarterly*, 38(2), 79–94.
- Kühnert, B., & Antolík, T. K. (2017). Patterns of articulation rate variation in English/French tandem interactions. In J. Volin & R. Skarnitzl (Eds.), *Pronunciation of English by Speakers of Other Languages*. Cambridge Scholars Publishing.
- Lewandowski, N., & Jilka, M. (2019). Phonetic Convergence, Language Talent, Personality and Attention. *Frontiers in Communication*, 4. <https://www.frontiersin.org/articles/10.3389/fcomm.2019.00018>
- Litcofsky, K. A., Tanner, D., & van Hell, J. G. (2016). Effects of language experience, use, and cognitive functioning on bilingual word production and comprehension. *International Journal of Bilingualism*, 20(6), 666–683. <https://doi.org/10.1177/1367006915579737>
- Margić, B. (2017). Communication courtesy or condescension? Linguistic accommodation of native to non-native speakers of English. *Journal of English as a Lingua Franca*, 6(1). <https://doi.org/10.1515/jelf-2017-0006>
- Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4), 2382–2393. <https://doi.org/10.1121/1.2178720>
- Piazza, G., Kalashnikova, M., & Martin, C. (2023). Phonetic accommodation in non-native directed speech supports L2 word learning and pronunciation. *Scientific Reports*, 13. <https://doi.org/10.1038/s41598-023-48648-7>
- Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America*, 123(2), 1104–1113. <https://doi.org/10.1121/1.2821762>
- Rodriguez-Cuadrado, S., Baus, C., & Costa, A. (2018). Foreigner talk through word reduction in native/non-native spoken interactions. *Bilingualism: Language and Cognition*, 21(2), 419–426. <https://doi.org/10.1017/S1366728917000402>
- Rothermich, K., Baker, R., Falkins, S. M., Kum, J., Capps, M. R., & Bobb, S. C. (2023). Reported Use of Second-Language Speech Accommodation in Everyday Interactions: The Role of Individual Differences. *Journal of Speech, Language, and Hearing Research*, 66(9), 3307–3327. [https://doi.org/10.1044/2023\\_JSLHR-23-00167](https://doi.org/10.1044/2023_JSLHR-23-00167)
- Rothermich, K., Harris, H. L., Sewell, K., & Bobb, S. C. (2019). Listener impressions of foreigner-directed speech: A systematic review. *Speech Communication*, 112, 22–29. <https://doi.org/10.1016/j.specom.2019.07.002>
- Scarborough, R. (2010). Lexical and contextual predictability: Confluent effects on the production of vowels. In *Lexical and contextual predictability: Confluent effects on the production of vowels* (pp. 557–586). De Gruyter Mouton. <https://doi.org/10.1515/9783110224917.5.557>
- Scarborough, R., Dmitrieva, O., Hall-Lew, L., Zhao, Y., & Brenier, J. (2007). An acoustic study of real and imagined foreigner-directed speech. *The Journal of the Acoustical Society of America*, 121(5), 3044–3044. <https://doi.org/10.1121/1.4781735>
- Weizheng, Z. (2019). Teacher-Student Interaction in EFL Classroom in China: Communication Accommodation Theory Perspective. *English Language Teaching*, 12(12), 99–111.
- Zuengler, J. (1991). Accommodation in native-nonnative interactions: Going beyond the “what” to the “why” in second-language research. In H. Giles, J. Coupland, & N. Coupland (Eds.), *Contexts of accommodation: Developments in applied sociolinguistics* (pp. 223–244). Cambridge University Press. <https://doi.org/10.1017/CBO9780511663673>

## 9. Language Resource References

- Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glott International* 5.9/10, 341–345. <https://doi.org/10.1097/AUD.0b013e31821473f7>
- Gao, C., Shinkareva, S. V., & Desai, R. H. (2023). SCOPE: The South Carolina psycholinguistic metabase. *Behavior Research Methods*, 55(6), 2853–2884. <https://doi.org/10.3758/s13428-022-01934-0>

# Enhancing Consumer Health Question Reformulation: Chain-of-Thought Prompting Integrating Focus, Type, and User Knowledge Level

Jooyeon Lee, Luan Huy Pham, Özlem Uzuner

Department of Information Sciences and Technology

George Mason University

{jlee252, lpham6, ouzuner}@gmu.edu

## Abstract

In this paper, we explore consumer health question (CHQ) reformulation, focusing on enhancing the quality of reformation of questions without considering interest shifts. Our study introduces the use of the website of the Genetic and Rare Diseases Information Center (GARD) at the National Institutes of Health (NIH) as a gold standard dataset for this specific task, emphasizing its relevance and applicability. Additionally, we developed other datasets consisting of related questions scraped from Google, Bing, and Yahoo. We augmented, evaluated and analyzed the various datasets, demonstrating that the reformulation task closely resembles the question entailment generation task. Our approach, which integrates the Focus and Type of consumer inquiries, represents a significant advancement in the field of question reformulation (QR). We provide a comprehensive analysis of different methodologies, offering insights into the development of more effective and user-centric AI systems for consumer health support.

**Keywords:** Consumer Health Question Answering, Chain-of-Thought Prompting, Entailment Question Generation, Question Reformulation

## 1. Introduction

Our research explores the field of consumer health question answering, a specialized area within question-answering systems that aims to provide medical knowledge to the general public. This area presents unique challenges, as it requires AI systems to communicate complex medical information in a clear and accessible manner to users who may have little to no medical training. Despite improvements in AI systems, consumers often face the burden of formulating effective queries to obtain the information they need. This process can involve a manual series of trial-and-error attempts where consumers refine their questions to learn more about their health concerns. Our study focuses on understanding how consumers reformulate their questions in the consumer health domain, drawing on insights from a study by [Chen et al.](#) that examined user behavior in Question Reformulation (QR). This study provides valuable understanding of the reasons and methods consumers use to modify their queries in search of health information. According to [Chen et al.](#) there are four primary reasons for QR: satisfaction with results, dissatisfaction leading to modification, user-initiated improvements for better alignment with search intent, and shifts to different Foci. Our research specifically addresses the first three reasons, excluding interest shifts, as a shift generally indicates a change in intent. We aim to enhance consumer satisfaction by providing suggestions that match their intended query Focus, such as maintaining the core focus while avoid-

ing shifts to new Foci. In this paper, the definition focus of a question indicates disease names as defined by [Roberts et al.](#) for the purpose of decomposing Consumer Health Questions (CHQs) and type is nondisease information of the question (e.g., symptom and treatment) ([Demner-Fushman et al., 2019](#)). Our experiments reveal a significant similarity between entailment and QR tasks, addressing a key challenge in CHQ reformulation - the lack of a standardized dataset for system evaluation. By aligning our task with entailment, we utilize multiple data sources to improve the robustness and relevance of our findings. Our main contributions are as follows: **[1.]** We propose datasets that are particularly suited for the task of CHQ reformulation. **[2.]** Our study compares three question generation methods: QR definition (ref-def) prompting, entailment definition (ent-def) prompting, and Chain-of-Thought (CoT) Prompting ([Reynolds et al., 2021](#)), to find the best way to reformulate CHQs. **[3.]** We introduce a CoT prompting technique that focuses on focus and type specifics for CHQ reformulation. This approach aims to make the reformulated questions (RQs) more relevant, accurate, and helpful.

## 2. Datasets

For the purposes of this work, we employed multiple datasets. This section provides a description of each dataset, including origin and characteristics.

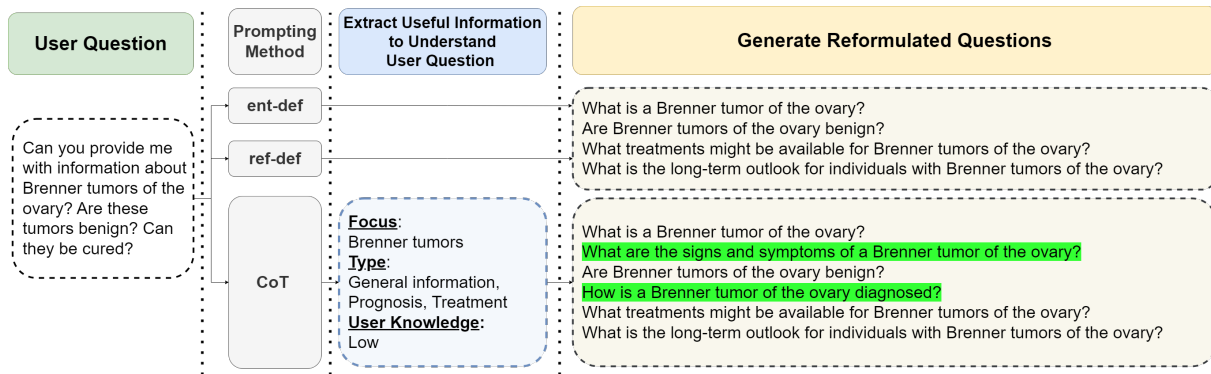


Figure 1: Prompt flow for standard prompting (ent-def and ref-def prompt) vs CoT prompting with example.

## 2.1. Entailment Question

In our study, we employed the Recognizing Question Entailment (RQE) dataset to develop algorithms for identifying similarities between CHQs and expert-answered queries (Ben Abacha and Demner-Fushman, 2016). This dataset includes real-world CHQs from the National Library of Medicine and FAQs from various NIH websites, encompassing a wide range of medical Foci. The goal of the RQE task is to determine if an existing FAQ answer can also respond to a new CHQ. We select the 129 true entailment pairs to test the QR models.

## 2.2. GARD Reformulated Questions

We created a dataset using content scraped from the website of the Genetic and Rare Diseases Information Center (GARD), a part of the US National Institutes of Health (NIH). GARD provides public information regarding rare diseases, including causes, treatments, and submitted CHQs paired with expert-provided responses (GARD). For instance, when we go to GARD website, section cysticercosis, the page has a "see answer" link under "GARD Answer", which redirects to a different page, that lists the CHQs, their answers with references, and the expert suggested questions are introduced with the phrase "The following information may help to address your question:". It is important to note that direct access to these expert suggestions is not available in the current version of the GARD website. However, an earlier version of the site, preserving these valuable expert insights, is accessible through Internet-in-a-Box, an innovative solution designed to provide offline access to various educational content, including medical resources (Internet-in-a-Box Team, 2016). This digital resource, available at [https://iiab.me/modules/en-nih\\_rarediseases/diseases/categories/index.html](https://iiab.me/modules/en-nih_rarediseases/diseases/categories/index.html), offers an archived version of the GARD site.

We scraped these questions upto 250 CHQ-RQ

pairs. Of the pairs CHQs in the dataset, 48 were modified to explicitly state the topic, which could be inferred from the section title on the GARD website but was not directly mentioned in the CHQ. For example, "this condition..." is changed to: "Myostatin-related muscle hypertrophy condition...".

## 2.3. Search Engine API

We developed QR datasets using the RQE and GARD CHQs using SerpApi. SerpApi is a commercially-available web engine scraping service which has People Also Ask/Related Questions APIs for Google, Bing, and Yahoo (jvmvik et al., 2024). According to the web traffic analysis website StatCounter, these three search engines comprise over 96% of the global search market in February 2024 (Chen et al., 2021). SerpApi furnishes code libraries in various programming languages. We used version 2.4.1, which allowed text queries to be submitted (specifying the search engine as a parameter) and the RQs be returned in JSON format. Our inputs were the GARD and RQE datasets.

## 3. Experiment Setup

Our experimental framework is structured around GPT-4 (OpenAI et al., 2024), a model recognized for its advanced language processing and generation proficiency. The methodology employed in this study was based on a one-shot prompting technique, which was consistently applied across all experiments.

### 3.1. Prompts

Our study involved a comparative analysis of three distinct prompts to assess their effectiveness in the reformulation of CHQs: 1) ref-def prompt, 2) ent-def prompt, and 3) Chain-of-Thought (CoT) Prompting. Actual prompts are shown in Table 1.

The ref-def prompt was designed for QR task and is aimed to rephrase a given CHQ into several

Method	Actual Prompt
ref-def Prompt	The goal is to reformulate the given consumer health question into several, clearer single-sentence questions that could potentially answer the original, given consumer health question to increase satisfaction of the consumer."
ent-def Prompt	The goal is to reformulate the given consumer health question into several clearer single-sentence questions which are in an entailment relationship to the original given consumer health question. The definition of an entailment relationship is 'when question A can answer question B partially or fully, then question B entails question A.'
CoT Prompt	<p>The goal is to reformulate the given consumer health question into several, clearer single-sentence questions that could potentially answer the original, given consumer health question to increase satisfaction of the consumer.</p> <p>(1) Identify the Main Health Topic (Focus): Determine the primary health condition in the user's question to establish the central subject of inquiry.</p> <p>(2) Assess User Knowledge Level (Knowledge): Evaluate the consumer's familiarity with the health topic based on the language and concepts used in their question. This assessment categorizes knowledge as LOW, MEDIUM, or HIGH. LOW Knowledge Level: Indicates a basic or minimal understanding of the health condition or topic. The user may be unfamiliar with the condition or its implications. This level typically includes general inquiries or seeks foundational information. Example questions might be: "What is [health condition]?" "What causes [health condition]?" "Are there common symptoms associated with [health condition]?" MEDIUM Knowledge Level: Suggests a moderate understanding of the condition. The user might know what the condition is or some of its symptoms, but seeks more detailed or specific information. This level often involves questions about management, treatment options, or lifestyle impacts. Example questions might be: "What are the treatment options for [health condition]?" "How does [health condition] typically progress over time?" "Can lifestyle changes impact the course of [health condition]?" HIGH Knowledge Level: Indicates an advanced understanding or familiarity with the health condition. Users at this level often have detailed knowledge about the condition and seek highly specific, nuanced, or recent information. This might include questions about the latest research, complex treatment options, or specific subtypes of the condition. Example questions might be: "What are the latest research findings on [health condition]?" "Are there new or experimental treatments for [health condition]?" "How does [health condition] interact with other coexisting conditions?"</p> <p>(3) Determine Information Needs (Type): Identify what specific aspects of the condition the user is interested in, such as symptoms, causes, treatments, prognosis, or lifestyle impacts.</p> <p>(4) Question Segmentation: Segment the consumer question into individual, focused questions using the determined Focus, Type and User Knowledge Level. Each question should address a single aspect of the Focus.</p>

Table 1: This table shows actual prompts used for the experiments.

clearer, single-sentence questions that could potentially provide answers to the original CHQ, thereby increasing consumer satisfaction.

The ent-def prompt was based on the question entailment definition as defined by [Ben Abacha and Demner-Fushman](#), particularly suited for question-answering tasks: "When question A can answer question B partially or fully, then question B entails question A."

For the CoT Prompting, we expanded on our prior research that underscored the significance of identifying the Focus and Type in the question entailment recognition task for CHQs ([Lee and Pham, 2022](#)). This approach integrates the CoT method, introduced by [Reynolds et al.](#), which enhances performance by incorporating reasoning steps. Our prompt has three chains: 1) extract the Focus; 2) determine the Type; and 3) evaluating the user's knowledge level about the Focus. 4) considering 1, 2, 3, reformulate the CHQ.

### 3.2. Evaluation Method

Our approach for evaluating the accuracy of generated questions involves text similarity measurement analysis. We aligned the output generated by our models with a predefined gold standard of questions. We consider the output of models as a single output by concatenating the list of questions and doing the same for the gold standard (GARD expert suggestion). This enables a direct comparison to obtain a similarity score. For the similarity metric, we utilized UniEval, a multifaceted tool designed for evaluating text generation tasks. It measures aspects like consistency, coherence, and relevance ([Zhong et al., 2022](#)). We also applied the ROUGE (R) metric, with an emphasis on R-1 for unigram overlap and R-L for the longest common subsequence ([Lin, 2004](#)), to analyze the lexical similarity and coherence of the generated research questions. This methodological blend offered a robust framework for assessing the effectiveness of our models.

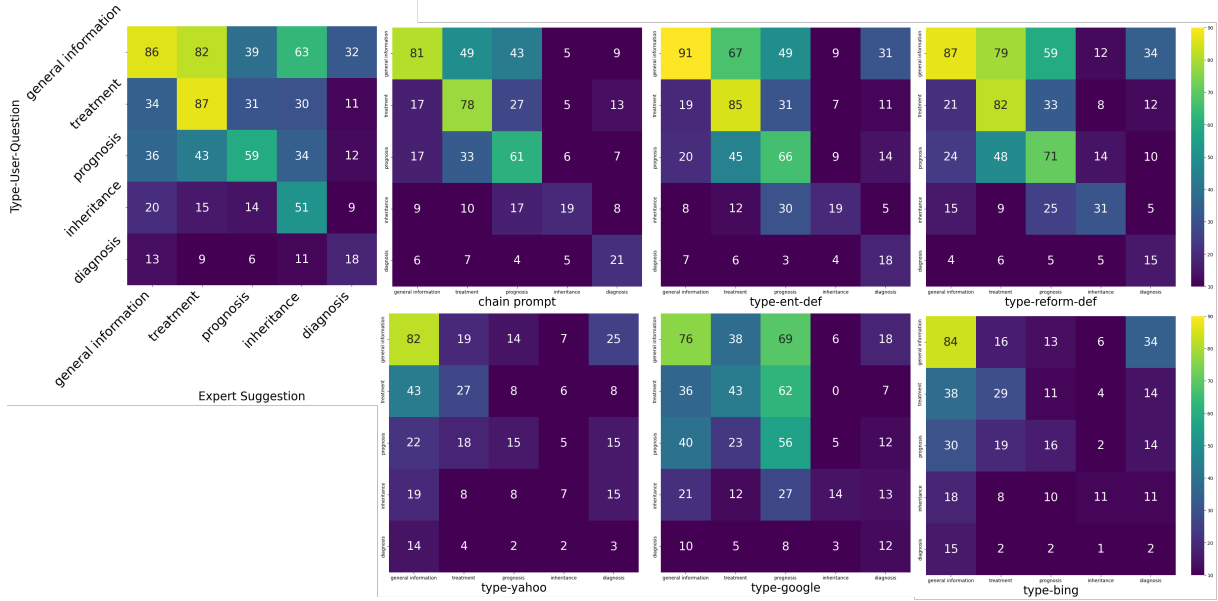


Figure 2: Heatmap of GARD dataset vs generated output (5 most frequent types only).

Data	Method	Consistency	Coherence	Relevance	R-1	R-L
GARD	CoT prompt	<b>0.9338</b>	<b>0.9604</b>	0.8945	<b>0.3115</b>	<b>0.2630</b>
	ref-def prompt	0.9003	0.9472	<b>0.8975</b>	0.2964	0.2344
	ent-def prompt	0.9087	0.9456	0.8833	0.2892	0.2300
	Bing	0.7978	0.8078	0.7475	0.1487	0.1231
	Google	0.7704	0.7262	0.6383	0.2101	0.1711
	Yahoo	0.7979	0.8037	0.7442	0.1572	0.1318
GARD**	CoT prompt	<b>0.9339</b>	<b>0.9620</b>	<b>0.9414</b>	<b>0.4269</b>	<b>0.4241</b>
	ref-def prompt	0.9068	0.9530	0.9345	0.4128	0.4061
	ent-def prompt	0.9151	0.9525	0.6512	0.1946	0.1918
	Bing	0.8193	0.8728	0.8690	0.3924	0.3910
	Google	0.7927	0.8325	0.7941	0.3523	0.3510
	Yahoo	0.8151	0.8784	0.8676	0.3634	0.3614
RQE	CoT prompt	0.7436	0.8058	0.7690	<b>0.1208</b>	<b>0.1152</b>
	ref-def prompt	0.7411	<b>0.8798</b>	<b>0.8506</b>	0.1097	0.1030
	ent-def prompt	0.7424	0.8532	0.8226	0.1118	0.1055
	Bing	0.7690	0.7937	0.7620	0.0743	0.0662
	Google	0.7060	0.7290	0.6934	0.0868	0.0819
	Yahoo	<b>0.7868</b>	0.8170	0.7864	0.0742	0.0698

Table 2: Evaluation using with UniEval (consistency, coherence and relevance), R-1 and R-L. GARD\*\* designates results on the GARD dataset where output and user questions possess identical Focus (to have a "fairer" comparison with search engine suggestions by excluding "intention shift").

Furthermore, we incorporated quantitative metrics for a comprehensive assessment. We calculate the frequency of Focus and Type elements in the responses, thereby measuring their alignment with the gold standard, which is shown in Table 3. We use the Euclidean Distance to measure the similarity between gold standard vs. prompt outputs and gold standard vs. search engine. The formula is  $Distance_i = \sqrt{(F_0 - F_i)^2 + (T_0 - T_i)^2}$  where,  $F_0$  and  $T_0$  are representing an average number of Focus and an average number of Type values, respectively, for the gold standard.  $F_i$  and  $T_i$  are the corresponding values for the row being compared.

## 4. Analysis

In this section, we provide an analysis of results.

### 4.1. Question Reformulation Analysis

The performance of prompts on the GARD dataset using ref-def, ent-def, and CoT prompting (which incorporates Focus and Type along with reformulation definition) was remarkably similar when tested against an expert-suggested dataset. This similarity indicates a strong alignment between the task and the dataset. Notably, CoT prompting demonstrated superior performance over other methods

	Focus	Type	Distance
GARD User Question	1.6447	2.0220	2.897
Gold standard	1.5311	4.8462	-
CoT Prompting	2.2308	4.3187	<b>0.707</b>
Ref Def Prompting	2.2418	<b>5.2601</b>	0.715
Ent Def Prompting	<b>1.7289</b>	3.2015	1.648
Bing API	2.3736	2.8425	2.034
Google API	2.7509	3.3663	1.596
Yahoo API	2.3040	2.8352	2.028

Table 3: Comparative analysis of prompting methods against the gold standard, using an average occurrence of Focus and Type. Method alignment with the gold standard calculated using the Euclidean distance (lower values = greater similarity).

in terms of consistency, coherence, and R-1 and R-L scores. This suggests that understanding the Focus and Type of a question before generating a reformulated version is crucial for this specific task.

## 4.2. Entailment Task Analysis

In our analysis of entailment prompts, we noted that the ref-def prompt and the ent-def prompt demonstrated remarkably similar patterns. As shown in Table 2, the performance scores of both prompts were closely aligned when applied to the GARD dataset and RQE dataset, across multiple evaluation methods. This parallel trend is also evident in Figure 2, which further corroborates our observation. These results align with the patterns of human behavior observed in previous surveys, as cited in (Chen et al., 2021). For future research, especially in scenarios where there is no “interest shift” in user intent, augmenting the reformulation task with question entailment datasets emerges as a promising strategy to overcome data limitations.

Regarding the entailment dataset, we conducted the same experiment using the RQE dataset to assess its similarity to the QR task. Considering that the RQE dataset consists of only 129 entailment question pairs and was originally designed for entailment recognition tasks, the reformulation pairs are limited to user questions with a single corresponding entailment question. Therefore, the evaluation might not fully represent real-world scenarios. However, similar to the findings with the GARD dataset, the CoT Prompt method outperformed others on the ROUGE metrics.

## 4.3. Search Engine Behavior Analysis

Our study also delved into the behavior of search engines in offering QR suggestions. Utilizing the GARD dataset, Google provided QR suggestions for 99.6% of queries, significantly outperforming Yahoo and Bing, with both providing only 81.2% each.

Interestingly, the Google’s heatmap in Figure 2, illustrates a more evenly distributed range of Foci compared to prompt results and expert suggestions. Also indicated in Table 3, the average number of Foci in Google’s suggestions exceeds those in other methods. This suggests that Google’s QR approach provides more varied results.

Given this indication that a large portion of questions suggested by search engine consist of interest shift suggestions, we explored excluding the ‘intent shift’ questions from the search engines, by selecting questions from the GARD dataset where both input and output were determined to have the same Focus. After filtering, only 11%, 9%, and 8% of questions remained in the Google, Bing, and Yahoo datasets, respectively - confirming our suspicions. In comparison, the prompting methods had higher Foci alignment as the GARD expert suggested RQs: 33% (ent-def prompt), 35% (ref-def prompt), and 34% (CoT prompt). We then ran our evaluation method upon this filtered dataset with results shown in Table 2, row with GARD\*\*. Compared to the full GARD dataset, we see that overall search engine accuracy significantly increased, yet still lower than our three prompting results. We conclude that filtering on Focus would not be a desirable dataset augmentation and that further investigation is required on how to separate interest shift task vs. QR with search engine dataset.

Also, while this diversity is advantageous for users seeking to shift the Focus of their inquiries, it may be undesirable for those who simply wish to rephrase their existing questions. In contrast, a mechanism to filter or categorize these options may be preferred. Such a system would enhance the user experience by streamlining the process of navigating through the multitude of suggestions, thereby catering more effectively to the specific needs of users, whether they seek focus diversity or question refinement.

## 5. Conclusion

In our study, we have made significant strides in the field of CHQ reformulation by conducting a comprehensive comparative analysis of three distinct question generation methodologies: ref-def, ent-def and CoT prompt. Our CoT prompting approach, which integrates Focus and Type specificity, represents a novel method tailored for CHQ reformulation. Furthermore, we have identified and recommended specific datasets that are instrumental for ongoing research in this domain. These datasets are poised to aid other researchers in conducting similar studies, thereby driving continuous innovation and exploration in the field. Our contributions lay the groundwork for future advancements in CHQ reformulation, setting a new benchmark for research.

## 6. Bibliographical References

- Asma Ben Abacha and Dina Demner-Fushman. 2016. [Recognizing question entailment for medical question answering](#). In *AMIA 2016, American Medical Informatics Association Annual Symposium, Chicago, IL, USA, November 12-16, 2016*.
- Jia Chen, Jiaxin Mao, Yiqun Liu, Fan Zhang, Min Zhang, and Shaoping Ma. 2021. [Towards a better understanding of query reformulation behavior in web search](#). In *Proceedings of the Web Conference 2021, WWW '21*, page 743–755, New York, NY, USA. Association for Computing Machinery.
- Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2019. [Consumer health information and question answering: helping consumers find answers to their health-related information needs](#). *Journal of the American Medical Informatics Association : JAMIA*.
- GARD. About GARD | genetic and rare diseases information center (GARD) – an NCATS program. .
- Internet-in-a-Box Team. 2016. Internet-in-a-box (iiab). <https://github.com/iiab/iiab>.
- jvmvik, hartator, ilyazub, Dmitry Zub, Lóric Pap, Kenneth Reitz, Lenny Fishler, Alexej, gbcfxs, Justin O'Hara, Manoj Nathwani, ajsierra117, and elizost. 2024. [serpapi/google-search-results-python](#).
- Jooyeon Lee and Luan Huy Pham. 2022. [Recognizing question entailment in consumer health using a query formulation approach](#). In *Proceedings of The First Workshop on Context-aware NLP in eHealth (WNLPe-Health 2022) co-located with The nineteenth International Conference on Natural Language Processing (ICON-2022), Delhi, India, December 15-18, 2022*, volume 3416 of *CEUR Workshop Proceedings*, pages 56–69. CEUR-WS.org.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madeleine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rameesh Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly



Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sasstry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lillian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. [Gpt-4 technical report](#).

Laria Reynolds, Jason Wei, Daphne Ippolito, Noah Fiedel, Emily Reif, Andy Coenen, Ann Yuan, Adam Roberts, and Colin Raffel. 2021. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Kirk Roberts, Halil Kilicoglu, Marcelo Fiszman, and Dina Demner-Fushman. 2014. [Decomposing consumer health questions](#). In *Proceedings of BioNLP 2014*, pages 29–37, Baltimore, Maryland. Association for Computational Linguistics.

Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, and William Yang Wang. 2023a. [SESCORE2: Learning text generation evaluation via synthesizing realistic mistakes](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5166–5183, Toronto, Canada. Association for Computational Linguistics.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023b. [INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. [Towards a unified multi-dimensional evaluator for text generation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

## A. Appendix

### A.1. Data Source Links

This section specifies the precise location of the data source.

#### A.1.1. GARD Questions

The dataset was obtained from the following website: <https://rarediseases.info.nih.gov/search?keyword=the%20following%20information%20may%20help%20address%20your%20question&page=1&filters=contentType%3Dgardcase>

However, access to the expert suggestions is no longer available on the current version of the site. Nevertheless, an older version of the website, which includes these expert suggestions, can be accessed via this link: [https://iiab.me/modules/en-nih\\_rarediseases/diseases/categories/index.html](https://iiab.me/modules/en-nih_rarediseases/diseases/categories/index.html). In this version, the expert responses can be found in the "GARD Answer" section.

For instance, to find information on cysticercosis, one can visit [https://iiab.me/modules/en-nih\\_rarediseases/diseases/8194/cysticercosis/index.html](https://iiab.me/modules/en-nih_rarediseases/diseases/8194/cysticercosis/index.html). In this section, clicking on the "see answer" link under "GARD Answer" redirects to a different page ([https://iiab.me/modules/en-nih\\_rarediseases/diseases/8194/cysticercosis/cases/26056/index.html](https://iiab.me/modules/en-nih_rarediseases/diseases/8194/cysticercosis/cases/26056/index.html)). This page lists the CHQ, suggested questions, their answers with references, and the expert suggested questions are introduced with the phrase "The following information may help to address your question:".

#### A.1.2. RQE Dataset

The RQE datasets, as referenced in [Ben Abacha and Demner-Fushman \(2016\)](#), are made publicly accessible at the following URL: [https://github.com/abachaa/RQE\\_Data\\_AMIA2016](https://github.com/abachaa/RQE_Data_AMIA2016)

#### A.1.3. Search Engine API

The dataset for the search engine was not obtained from a repository; instead, we directly scraped search results using SerpAPI, as detailed in

(jvmvik et al., 2024). The relevant information and resources for this API can be found at the following URL: <https://github.com/serpapi/google-search-results-python?tab=readme-ov-file>.

## **A.2. Examples**

In this section, we provide samples of each dataset and prompt results, which are listed in Table 4, Table 5, Table 6.

CHQ	How is fibrosing mediastinitis treated? How can I learn more about research?
Focus	['Fibrosing Mediastinitis']
Type	['Treatment', 'Research']
User Knowledge Level	MEDIUM
CoT Prompt Result	['What is the treatment for Fibrosing Mediastinitis?', 'Where can I find recent research on Fibrosing Mediastinitis?']
GARD Expert Suggestion	How might fibrosing mediastinitis be treated? Are there any research studies or clinical trials enrolling people with fibrosing mediastinitis?

Table 4: CoT Prompt Good Result.

CHQ	My son presented with a sudden onset of partial complex seizures and a friend provided us with information about neurocysticercosis. My son has not yet been tested, but I would like to know if this disease has an affect on a person's appetite or weight?
Focus	['Neurocysticercosis']
Type	['Symptoms', 'Effects']
User Knowledge Level	MEDIUM
CoT Prompt Result	['What is neurocysticercosis?', 'What are the symptoms of neurocysticercosis?', "Does neurocysticercosis affect a person's appetite?", "Does neurocysticercosis affect a person's weight?"]
GARD Expert Suggestion	What is cysticercosis? What is neurocysticercosis? Can neurocysticercosis affect a person's appetite or weight? With whom should I speak if I suspect neurocysticercosis may be affecting someone I know?

Table 5: CoT Prompt Error Result. Expert suggested a related question about cysticercosis, but this Focus is not shown in CHQ nor prompt output. Cysticercosis and neurocysticercosis are related but distinct conditions, both caused by the same parasitic infection but affecting different parts of the body. Thus, it is useful information to users who do not have a HIGH knowledge level about the disease. This is a question the user may ask without an "interest shift" intention.

CHQ	I have a Robertsonian translocation 13;14. My husband and I just had our trisomy 13 baby boy pass at 24 weeks gestation. My question is, does the type of trisomy 13 a baby has correlate with how long it will live in utero? For example, does full trisomy 13 have a better chance of survival than translocation trisomy 13? Or do they all have an equal prognosis?
Focus	['Robertsonian translocation 13;14', 'trisomy 13', 'survival in utero', 'prognosis']
Type	['cause', 'prognosis']
User Knowledge Level	MEDIUM
CoT Prompt Result	['Does the type of trisomy 13 a baby has correlate with how long it will live in utero?', 'Does full trisomy 13 have a better chance of survival than translocation trisomy 13?', 'Do all types of trisomy 13 have an equal prognosis?']
GARD Expert Suggestion	Does the prognosis differ depending on the type of trisomy 13 a fetus has? How can I find a genetics professional in my area?

Table 6: CoT Prompt Bad Result. Prognosis belongs to Type, not Focus. Survival in utero is also not a disease name. This is a case of multiple and incorrect Foci, which led to incorrect final results.

# Exploring the Challenges of Behaviour Change Language Classification: A Study on Semi-Supervised Learning and the Impact of Pseudo-Labelled Data

Selina Meyer\*, Marcos Fernández-Pichel†, David Elswailer\*, David E. Losada†

\*Regensburg University

Universitätsstraße 31, 93053 Regensburg

†Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS),

Universidade de Santiago de Compostela

Rúa Jenaro de la Fuente s/n, 15782 Santiago de Compostela

{selina.meyer, david.elswailer}@ur.de

{marcosfernandez.pichel, david.losada}@usc.es

## Abstract

Automatic classification of behaviour change language can enhance conversational agents' capabilities to adjust their behaviour based on users' current situations and to encourage individuals to make positive changes. However, the lack of annotated language data of change-seekers hampers the performance of existing classifiers. In this study, we investigate the use of semi-supervised learning (SSL) to classify highly imbalanced texts around behaviour change. We assess the impact of including pseudo-labelled data from various sources and examine the balance between the amount of added pseudo-labelled data and the strictness of the inclusion criteria. Our findings indicate that while adding pseudo-labelled samples to the training data has limited classification impact, it does not significantly reduce performance regardless of the source of these new samples. This reinforces previous findings on the feasibility of applying classifiers trained on behaviour change language to diverse contexts.

**Keywords:** Behaviour Change, Semi-Supervised Learning, Low Resource Application Areas

## 1. Introduction

The way people talk about change can be an indicator for future success of their attempts to alter their behaviour (Magill et al., 2018; Moyers et al., 2007). Different types of language indicate varying levels of intent to change (Resnicow et al., 2012) and being able to automatically differentiate these language types could improve conversational agents (CAs) with the purpose of assisting behaviour change. For example, a CA could adapt its behaviour to a user's current situation and motivational level and elicit more favourable utterances in order to increase the user's resolve to change. Additionally, obtaining such information from patient texts, such as journals, could serve as a meaningful resource for practitioners, helping them gain deeper understandings of the patient's current situation (Kim et al., 2023)

Current CAs fail to use this information (Xu and Zhuang, 2022), not least because of a lack of annotated text around behaviour change. The availability of new datasets with labelled utterances would facilitate the construction of supervised learning solutions (Meyer and Elswailer, 2022). However, such data is difficult to obtain for two main reasons. First, it is commonly sourced from transcripts of therapy sessions or counsellor training materials, leading to privacy and data security concerns that complicate the publication of datasets. This natu-

rally limits the size of the data. Second, the costs of annotation and the necessary training of the assessors hinder the creation of large datasets with fine-grained annotations (Pérez-Rosas et al., 2016; Wu et al., 2022).

Additionally, behaviour change language exhibits certain peculiarities. It can be applied to different kinds of unrelated *target behaviours*, e.g. increasing physical activity or smoking cessation, which demands the ability of a classification algorithm to be able to transfer between contexts. Talking about behaviour change also naturally leads to highly imbalanced data. Certain types of utterances, such as reasons for or against change, tend to appear often while others, such as statements about specific commitments for the future, are less frequent (Lord et al., 2015).

Semi-supervised learning (SSL) has commonly been used to alleviate the limitations imposed on classification performance by a scarcity of training data. SSL has shown to be particularly successful in popular benchmarks (Van Engelen and Hoos, 2020; Duarte and Berton, 2023). In this paper, we explore the feasibility of using SSL in the context of classifying highly imbalanced text about behaviour change. We explore the inclusion of pseudo-labelled data, both from the original source and from new sources covering different behaviour change contexts. We evaluate to

what extent including new pseudo-labelled data from different sources impacts the classifier’s ability to correctly predict the utterance class of short texts. Furthermore, we work with several out-of-context test datasets and explore the trade-off between the amount of pseudo-labelled data added and the strictness of the inclusion criteria for the pseudo-labelled samples<sup>1</sup>.

## 2. Background & Related Work

### 2.1. SSL for Text Classification

In their survey paper, [Van Engelen and Hoos \(2020\)](#) give an extensive overview of different SSL models and common application areas. A more recent review, by [Duarte and Berton \(2023\)](#), focuses specifically on the application of SSL methods to text classification. According to them, one of the most commonly explored types of SSL for text classification is self-learning, where a baseline classifier is used to assign pseudo-labels to new, unlabelled data. These pseudo-labelled examples are then included in the training data of the classifier. This classic approach is simple and has the advantage of being suited to be used in combination with any base learner ([Van Engelen and Hoos, 2020](#)).

Past work has shown the effectiveness of SSL in various domains, including health and well-being. For instance, [Varma and Ré \(2018\)](#) presented a tool for automatically generating weak supervision rules for data labelling. The authors demonstrated the effectiveness of this method in spam classification and medical diagnosis. In the same vein, [Ratner et al. \(2020\)](#) presented a tool to streamline the process of creating training data with weak supervision techniques. The usefulness of this tool, which allowed users to rapidly define labelling functions, was demonstrated in real-life applications such as medical information extraction and knowledge base construction.

Other studies have focused on mitigating weaknesses frequently associated with SSL techniques. For instance, there is often an inherent proneness to class imbalance, which is observable even when the baseline classifier is trained on balanced data ([Wang et al., 2022](#)). Real-world data is rarely balanced. [Guo and Li \(2022\)](#) addressed this problem by introducing a framework that supports adaptive thresholding for different classes. Their approach is effective without prior knowledge of a dataset’s class distribution.

SSL has been frequently applied to publicly available and widely researched benchmarks. These experiments often yielded solid results ([Van Engelen and Hoos, 2020](#); [Duarte and Berton, 2023](#)). However, recent studies have argued that

performance on these datasets does not always equal reliability and robustness in real-world applications ([Kiela et al., 2021](#); [Schlegel et al., 2022](#); [Church and Kordoni, 2022](#)). It is hard to predict to which extent SSL is beneficial for a given situation ([Van Engelen and Hoos, 2020](#)), with many studies even reporting decreases in classification performance ([Oliver et al., 2018](#); [Li and Zhou, 2014](#)). Because of this potential for deterioration, we chose to first evaluate the effect of SSL for behaviour change language using self-training, and leave the exploration of other, more sophisticated SSL methods to future work.

### 2.2. Behaviour Change Language

One way to formalise talk about behaviour change is the Motivational Interviewing Skill Code (*MISC*) ([Miller et al., 2003](#)). It helps to categorise utterances into different valences and topics around behaviour change across multiple target behaviours. While Motivational Interviewing (MI) was initially developed for addiction counselling, it has since been used for various topics, ranging from smoking cessation, over nutrition and fitness, to work-related behaviour ([Miller and Rollnick, 2002](#); [Clifford and Curtis, 2016](#); [Page and Tchernitskaia, 2014](#); [Güntner et al., 2019](#)).

The *MISC* defines different categories for utterances, which we outline in Table 1. Based on the *MISC*, each user utterance that is not Follow/Neutral is assigned a valence and a topic. If the topic is *Reason*, the utterance is also assigned a reason type. This annotation framework can help to infer a person’s intensity of commitment to behaviour change ([Resnicow et al., 2012](#)). For example, the *MISC* helps to understand how confident people feel about change, what type of rationale leads them to pursue change and whether they have already become active or are planning to do so in the near future.

Past research on classifying these behaviour codes has largely focused on the distinction between Change Talk, Follow/Neutral and Sustain Talk, and the few existing public MI-datasets do not contain topic and reason type annotations ([Wu et al., 2022](#); [Pérez-Rosas et al., 2016](#)). This lack of fine-grained annotations hinders the development of more sophisticated classifiers that take into account the topic of user utterances and the types of reasons they voice for making a change.

An exception to this is the GLoHBCD, a German dataset that contains written forum data annotated with valences, topics, and reason types based on the *MISC* ([Meyer and Elsweiler, 2022](#)). The creators of the GLoHBCD demonstrated the feasibility of training transformer-based classifiers on the data, reaching macro F1 scores between 70% and 77% depending on the label-level. How-

---

<sup>1</sup>We make our code available on [GitHub](#).

Level	Label	Description
Valence	Change Talk (+)	Utterances in favour of behaviour change
	Sustain Talk (-)	Utterances in favour of status quo
Topic	Reason	Reasons for/against change
	Taking Steps	Specific steps taken in the recent past
	Commitment	Agreement, intention, or obligation for the near future
Reason Type	Ability	Ability and degree of difficulty of the change
	Need	Necessity of change, or maintaining the status quo
	Desire	Desire for change, or current behaviour
	General	General justifications, incentives, or justifications
Follow/Neutral (FN)		Utterances not related to behaviour change

Table 1: Description of utterance classifications, based on (Miller et al., 2003; Meyer and Elswailer, 2022)

ever, these experiments also showed that some label-levels are harder to classify and that the imbalanced nature of the data can be problematic. In further experiments, the same team showed that the classification of these utterances transfers to a certain extent between different target behaviours and conversational contexts (Meyer and Elswailer, 2023).

With macro F1 scores ranging mostly between 60% and 90%, the classification results reached on out-of-context datasets suggest a certain degree of stability, but still leave much room for improvement. The GLoHBCD consists of only 4724 data points relevant to behaviour change, and the less represented classes include less than 200 samples, which makes it likely that introducing more data would lead to improved classification.

### 3. Datasets

In this paper, we intend to build on the results presented by Meyer and Elswailer (2022, 2023) and determine the feasibility of applying SSL approaches to the GLoHBCD. We aim at increasing classification performance on the original dataset and, additionally, employing the classifiers on external chat-like conversational data about different target behaviours. To explore this, we have collected new data from the same source as the GLoHBCD, as well as from other sources. In this section, we first outline the main properties of the GLoHBCD (§3.1) and then give an overview of the data sources used for pseudo-labelling (§3.2) and the test sets used to evaluate transfer learning capabilities (§3.3).

#### 3.1. GLoHBCD

The GLoHBCD is a dataset of forum posts, written by people trying to lose weight, which was annotated with labels based on the *MISC* (Meyer and Elswailer, 2022). The data was collected in August 2020 and written between May 2006 and July

2020. It stems from two subforums of a large-scale German weight loss forum, which were initially pre-screened for utterances around motivation for weight loss, after which relevant posts were annotated on a sentence-level basis. Each data point consists of a single sentence from the forum, together with a valence, a topic, and, if the topic is reason, a reason type annotation, as defined in Table 1.

#### 3.2. Data Used for Pseudo-labelling

We used three different datasets as sources for pseudo-labelled text, one of them stemming from the same source as the GLoHBCD, another coming from a different source with the same conversational context, and a third being sourced from spoken interactions about a variety of target behaviours. This allowed us to explore to what extent adding new data from different contexts, which likely introduces more linguistic variety, can be used to improve classification of new data.

**Weight Loss Forum Data** For the Weight Loss Forum Data, we collected new posts from the same source as the GLoHBCD. We collected all posts published after the extraction date of the initial GLoHBCD data (August 2020). There was no manual pre-filtering of this new data, which consists of 992 sentences and serves as in-domain data for pseudo-labelling.

**Smoking Cessation Forum Data** The Smoking Cessation Forum Data consists of data that is similar to the GLoHBCD, in the sense that it also consists of forum data. However, this dataset consists of reports of people attempting to quit smoking. As such, it represents data from the same type of source, but from a different context as the original dataset. The dataset was created by Meyer and Elswailer (2023) and includes ground truth *MISC* annotations for each of the 662 sentences in the dataset. We can use these annotations to evaluate the effect of adding pseudo-labelled samples

Dataset	Domain/Target behaviour	Context	# sentences
Health Coaching Dialogue Corpus <sup>2</sup>	step count increase	Text conversations with health coach	508
Optifast Mock-Chatbot	weight loss	Text conversation with simulated motivational chatbot	90
DARN-CT-based Wizard of Oz Dialogues	New Year’s resolutions	Text conversations with simulated motivational chatbot	80
Synthetic GPT-3 Data <sup>3</sup>	weight loss	User simulation through eliciting questions	74
GLoHBCD (test split)	weight loss	Forum - Interaction between peers	924
Smoking Cessation Forum (test split)	smoking cessation	Forum - Interaction between peers	199

Table 2: Overview of test datasets introduced in Meyer and Elswailer (2023) to evaluate domain transfer capabilities of classifiers (table adapted from Meyer and Elswailer (2023))

to the training data on the classification of data from different sources as the original training data. To explore this, we use 10% of this data collection as a test set to evaluate the performance of the final model. The remaining sentences are used as a source for pseudo-labelled data.

**AnnoMI** The AnnoMI dataset is a collection of transcribed MI sessions across a variety of behaviour change contexts (Wu et al., 2022). While still being language data related to behaviour change uttered by humans, this dataset differs both in context (topics range from weight loss and smoking cessation, across alcohol abuse to other issues) and source type, as the data is transcribed from spoken counselling sessions, whereas the GLoHBCD consists of peer-to-peer conversations in a written forum. As such, this dataset is the furthest away from the original dataset and could thus offer the largest increase in linguistic variance. Since the dataset only includes valence annotations of client utterances, we use all client utterances which are not annotated as Follow/Neutral as data to pseudo-label for our experiments. Since for the remaining datasets used in this study each sentence constitutes a single data point, we separate the utterances in the AnnoMI into sentences following the same approach as for the other datasets, resulting in 2481 sentences.

### 3.3. Data used for Testing the SSL Classifier

Finally, we use multiple test sets to evaluate the ability of the SSL classifier to predict the type of behaviour change utterance. This includes a broad range of collections, ranging from written chat-like conversations to forum and spoken interactions, assembled by Meyer and Elswailer (2023) to evaluate transfer learning capabilities of classifiers trained on GLoHBCD data. In this way, we

can evaluate the transfer learning capabilities of the SSL classifier. This is intended to give insights about the effects of adding pseudo-labelled data from the original source (and from other sources) on the ability of the classifier to recognise utterances under varying conditions. Introducing test data from such a broad variety of contexts tells us to which extent adding pseudo-labelled data from multiple sources benefits or hinders classification of new data with varying degrees of closeness to the GLoHBCD.

In Table 2 we give an overview of the datasets used for testing, their conversational context, and behaviour change domain. Following Meyer and Elswailer (2023), we included synthetically generated chat data, which can be seen as stereotypical utterances about change. This acts as a sanity check, since a decrease in performance on this dataset after adding pseudo-labelled data would indicate a significant increase in noise. We also create an 80%-20% split of the GLoHBCD, using the 20% as a final test set, whereas the remaining 20% are used for training the baseline classifiers.

## 4. Experimental Setup

We ran experiments across four stages, which we will outline in this section. The first three stages are made up of fine-tuning experiments, whereas the fourth stage applies the findings to the test sets. For fine-tuning, we followed the following methodology: In 10-fold cross-validation, i) a BERT-based classifier is fine-tuned on the GLoHBCD training data (baseline classifier), ii) new data is pseudo-labelled, iii) GLoHBCD training data and pseudo-

<sup>1</sup>data based on Gupta et al. (2020) with annotations by Meyer and Elswailer (2023)

<sup>2</sup>based on Meyer et al. (2022) with annotations by Meyer and Elswailer (2023)

Valence		Topic		Reason Type	
System	Macro F1	System	Macro F1	System	Macro F1
baseline	72.65 (1.96)	baseline	74.05 (3.17)	baseline	75.63 (2.93)
NP, CT(0.5)	73.98 (2.29)	NP, CT(0.95)	75.89 (3.82)	P(0.7), CT(0.5), min	76.7 (2.96)
NP, CT(0.5), equal	73.73 (1.85)	P(0.5), CT(0.95)	75.46 (3.47)	P(0.7), CT(0.5)	76.28 (4.33)
NP, CT(0.1)	73.55 (2.36)	P(0.7), CT(0.95)	75.35 (3.15)	P(0.7), CT(0.4)	76.25 (3.35)

Table 3: Comparison of classification setups on gLoHBCD cross-validation splits with baseline (no SSL). Variants include Pre-filtering (P(.)) and No Prefiltering (NP). Confidence thresholds for sample incorporation and classification indicated as P(t) and CT(t) respectively. If threshold < 0.5, points labeled minority class if predicted confidence > threshold. Equal: equal samples, Min: only new minority class samples included.

labelled data are combined to fine-tune an SSL classifier, and iv) the SSL classifier and baseline classifier are evaluated against the validation split of the cross-validation. This process is repeated for each label-level (see Table 1). Figure 1 provides a visual overview of the experimental setup.

#### 4.1. Stage 1: Pre-filtering and Confidence Thresholds

The careful selection of new data, for example, by excluding data points with low-confidence classifications with the help of a baseline classifier, has been shown to be essential for successfully applying SSL methods (Van Engelen and Hoos, 2020). To achieve this, i) we use a relevance filter supplied by the GLoHBCD authors<sup>4</sup> to weed out change-unrelated (Follow/Neutral) sentences, and ii) we compare different confidence thresholds for pre-filtering and pseudo-labelling. We test all combinations of three confidence thresholds (0.5, 0.7, 0.95) for both the relevance filter and the baseline classifier that is used to pseudo-label new data. To avoid noise, we use only Weight Loss Forum Data (§3.2) as a source for pseudo-labelled data at this stage, as it stems from the same source as the GLoHBCD.

#### 4.2. Stage 2: Class Imbalance

Pseudo-labelled data is prone to class-imbalance even with a balanced baseline classifier (Wang et al., 2022). Such imbalance can severely impact performance (Guo and Li, 2022). In our first experimental stage, the majority class dominated pseudo-labelling, possibly suppressing SSL improvements. To address this, we tested additional strategies to boost minority-class representation.

From stage 1, we selected optimal pre-filter threshold combinations for each label-level. We then test the following variants: i) adding only minority-class pseudo-labelled samples, ii) adding

equal amount of pseudo-labelled samples for all classes, based on the number of minority-class samples, and iii) pseudo-labelling as minority class even with low confidence (thresholds: 0.4, 0.3, 0.2, 0.1).

#### 4.3. Stage 3: Amount and Domain of Pseudo-labelled Data

After initial proofs considering only Weight Loss Forum data, we wanted to assess to what extent the amount of new data added and the domain of pseudo-labelled data impact classification performance and transfer learning. To this end, we included the two other datasets described in §3.2, and tested the following combinations of datasets as providers of pseudo-labelled samples: i) Weight Loss Forum only, ii) Smoking Cessation Forum only, iii) AnnoMI only, iv) Weight Loss Forum + Smoking Cessation Forum, and v) Weight Loss Forum + Smoking Cessation Forum + AnnoMI.

At this stage, the confidence thresholds were set to those that yielded the best results in Stages 1 and 2. We added varying shares of pseudo-labelled data to the original training data (between 20%-100% in 20% increments).

Since the data from the Smoking Cessation Forum contains ground truth labels, we incorporated these examples to the 10-fold cross-validation experiments (at each round 90% of them were pseudo-labelled and fed to the classifier and the remaining 10% of them were included into the validation fold along with the GLoHBCD validation data).

#### 4.4. Stage 4: Application to Test Sets

In this final stage, we combine insights from stages 1-3 and applied the best performing system for each label-level to the independent test sets. Examining the SSL approach on data derived from chat-like conversational contexts and spanning various behaviour change domains aids in gauging its effectiveness and transfer learning capabilities. The main goal was to determine what kind of out-of-context data might benefit the most from

<sup>4</sup>[https://huggingface.co/selmey/behaviour\\_change\\_prefilter\\_german](https://huggingface.co/selmey/behaviour_change_prefilter_german)



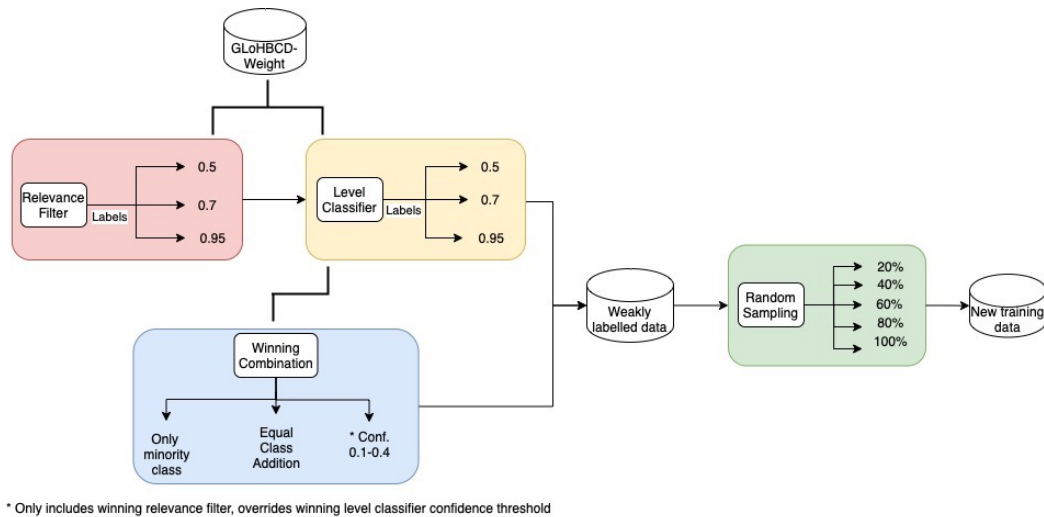


Figure 1: Framework plot of experimental setup.

SSL. We also expected to build on previous results (Meyer and Elsweiler, 2023) and gain additional insights about the interaction between a dataset’s properties and the difficulty of mining change behaviour cues from it.

## 5. Results

In Table 3 we summarise the classification results of the top three conditions from stages 1 and 2, in which only Weight Loss Forum data was pseudo-labelled, on the cross-validation splits of the GLoHBCD for each label-level compared to the baseline classifier. These results suggest that adding pseudo-labelled samples from the same source as the GLoHBCD has a minor yet discernible positive effect on classification. Most of the tested variants led to some improvements compared to the baseline, although the improvements were modest, and we did not observe any statistically significant differences between conditions. This outcome could potentially be attributed to the low amount of available new data.

Although there was no significant improvement in performance, none of the classification tasks experienced a decrease in performance when new pseudo-labelled data was added. When analysing the class-specific F1 scores and the amount of new data points added per class, we noticed that the F1 scores of the minority classes vary more than those of the majority classes. The amount of data labelled as the minority class is generally small, even in conditions where the confidence threshold for a sample to be labelled as the minority class was set lower than 0.5. In Figure 2, we show that this effect can be observed across all classification experiments (valence, topic, and reason type). Based on the results of stage 1 and

2, the systems chosen for the next stage of experiments were the following:

**valence level:** no prefilter, confidence threshold 0.5 (NP, CT(0.5)),

**topic level:** no prefilter, confidence threshold 0.5 (NP, CT(0.95)),

**reason type level:** prefilter with confidence 0.7, confidence threshold 0.5, and adding only minority samples (P(0.7), CT(0.5), min).

Applying those systems in stage 3 of experiments led to more stable classification results for the GLoHBCD validation sets compared to the Smoking Cessation Forum validation sets. This was expected since the smoking cessation data is from a different source and domain than the original training data, and has fewer samples.

However, regardless of the validation set and the type of pseudo-labelled data added, the results do not show a clear increase of performance when more data is added. Only in a few instances did adding out-of-domain data lead to improvements of in-domain classification. The effects of SSL seems to be slightly more apparent on the reason type level. For example, adding weight loss forum data led to slight improvements in reason type classifications of the GLoHBCD and Smoking Cessation Forum validation sets. The reason type classifiers work with few labelled data points from the original training data, thus presumably allowing pseudo-labelled samples to have more influence.

Based on the results obtained in stage 3, we included different shares of pseudo-labelled data from Weight Loss Forum, Smoking Cessation Forum and AnnoMI to predict on the test sets in Stage 4 (see Table 3). For valence and reason type classification, we included 20% of the pseudo-labelled

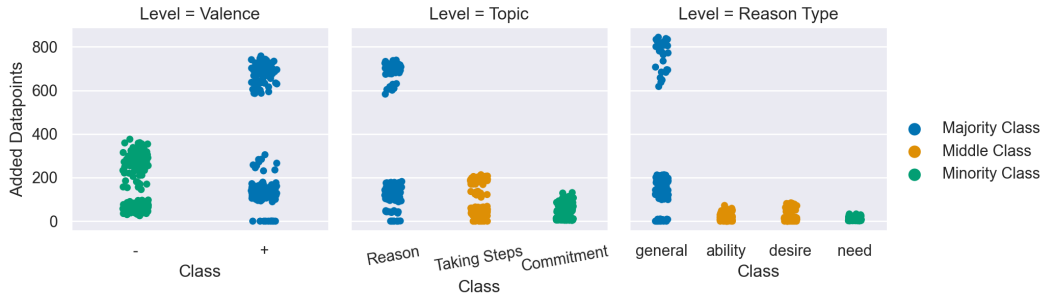


Figure 2: Amount of pseudo-labelled data points added to the training data across conditions by class.

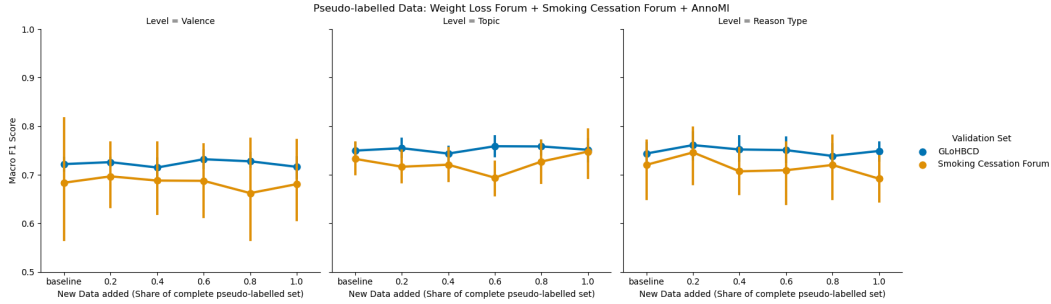


Figure 3: Change in classification performance when more pseudo-labelled data is added to training.

data, and for topic classification we included all pseudo-labelled samples. In Figure 4, we compare the classification performance of the baseline classifier without pseudo-labelled data and the best performing system from stages 1-3 for each label level.

For the topic level, all datasets but the Health Coaching Dialogue Corpus experienced some improvement in performance. Effects on valence and reason type were larger and more varied. The valence predictions by the SSL classifier on the Smoking Cessation Forum, the synthetic GPT-3 data, and the Health Coaching Dialogue Corpus were better than those of the baseline. Still, the SSL classifier produced poorer valence results for the Wizard of Oz dialogues, GLoHBCD and Optifast Data. For reason type, decreases in performance were observed for the synthetic GPT-3 data and the GLoHBCD test set, while performance on Optifast Data remained the same and all other datasets benefited from the inclusion of the pseudo-labelled data.

## 6. Discussion

Weak supervision has shown promising results in multiple previous studies working with curated benchmark datasets (Van Engelen and Hoos, 2020; Duarte and Berton, 2023). However, its effects appear to be more elusive when applied to imbalanced data. Although we found some slight improvements when applying the SSL-classifiers

to test datasets, transfer learning did not improve for all out-of-context data. With the baseline classifier reaching F1 scores between 70% and 80% on in-context data, one potential reason for the lack of stable classification improvements could be unsteady behaviour of the baseline classifier when labelling new data.

In their survey study, Longpre et al. (2020) highlighted that simply augmenting the training data of large pre-trained transformer models is insufficient to enhance classification performance. The reason behind this limitation lies in the fact that augmentation alone does not introduce the necessary linguistic variety to impart new knowledge to these powerful models. Drawing from this argument, one possible explanation for the minimal impact observed when incorporating pseudo-labelled samples, regardless of their source or label level, could be attributed to the uniformity of language surrounding behaviour change across different conversational contexts and behaviour change topics.

This observation aligns with the findings presented by Meyer and Elswiler (2023), who explored the transfer learning capabilities of classifiers for behaviour change language. In such a context, the addition of new training data, even from divergent sources, may not produce the required “newness” to improve classification performance. This is further exemplified by the fact that the AnnoMI, the largest dataset added during pseudo-labelling, not only stems from vastly differ-

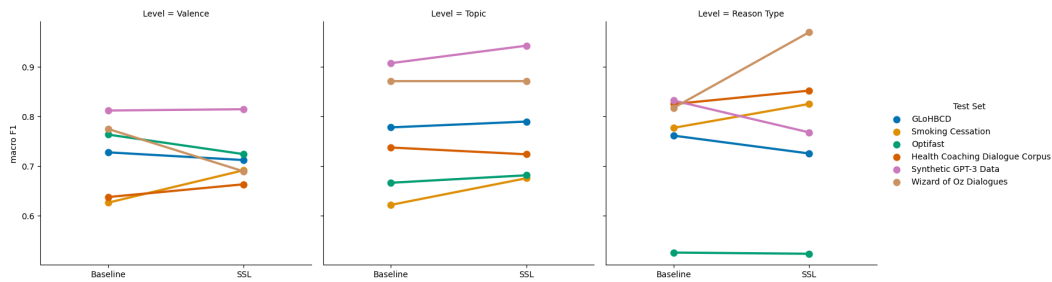


Figure 4: Comparison of classification results on multiple test sets. The plots present the performance of the baseline classifier (no semi-supervised learning) and a semi-supervised learning (SSL) classifier whose configuration was set based on the experiments of stages 1-3.

ent change scenarios, but even from a different modality (spoken conversation) compared to the GLoHBCD and this did not have a stable negative impact on classification results.

Generally, we discovered that the classification of behaviour change language remains stable and is not harmed by the inclusion of new data from alternative sources. These results speak in favour of the application of classifiers trained on behaviour change language to novel contexts. That being said, while our experiments do confirm that there are large parts of the data that seem to be very similar across contexts and target behaviours, there also seem to be some utterances that are more context-specific and might not be picked up correctly by the baseline classifiers used for pseudo-labelling.

<p><b>Sentence:</b> Aber dennoch heißt es heute, ganz besonders acht geben auf mich. (But still, the motto today is to take extra special care of myself.)</p> <p><b>Potential Codes:</b> C+, Rn+</p> <p><b>Sentence:</b> So komme ich wieder auf ein Fahrrad und mache mich etwas fitter. (So I get back on a bike and get a bit fitter.)</p> <p><b>Potential Codes:</b> C+, R+</p>
--

Table 4: Example of an ambiguous sentence from the training data

Behaviour change language itself could also be a limiting factor for the success of the approach. Although this type of language has been shown to be rather stable across domains and target behaviours (Meyer and Elsweiler, 2023), the inter-rater reliability when labelling such data is often low compared to other annotation tasks even among trained professionals (Meyer and Elsweiler, 2022; Wu et al., 2022; Hershberger et al., 2021; Tanana et al., 2016; Pérez-Rosas et al., 2016). A task in which even human annotators with extensive training do not reach high consensus is likely to produce many samples that are highly contestable, or could even be correctly attributed to multiple classes (see Table 4). As such, relying on only one prediction per data point might never lead to excellent F1 scores, as they can be found in easier classification tasks.

Lastly, all test sets are annotated on a sentence to sentence basis, and no context is passed to the classifier. Especially in the case of chat-data, where some utterances might be replies to questions from the conversational partner, this way of labelling could lead to important information being missed by the classifier. This could additionally hinder robust classification and the potential of SSL-learning.

These results leave us pondering over the oft-debated issue of whether the current emphasis on SOTA-chasing (Church and Kordoni, 2022) is indispensable or advantageous for the effective deployment of algorithms in practical settings. In some domains, especially those with a high number of debatable labels, it might be preferable to accept mid-range classification performance. In our future work, we plan to explore to what extent the current effectiveness of the models is sufficient for practical applications.

## 7. Limitations

We did not add extensive amounts of data, and the size of each dataset used as a source for weakly labelled data was smaller than the size of the original dataset. We consider this as one of the main limitations of this work and intend to approach this problem in future work by adding large quantities of weakly labelled data from various sources. Our experiments so far have suggested that the source of pseudo-labelled data does not have a significant impact on classification performance. Consequently, we intend to explore the possibility of using web sources, such as relevant Reddit forums. These new sources could provide large amounts of textual data, although the noisy nature of these sources may necessitate a re-evaluation of our selection criteria, including the recalibration of confidence thresholds.

Another limitation was that some test sets, used in the final stage of experiments, were very small and in some cases included only few to no data points for the smaller classes. This could poten-

tially have distorted our results and might have made the metrics more prone to outliers. Nonetheless, it is important to recognise that such imbalanced conditions may naturally occur when deploying these classifiers in real-life scenarios. In any case, we want to further explore the transfer capabilities of the solutions introduced here. For instance, by collecting and evaluating a larger dataset based on chat-like conversations around different target behaviours.

## 8. Conclusion

In this paper, we have attempted to shed light on the effectiveness of semi-supervised learning to increase both in-domain and transfer classification of written utterances concerning behaviour change. This is a low-resource classification task, where the learned classifiers can potentially be applied to data from various topics and across conversational contexts.

We found that adding pseudo-labelled data to the training sets had a stronger effect on the classification of smaller classes, whereas classification performance of the majority class remained fairly stable, regardless of the pre-filtering method or confidence thresholds. Observed effects were not stable across conditions, and adding larger amounts of data did not necessarily mean increased performance.

The transfer capabilities of the classifiers exhibited promising results in certain test scenarios. However, no consistent patterns or trends emerged when considering different label levels and target domains. Despite the lack of substantial performance enhancement through semi-supervised learning, there were also no noticeable deteriorations. This held true even when incorporating pseudo-labelled data from significantly distinct contexts, as evidenced by the AnnoMI collection. These findings highlight the robustness of the baseline classifier and its ability to effectively apply pseudo-labels to new data. Such outcomes could be attributed to the linguistic stability observed in the language pertaining to behaviour change across various contexts. These experiments underline the issue of unreliability of annotations in this domain hindering highly effective classification, leading us to question the need for high F1 scores in application areas like these.

## 9. Acknowledgements

The second and the fourth author thank: i) the financial support supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditation 2019-2022 ED431G-2019/04, ED431C 2022/19) and the European

Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System, ii) the financial support supplied by project PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; supported by the European Regional Development Fund), iii) the financial support supplied by project PLEC2021-007662 (MCIN/AEI/10.13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next Generation EU).

The fourth author also thanks the financial support obtained from project SUBV23/00002 (Ministerio de Consumo, Subdirección General de Regulación del Juego).

## 10. Bibliographical References

- Kenneth Ward Church and Valia Kordoni. 2022. Emerging trends: Sota-chasing. *Natural Language Engineering*, 28(2):249–269.
- Dawn Clifford and Laura Curtis. 2016. *Motivational Interviewing in Nutrition and Fitness*. Guilford Publications.
- José Marcio Duarte and Lilian Berton. 2023. A review of semi-supervised learning for text classification. *Artificial Intelligence Review*, pages 1–69.
- Amelie V Güntner, Paul C Endrejat, and Simone Kauffeld. 2019. Guiding change: using motivational interviewing within organizations. *Gruppe. Interaktion. Organisation. Zeitschrift für Angewandte Organisationspsychologie (GIO)*, 50:129–139.
- Lan-Zhe Guo and Yu-Feng Li. 2022. Class-imbalanced semi-supervised learning with adaptive thresholding. In *International Conference on Machine Learning*, pages 8082–8094. PMLR.
- Itika Gupta, Barbara Di Eugenio, Brian Ziebart, Aiswarya Baiju, Bing Liu, Ben Gerber, Lisa Sharp, Nadia Nabulsi, and Mary Smart. 2020. [Human-Human Health Coaching via Text Messages: Corpus, Annotation, and Analysis](#). In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 246–256, 1st virtual meeting. Association for Computational Linguistics.

- Paul J Hershberger, Yong Pei, Dean A Bricker, Timothy N Crawford, Ashutosh Shivakumar, Miteshkumar Vasoya, Raveendra Medaramitta, Maria Rechten, Aishwarya Bositty, and Josephine F Wilson. 2021. Advancing Motivational Interviewing Training with Artificial Intelligence: ReadMI. *Advances in Medical Education and Practice*, 12:613.
- Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. *Dynabench: Rethinking Benchmarking in NLP*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.
- Taewan Kim, Seolyeong Bae, Hyun Ah Kim, Su-woo Lee, Hwajung Hong, Chanmo Yang, and Young-Ho Kim. 2023. Mindfuldiary: Harnessing large language model to support psychiatric patients' journaling. *arXiv preprint arXiv:2310.05231*.
- Yu-Feng Li and Zhi-Hua Zhou. 2014. Towards making unlabeled data never hurt. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):175–188.
- Shayne Longpre, Yu Wang, and Chris DuBois. 2020. How Effective is Task-Agnostic Data Augmentation for Pretrained Transformers? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4401–4411.
- Sarah Peregrine Lord, Doğan Can, Michael Yi, Rebecca Marin, Christopher W Dunn, Zac E Imel, Panayiotis Georgiou, Shrikanth Narayanan, Mark Steyvers, and David C Atkins. 2015. Advancing methods for reliably assessing motivational interviewing fidelity using the motivational interviewing skills code. *Journal of substance abuse treatment*, 49:50–57.
- Molly Magill, Timothy R Apodaca, Brian Borsari, Jacques Gaume, Ariel Hoadley, Rebecca EF Gordon, J Scott Tonigan, and Theresa Moyers. 2018. A meta-analysis of motivational interviewing process: Technical, relational, and conditional process models of change. *Journal of consulting and clinical psychology*, 86(2):140.
- Selina Meyer and David Elswailer. 2022. GLoHBCD: A Naturalistic German Dataset for Language of Health Behaviour Change on Online Support Forums. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2226–2235.
- Selina Meyer and David Elswailer. 2023. Towards Cross-Content Conversational Agents for Behaviour Change: Investigating Domain Independence and the Role of Lexical Features in Written Language Around Change. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–13.
- Selina Meyer, David Elswailer, Bernd Ludwig, Marcos Fernandez-Pichel, and David E Losada. 2022. Do We Still Need Human Assessors? Prompt-Based GPT-3 User Simulation in Conversational AI. In *Proceedings of the 4th Conference on Conversational User Interfaces*, pages 1–6.
- William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. 2003. Manual for the motivational interviewing skill code (MISC). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*.
- William R Miller and Stephen Rollnick. 2002. *Motivational Interviewing, Second Edition: Preparing People for Change*. Applications of Motivational Interviewing Series. Guilford Publications.
- Theresa B Moyers, Tim Martin, Paulette J Christopher, Jon M Houck, J Scott Tonigan, and Paul C Amrhein. 2007. Client language as a mediator of motivational interviewing efficacy: where is the evidence? *Alcoholism: clinical and experimental research*, 31:40s–47s.
- Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31.
- Kathryn M Page and Irina Tchernitskaia. 2014. Use of motivational interviewing to improve return-to-work and work-related outcomes: a review. *The Australian Journal of Rehabilitation Counselling*, 20(1):38–49.
- Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51.
- Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2020. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2-3):709–730.

- Ken Resnicow, Fiona McMaster, and Stephen Rollnick. 2012. Action reflections: a client-centered technique to bridge the WHY–HOW transition in motivational interviewing. *Behavioural and cognitive psychotherapy*, 40(4):474–480.
- Viktor Schlegel, Erick Mendez-Guzman, and Riza Batista-Navarro. 2022. Towards Human-Centred Explainability Benchmarks For Text Classification. *arXiv preprint arXiv:2211.05452*.
- Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. 2016. A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing. *Journal of Substance Abuse Treatment*, 65:43–50.
- Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440.
- Paroma Varma and Christopher Ré. 2018. Snuba: Automating weak supervision to label training data. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 12, page 223. NIH Public Access.
- Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. 2022. Debaised learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14647–14657.
- Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181. IEEE.
- Bei Xu and Ziyuan Zhuang. 2022. Survey on psychotherapy chatbots. *Concurrency and Computation: Practice and Experience*, 34(7):e6170.

# Development of a Benchmark Corpus for Medical Device Adverse Event Detection

Susmitha Wunnava<sup>\*,†</sup>, David Harris<sup>†</sup>, Florence T. Bourgeois<sup>†,‡</sup>, Timothy A. Miller<sup>†,‡</sup>

<sup>\*</sup>Harvard-MIT Center for Regulatory Science, Harvard Medical School, Boston, MA, USA

<sup>†</sup>Computational Health Informatics Program, Boston Children’s Hospital, Boston, MA, USA

<sup>‡</sup>Department of Pediatrics, Harvard Medical School, Boston, MA, USA

susmitha\_wunnava@hms.harvard.edu

{david.harris, florence.bourgeois, timothy.miller}@childrens.harvard.edu

## Abstract

The U.S. Food and Drug Administration (FDA) collects real-world adverse events, including device-associated deaths, injuries, and malfunctions, through passive reporting to the agency’s Manufacturer and User Facility Device Experience (MAUDE) database. However, this system’s full potential remains untapped given the extensive use of unstructured text in medical device adverse event reports and lack of FDA resources and expertise to properly analyze all available data. In this work, we focus on addressing this limitation through the development of an annotated benchmark corpus to support the design and development of state-of-the-art NLP approaches towards automatic extraction of device-related adverse event information from FDA Medical Device Adverse Event Reports. We develop a dataset of labeled medical device reports from a diverse set of high-risk device types, that can be used for supervised machine learning. We develop annotation guidelines and manually annotate for nine entity types. The resulting dataset contains 935 annotated adverse event reports, containing 12252 annotated spans across the nine entity types. The dataset developed in this work will be made publicly available upon publication.

**Keywords:** medical devices, adverse event, natural language processing

## 1. Introduction

Medical device adverse events are undesirable, unexpected events that occur during or after the use of a medical device. The United States Food and Drug Administration (FDA) uses a multifaceted approach to monitor the safety and effectiveness of marketed devices. The Manufacturer and User Facility Device Experience (MAUDE) database is a passive surveillance system and the FDA’s primary post-market surveillance tool to capture real-world device-related deaths, serious injuries, and malfunctions. Other data sources include premarket clinical trials, and analysis of real-world data such as information from electronic health records (EHRs) (FDA, 2018). To support large-scale use of EHRs similar to what has been achieved in pharmacovigilance for drugs, the FDA introduced the unique device identification (UDI) system to enable identification and monitoring of devices. However, the program remains in an early phase given the slow adoption of UDIs by healthcare systems and the inability to efficiently identify and track device use (Kinard and McGiffert, 2020; Concato and Corrigan-Curay, 2022; Salazar and Redberg, 2020). Thus, the FDA continues to seek additional methodologies to support device surveillance activities.

Adverse event reporting enables FDA to take corrective action on problematic devices when safety concerns are identified (Levinson, 2009). The FDA’s MAUDE database, which contains all device-

related adverse event reports dating back to 1991, is publicly available on the FDA’s website. The FDA makes the MAUDE database available to “provide patients and health care professionals with important information they can utilize to make more informed medical decisions.” While spontaneous reports have limitations, most notably underreporting, many important safety signals have been initially identified using this information (Chung et al., 2020). The MAUDE database remains the primary mechanism for identifying safety signals for devices that require enforcement action, and was the most frequent source of device safety information leading to Medical Device Safety Warnings issued from 2011 to 2019 (Tau and Shepshelovich, 2020; Tomes, 2020). The database has also been used by investigators to assess the safety of specific devices across medical specialties (Coelho and Tampio, 2017; Tambyraja et al., 2005; Mahmoud et al., 2021).

Existing methods for safety signal detection from adverse event reports use statistical data mining methods such as disproportionality analysis, statistical process control, and sequential probability tests. These methods depend on structured data in the reports. While the data are rich in details regarding the specifics of the adverse event, most of it is free-form, unstructured text that requires processing and conversion into structured information for analysis. The few studies addressing device adverse event information extraction from text

BERLIN HEART GMBH BERLIN HEART EXCOR PEDIATRIC VAD; VENTRICULAR ASSIST DEVICE	<a href="#">Back to Search Results</a>
<p><b>Device Problem</b> Adverse Event Without Identified Device or Use Problem (2993)</p> <p><b>Patient Problem</b> Ischemia (1942)</p> <p><b>Event Date</b> 02/27/2021</p> <p><b>Event Type</b> Injury</p> <p><b>Manufacturer Narrative</b></p> <p>On (b)(6) 2021, a repeat head ct scan showed evolution of right territory distribution, laminar necrosis along superior right frontal component and right basal ganglia hyperdensity; all unchanged. Persistent apparent hypoattenuation in the left occipital and posterior temporal lobes was noted and may represent an acute infarct. The patient's asymmetrical facial and left leg weakness had improved.</p> <p><b>Event Description</b></p> <p>Berlin heart was informed by the site on 3/1/2021 that a patient being supported with the excor pediatric vad system in the lvad configuration had an ischemic cva event. The pump was full fill and ejection, but fibrin was noted in the pump. On (b)(6) 2021, the patient was found to have a left facial droop and an inability to move the left arm and leg. A head ct conducted on (b)(6) 2021 found a large right mca ischemic stroke with cerebral edema. Anticoagulation was stopped. The patient was started on keppra. A repeat head ct scan on (b)(6) 2021 showed no changes from the previous scan. On (b)(6) 2021, a pump change occurred for thrombus.</p> <p><a href="#">Search Alerts/Recalls</a></p>	

Figure 1: Example of adverse event report narrative from MAUDE

use rule-based methods consisting of user-defined rules for pattern matching to the raw text for information extraction (Alemzadeh et al., 2016; Penz et al., 2007). Rule-based systems are valued for their interpretability and ability to incorporate domain knowledge, but manually creating rules covering all possible information categories is labor intensive and requires high-level human expertise. The rules also apply to a small number of event types, making generalization expensive. Even fewer studies have applied supervised machine learning-based approaches towards device adverse event information extraction (Xie et al., 2018; Callahan et al., 2019). The few studies that automate information extraction have focused on specific device(s), limited data types, and a pre-determined set of basic adverse events.

Application of natural language processing (NLP) techniques to adverse event information extraction may provide an effective way to augment current approaches for post-marketing safety monitoring (Harpaz et al., 2014; Karimi et al., 2015). In this work we describe the development of a new dataset that will allow for fine-grained device-related adverse event information extraction, including important data types such as patient problems, device problems, reported patient outcomes and device information mentioned in the reports.

## 2. Background

In the context of drug safety surveillance and pharmacovigilance, many open challenges and shared tasks were conducted to assess and advance the state of the art in NLP for extraction of adverse drug events from clinical narratives (Uzuner et al., 2011; Henry et al., 2019; Jagannatha et al., 2019; Weissenbacher et al., 2019). Besides providing a venue for researchers to develop comparative systems on the same data and tasks, the challenges also made

a variety of annotated adverse drug events datasets available for future researchers to learn and build on the state-of-the-art systems. On the other hand, NLP for medical device adverse event detection is unexplored. Research in this area is also impeded by a lack of curated medical device adverse event detection datasets for developing NLP models, and limited research in device signal detection methods from unstructured text. This work is therefore addressing an unmet need, since it is the first to describe the creation of a novel medical device adverse event detection NLP benchmark dataset, a data genre that is medical but different from adverse drug events, EHRs, and other biomedical text.

## 3. Data and Preparation

### 3.1. Data Source

We use the FDA's MAUDE database, a publicly accessible resource with over 10 million records on medical device safety. Each report has structured fields that capture patient problem and device problem codes, but also two unstructured fields – manufacturer narrative and adverse event description (Figure 1). The adverse event information in the MAUDE reports might not be well-captured by the structured data. Detailed information about the adverse event in the unstructured part of the reports may play a key role in identifying additional events and safety signals that are missed in the structured data (Figure 2).

Natural language annotation (i.e., tagging text such as patient problems, product problems, and patient outcomes) is a key step for training machine learning models to automatically extract adverse event information from large-scale corpora. This requires the following steps we detail below: 1) Identifying important information from the reports, defining the entities that reflect this information, and



## Structured field → Patient Problems: Pain; Uterine Perforation

**Event Description:** THIS SPONTANEOUS CASE WAS REPORTED BY A LAWYER AND DESCRIBES THE OCCURRENCE OF **PELVIC PAIN ('PELVIC PAIN') AND UTERINE PERFORATION ('ORGAN PERFORATION')** IN A (B)(6) YEAR OLD FEMALE PATIENT WHO HAD ESSURE INSERTED FOR CONTRACEPTION. THE OCCURRENCE OF ADDITIONAL NON-SERIOUS EVENTS IS DETAILED BELOW. IN 2015, THE PATIENT HAD ESSURE INSERTED. ON (B)(6) 2018, THE PATIENT EXPERIENCED **ABDOMINAL PAIN ('ABDOMINAL PAIN')**. ON (B)(6) 2018, THE PATIENT EXPERIENCED **MENORRHAGIA ('EXCESSIVE BLEEDING / HYPERMENORRHEA')**. ON (B)(6) 2019, THE PATIENT EXPERIENCED DEVICE INTOLERANCE ('ESSURE INTOLERANCE'). ON AN UNKNOWN DATE, THE PATIENT EXPERIENCED **PELVIC PAIN** (SERIOUSNESS CRITERIA MEDICALLY SIGNIFICANT AND INTERVENTION REQUIRED), **UTERINE PERFORATION** (SERIOUSNESS CRITERIA MEDICALLY SIGNIFICANT AND INTERVENTION REQUIRED), **SWELLING ('SWELLING')**, **HYPERSENSITIVITY ('ALLERGIC REACTION')**, **HEADACHE ('HEADACHES')**, **BACK PAIN ('LOW BACK PAIN')**, **FATIGUE ('TIREDNESS')**, **ALOPECIA ('HAIR LOSS')**, **ANXIETY ('ANXIETY')**, **DEPRESSION ('DEPRESSION')**, **LIBIDO DECREASED ('LIBIDO DECREASE')** AND **ANGER ('ANGER REACTIONS')**. THE PATIENT WAS TREATED WITH SURGERY (TOTAL ABDOMINAL HYSTERECTOMY AND DOUBLE SALPINGECTOMY). ESSURE WAS REMOVED ON (B)(6) 2019. IN (B)(6) 2019, THE PELVIC PAIN, UTERINE PERFORATION, SWELLING, MENORRHAGIA, HYPERSENSITIVITY, HEADACHE, BACK PAIN, FATIGUE, ALOPECIA, ANXIETY, DEPRESSION, LIBIDO DECREASED, ANGER, ABDOMINAL PAIN AND DEVICE INTOLERANCE HAD RESOLVED. THE REPORTER CONSIDERED **ABDOMINAL PAIN, ALOPECIA, ANGER, ANXIETY, BACK PAIN, DEPRESSION, DEVICE INTOLERANCE, FATIGUE, HEADACHE, HYPERSENSITIVITY, LIBIDO DECREASED, MENORRHAGIA, PELVIC PAIN, SWELLING AND UTERINE PERFORATION** TO BE RELATED TO ESSURE. THE REPORTER COMMENTED: THE START DATE OF THE EVENTS WAS REPORTED AS 2015 (UNSPECIFIED). BASED ON THE AVAILABLE INFORMATION, A REVIEW OF OUR COMPLAINT RECORDS AND OTHER RELEVANT DATA WILL BE CONDUCTED; ANY NEW AND REPORTABLE INFORMATION THAT BECOMES AVAILABLE FROM OUR INVESTIGATION WILL BE PROVIDED IN A SUPPLEMENTARY REPORT

Figure 2: A sample report showing potential adverse events described in an unstructured “Event Description” field. Yellow highlighting indicates events that overlap with the structured data for the report, while blue indicates adverse events without corresponding structured data, and hence potential safety signals that were missed in the structured data.

creating annotation standards for annotators on the entities 2) Manually annotating a sample of reports with these entities. The key data extracted from the MAUDE database for this work is the unstructured device adverse event report narratives submitted to the FDA.

### 3.2. Dataset Creation

We create a large, diverse dataset of class III (high-risk) medical device adverse event reports from the FDA MAUDE database. Class III devices (e.g., pacemakers, blood vessel stents, cochlear implants) are implantable and/or life-sustaining devices that require premarket clinical safety and effectiveness data for approval. Any problems with these devices could lead to significant adverse outcomes for the patients. While class III devices constitute only 6.7% of all the devices, they make up more than 35.2% of device adverse event reports. Our sample includes reports of Class III devices with clinical safety and effectiveness data to maximize data usefulness for subsequent tasks. Devices are assigned to one of 491 “product categories”, to ensure a representative sample of devices, we include all product categories with at least one adverse event report and include up to a maximum of six reports per product category. Finally, we select reports that include narrative descriptions of the adverse event.

## 4. Annotation Protocol

### 4.1. Named Entity Annotations

We created annotation guidelines for the following nine named entities:

1. **Manufacturer.** The manufacturer of a device
2. **Device.** Type of device. Common/Generic name of device implanted/explanted, used in the diagnosis, cure, mitigation, treatment, or prevention of disease. And/Or The Proprietary/Trade/Brand name of the medical device (as used in device labeling or in the catalog).
3. **Device Problem.** The product problems that were reported to the FDA if there was a concern about the quality, authenticity, performance, or safety of any medication or device.
4. **Treatment.** Treatment of event the patient received. Medications/Device Therapy/Surgery in response to the adverse event. Name(s) of the drugs/ devices/ therapies mentioned in the treatment of the adverse event.
5. **Procedure.** Medical procedure for/during which the device is used. A device is either implanted, explanted, replaced, or applied.
6. **Adverse Event.** Adverse side-effects of the device on the patient (a.k.a. patient problems). These are medical conditions, signs or symptoms resulting from use (implanting/explanting/application) of the device.
7. **Indication.** Medical sign or symptom that is the basis or direct cause of treatment. Alternatively, it can be described as a medical condition for which a device implant/explant has been prescribed in the past or present.
8. **Other Medical Conditions (OMC).** Medical signs, symptoms, or disease names that are

neither being actively treated (Indications) nor are they adverse side effects (patient problems) of using a device.

9. **Outcome.** Outcome associated with the adverse event for a patient.

Manufacturer and Device categories are important for device name normalization because device names can be difficult to parse and are not consistently used across reports. Adverse Events, Indications, and Other Medical Conditions are all essentially “medical problems,” and so can appear superficially similar, but have crucial differences in how they should be interpreted. As a result, distinguishing them may make for a challenging task, but one that is vital for truly understanding the report. Treatment and Procedure categories are important to extract and distinguish since they are also superficially similar, yet have different interpretations, and which can also relate to the Outcome category. Overall, this set of categories attempts to capture the most important pieces of information in a report, potentially allowing for a variety of downstream applications.

## 4.2. Annotation Quality Control

We developed annotation guidelines and provided them to the annotators. We created rigorous annotation guidelines in an iterative process. The first draft guidelines included entity type descriptions, examples, and detailed instructions for challenging scenarios. Any ambiguous situations that arose during the annotation conflict resolution exercises were documented as examples for the guidelines. To ensure consistency and correctness, two annotators (a dedicated staff annotator with expertise in medical coding, and a regulatory scientist with experience working in the biomedical text mining domain) independently annotated a sample of reports after training, performed a check for agreement, and adjusted the annotation instructions to improve subsequent annotations. We use a web-based annotation tool called Label Studio (Maxim Tkachenko et al., 2023) to label the reports.

## 5. Annotation Results

To assess the quality of the manual annotations, we measure the inter-annotator agreement between the annotators using precision, recall, and F-measure, the performance metrics commonly applied in information retrieval tasks (Hripcsak and Rothschild, 2005). The two annotators labeled 130 reports with 2606 entity labels. The inter-annotator agreement yielded a precision of 0.71 and a recall of 0.68. In total, we labeled 935 adverse event reports with a total of 12252 labels spanning the nine

Entities	#Labels	Avg #Labels Per Report
Adverse Event	2993	4.17
Device	3410	3.99
Device Problem	964	2.52
Indication	385	2.01
Manufacturer	280	1.56
OMC	461	3.27
Outcome	70	1.46
Procedure	3144	4.05
Treatment	545	2.75

Table 1: Number of annotations per entity type in the dataset.

entities. We further split the corpus into train/test, resulting in 822, and 113 reports, respectively. The training/test set split is stratified such that the test set consists of devices that were not part of the training set. This split allows for domain adaptation-style experiments where an evaluation can be broken down into performance on devices that have been previously seen versus those that are new. All annotations are stored as JSON files as well as in CONLL2003 (Sang and De Meulder, 2003) data format suitable for the named entity recognition task. We report statistics on the labels. The occurrence of each named entity type is provided in Table 1.

## 6. Conclusions

Medical devices are more complex than pharmaceutical drugs, and faulty design and manufacturing are often the cause of device-related injuries. New devices are less likely to have their safety established clinically before they are marketed. Effective postmarket surveillance of high-risk medical devices is vital for early warning about safety issues. Reportable adverse events suggest that the device may have caused or contributed to a death or serious injury. Spontaneous reporting of adverse events is an important surveillance tool. Natural Language Processing (NLP) techniques can provide an effective way of post-marketing safety monitoring, but large domain-specific corpora are needed to train and assess high-performance NLP models. This work aims to address this unmet need by developing a benchmark corpus and annotated dataset for training and evaluating NLP approaches to extract adverse event information from medical device safety reports and help in improving the medical device safety surveillance process. The dataset can also be used for other natural language processing tasks such as text classification or question answering, among others. The dataset developed in this work will be made publicly available upon publication.

## 7. Bibliographical References

- Homa Alemzadeh, Jaishankar Raman, Nancy Leveson, Zbigniew Kalbarczyk, and Ravishankar K. Iyer. 2016. [Adverse Events in Robotic Surgery: A Retrospective Study of 14 Years of FDA Data](#). *PLOS ONE*, 11(4):e0151470.
- Alison Callahan, Jason A. Fries, Christopher Ré, James I. Huddleston, Nicholas J. Giori, Scott Delp, and Nigam H. Shah. 2019. [Medical device surveillance with electronic health records](#). *npj Digital Medicine*, 2(1):94.
- Gary Chung, Katherine Etter, and Andrew Yoo. 2020. [Medical device active surveillance of spontaneous reports: a literature review of signal detection methods](#). *Pharmacoepidemiology and Drug Safety*, 29(4):369–379.
- Daniel H. Coelho and Alex J. Tampio. 2017. [The Utility of the MAUDE Database for Osseointegrated Auditory Implants](#). *Annals of Otolaryngology & Laryngology*, 126(1):61–66.
- John Concato and Jacqueline Corrigan-Curay. 2022. [Real-World Evidence — Where Are We Now?](#) *New England Journal of Medicine*, 386(18):1680–1682.
- FDA. 2018. [FDA Medical Device Safety Action Plan: Protecting Patients, Promoting Public Health](#).
- Rave Harpaz, Alison Callahan, Suzanne Tamang, Yen Low, David Odgers, Sam Finlayson, Kenneth Jung, Paea LePendou, and Nigam H. Shah. 2014. [Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art](#). *Drug safety : an international journal of medical toxicology and drug experience*, 37(10):777–790.
- Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2019. [2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records](#). *Journal of the American Medical Informatics Association : JAMIA*, 27(1):3–12.
- George Hripcsak and Adam S. Rothschild. 2005. [Agreement, the F-Measure, and Reliability in Information Retrieval](#). *Journal of the American Medical Informatics Association : JAMIA*, 12(3):296–298.
- Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. 2019. [Overview of the First Natural Language Processing Challenge for Extracting Medication, Indication, and Adverse Drug Events from Electronic Health Record Notes \(MADE 1.0\)](#). *Drug safety*, 42(1):99–111.
- Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015. [Text and Data Mining Techniques in Adverse Drug Reaction Detection](#). *ACM Computing Surveys*, 47(4):1–39.
- Madris Kinard and Lisa McGiffert. 2020. [Medical Device Tracking—How It Is and How It Should Be](#). *JAMA Internal Medicine*.
- Daniel R Levinson. 2009. [Adverse event reporting for medical devices](#). *Office of Inspector General*.
- Karim Mahmoud, Sreenivasulu Metikala, Kathryn M. O'Connor, and Daniel C. Farber. 2021. [Adverse events related to total ankle replacement devices: an analysis of reports to the United States Food and Drug Administration](#). *International Orthopaedics*.
- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2023. [Label Studio: Data labeling software](#). Original-date: 2019-06-19T02:00:44Z.
- Janet F. E. Penz, Adam B. Wilcox, and John F. Hurdle. 2007. [Automated identification of adverse events related to central venous catheters](#). *Journal of Biomedical Informatics*, 40(2):174–182.
- James W. Salazar and Rita F. Redberg. 2020. [Leading the Call for Reform of Medical Device Safety Surveillance](#). *JAMA Internal Medicine*, 180(2):179.
- Erik F Sang and Fien De Meulder. 2003. [Introduction to the conll-2003 shared task: Language-independent named entity recognition](#). *arXiv preprint cs/0306050*.
- Rabindra R. Tambyraja, Michael A. Gutman, and Cliff A. Megerian. 2005. [Cochlear Implant Complications: Utility of Federal Database in Systematic Analysis](#). *Archives of Otolaryngology–Head & Neck Surgery*, 131(3):245.
- Noam Tau and Daniel Shepshelovich. 2020. [Assessment of Data Sources That Support US Food and Drug Administration Medical Devices Safety Communications](#). *JAMA Internal Medicine*, 180(11):1420–1426.
- Madris Tomes. 2020. [Identification and Market Removal of Risky Medical Devices](#). *JAMA Internal Medicine*, 180(11):1426–1427.
- Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. [2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text](#). *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–556.

- Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. [Overview of the Fourth Social Media Mining for Health \(SMM4H\) Shared Tasks at ACL 2019](#). In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 21–30, Florence, Italy. Association for Computational Linguistics.
- Jiaheng Xie, Xiao Liu, and Daniel Dajun Zeng. 2018. [Mining e-cigarette adverse events in social media using Bi-LSTM recurrent neural network with word embedding representation](#). *Journal of the American Medical Informatics Association*, 25(1):72–80.

# Using BART to Automatically Generate Discharge Summaries from Swedish Clinical Text

**Nils Berg, Hercules Dalianis**

Department of Computer and Systems Sciences (DSV)  
Stockholm University, Kista, Sweden  
nilsjv.berg@gmail.com, hercules@dsv.su.se

## Abstract

Documentation is a regular part of contemporary healthcare practices and one such documentation task is the creation of a discharge summary, which summarizes a care episode. However, to manually write discharge summaries is a time-consuming task, and research has shown that discharge summaries are often lacking quality in various respects. To alleviate this problem, text summarization methods could be applied on text from electronic health records, such as patient notes, to automatically create a discharge summary. Previous research has been conducted on this topic on text in various languages and with various methods, but no such research has been conducted on Swedish text. In this paper, four data sets extracted from a Swedish clinical corpora were used to fine-tune four BART language models to perform the task of summarizing Swedish patient notes into a discharge summary. Out of these models, the best performing model was manually evaluated by a senior, now retired, nurse and clinical coder. The evaluation results show that the best performing model produces discharge summaries of overall low quality. This is possibly due to issues in the data extracted from the Health Bank research infrastructure, which warrants further work on this topic.

**Keywords:** Patient Discharge Summaries, text summarization, clinical text, Natural Language Processing, Transformer, BART, synthetic text, negative results

## 1. Introduction

For clinicians in contemporary healthcare, documentation is a regular part of the daily tasks. One documentation task is the writing of discharge summaries. A discharge summary is a document created at the end of a care episode, such as a hospital admission, and documents that care episode (Scarfield et al., 2022). In this way, the discharge summary serves as one of the main tools of communication between secondary and primary care (Unnewehr et al., 2015).

Unfortunately, manually writing discharge summaries is a time-consuming process (Unnewehr et al., 2015), and as a consequence they are not always produced in time (Kripalani et al., 2007; Horwitz et al., 2013). Moreover, even when discharge summaries are made available in a timely manner, they are often of lacking quality in various respects (Kripalani et al., 2007; Unnewehr et al., 2015; Yemm et al., 2014; Callen et al., 2008; O’Leary et al., 2006; Braet et al., 2016).

Text summarization could potentially be applied to automatically summarize the text(s) which make up a hospital care episode, such as patient notes, into a discharge summary.

In recent years, state-of-the-art results have been achieved in text summarization with the use of Transformer-based language models (Alomari et al., 2022). One such model, and one which has achieved state-of-the-art results, is the Bidirectional

Auto-Regressive Transformers (BART) model (Alomari et al., 2022), which is the model used in this paper.

## 2. Related Research

Previous research employing extractive text summarization (ETS) to summarize patient notes into discharge summaries has been conducted on Finnish data using various language independent methods such as distributional semantics and specifically the random indexing method (Moen et al., 2016). Using Chinese data, previous research has employed various neural network based methods for ETS (Xiong et al., 2019).

Previous research using abstractive text summarization (ATS) has been conducted more frequently. Here, the MIMIC-III data set (Johnson et al., 2016) has been frequently explored with various methods, such as using recurrent neural networks (Diaz et al., 2020) and different Transformer-based language models (Hartman and Campion, 2022; Zhu et al., 2023; Pal, 2022). Summarizing data from Japanese electronic health records (EHRs) has also been explored (Ando et al., 2022).

In addition to the research done on this topic with ETS or ATS, research with hybrid text summarization (HTS) has also been performed. Here, the MIMIC-III data set has also been explored using different combinations of recurrent neural networks, Bidirectional Encoder Representations from Transformers (BERT) models, and BART models (Shing

et al., 2021).

### 3. Knowledge Gap

As described in the *Related research* section, the task of summarizing patient notes into a discharge summary has been explored in various languages in previous research. However, to the best of the authors' knowledge, no research has previously been conducted on summarizing Swedish patient notes into discharge summaries.

## 4. Methods

### 4.1. Data

In this study, data from the research infrastructure Swedish Health Record Research Bank<sup>1</sup> (Health Bank), held by the Department of Computer and Systems Sciences (DSV) at Stockholm University (Dalianis et al., 2015), was used for fine-tuning a BART language model. Health Bank covers patient data from over two million patient, extracted from the Karolinska University Hospital, Sweden, between 2006 and 2014. All patient notes used in this research had been automatically de-identified and anonymized.

#### 4.1.1. Data Set Structure

The data from Health Bank used in this paper is the so-called *Stockholm EPR Gastro ICD-10 Pseudo Corpus II* (hereinafter *Corpus II*) data set.<sup>2</sup> (Lamproudis et al., 2023), which consists of 351 730 patient notes, one row per note, in total 65 258 438 tokens, and encompasses 120 929 patients. Of these 351 730 notes, around 79 006 (22.4%) are discharge summaries. Each note is comprised of 6 columns, which are described below:

- `patientnr`, a unique serial number identifier for the patient which this patient note concerns. This identifier has no connection to any real-life identifier for the patient.
- `template_name`, a string identifier for the template used in the system which was used to create this patient note.
- `template_id`, an integer identifier for the template used in the system which was used to create this patient note.
- `recordnote_id`, a unique serial number identifier for this particular patient note.
- `codes`, the ICD-10 code(s) associated with this patient note, such as a diagnosis given to the patient at discharge.
- `full_note`, the free-text note written by the author of this patient note.

---

<sup>1</sup>Health Bank, <http://www.dsv.su.se/healthbank>

<sup>2</sup>This research has been approved by the Swedish Ethical Review Authority under permission, Dnr 2022-02386-02

Furthermore, the data set used in this paper did not have any internal structure or relationships, such as grouping of patient notes into care episodes, or any connection between patient notes and their corresponding discharge summary.

### 4.2. Model Used to Generate Discharge Summaries

As stated in the *Introduction* chapter, this paper makes use of a BART model to generate discharge summaries from patient notes. Specifically, a publicly available BART model<sup>3</sup> pre-trained on around 80 GB of Swedish text and developed by the Swedish National Library was fine-tuned on data from the Health Bank to perform the task of summarizing patient notes into discharge summaries. This model is referred to as *KB-BART* in this paper.

### 4.3. Data Pre-Processing

As described in section 4.1.1, there were no explicit relationships between patient notes, or between patient notes and discharge summaries. Thus, establishing what patient notes together form a care episode, and what discharge summary is related to that care episode, had to be done before the *KB-BART* model could be fine-tuned.

This task was performed first by sorting all patient notes first by the `patientnr` column, and then by the `recordnote_id` column in order to group patient notes belonging to one patient in chronological order. Then, all patient notes occurring chronologically between two discharge summaries where grouped as a care episode, and paired with the latter discharge summary.

After this pairing had been performed, the `full_note` column of all patient notes in a care episode associated with a discharge summary were concatenated together to form one text containing the entire care episode. Then, pairs where the text of the discharge summary was longer than the text of the care episode were discarded as this signified that the discharge summary did not actually summarize the care episode.

Once the pre-processing described above had been performed, the resulting care episode-discharge summary data set was used for fine-tuning the *KB-BART* model. Thus, the data set resulting from this pre-processing will be referred to as the *fine-tuning set* hereafter in this paper.

### 4.4. Fine-Tuning of the KB-BART Model

The process of fine-tuning the *KB-BART* model for the task of summarizing patient notes into discharge summaries was done in several steps. First, the fine-tuning set was split into four subsets via

---

<sup>3</sup>KB-BART, <https://huggingface.co/KBLab/bart-base-swedish-cased>

different methods of filtering out low quality samples. These subsets are in this paper referred to as *FULL*, *FILT1*, *FILT2* and *METR*, and were created in the following ways:

- The *FULL* subset was created by including all care episode-discharge summary pairs. Thus, it is identical to the full fine-tuning set.
- The *FILT1* subset was created by picking the pairs in the full fine-tuning dataset where at least one term in the care episode also existed in the corresponding discharge summary.
- The *FILT2* subset was created was created in the same way as *FILT1*, but instead of one term needing to be present in both the care episode and the discharge summary, 15% of all terms in the care episode needed to be present in the discharge summary in order to be included in the *FILT2* subset.
- The *METR* (short for "metrics" subset) was created by applying the three metrics *Semantic coherence*, *Topic similarity*, and *Redundancy* (Bommasani and Cardie, 2020). Care episode-discharge summary pairs which fell under certain thresholds in these three metrics were filtered out to create the *METR* subset, as this indicated low quality samples in text summarization (Bommasani and Cardie, 2020).

Additionally, for both *FILT1* and *FILT2*, all patient must have been created from more than one patient note.

Each subset was split into a training set, a validation set, and a test set, consisting of 80%, 10%, and 10% of the data respectively. For each of these constructed subsets, a *KB-BART* model, identical for each subset, was fine-tuned based on the training set of that subset. In this way, four different fine-tuned *KB-BART* models were created, in order to see what subset produced the fine-tuned *KB-BART* model with the highest performance.

#### 4.5. Evaluation of the Model Performance

In order to evaluate the performance of the four fine-tuned models, two types of evaluation were used: one automatic evaluation and one manual evaluation.

The automatic evaluation was based on Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004), calculated on the test set of the subset which the model was trained on. The average ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-S scores of the discharge summaries generated from the test set by the models were the main measurement of this evaluation. Furthermore, the results of the model which achieved the highest ROUGE scores were then compared against two benchmarks: Oracle and Random, derived from a similar work (Moen et al., 2016).

The Oracle benchmark is based on extractive text summarization and generates summaries by picking the sentences which maximize the ROUGE-2 F-score in regards to an existing summary, until a given length threshold is reached (Moen et al., 2016). This benchmark is not usable in any real-life scenario as it relies on the presence of an existing summary (which defeats the purpose of generating a summary) but it has its' uses as a benchmark to beat as it likely generates summaries with high ROUGE scores.

The Random benchmark is also based on extractive summarization, and generates summaries by randomly combining sentences until a given length threshold is reached (Moen et al., 2016). This makes it a benchmark which any sensible summarization method should be able to outperform. In this comparison, discharge summaries were generated by the benchmarks from the test set of the best performing model. Then, ROUGE scores were calculated based on these discharge summaries in order to compare them to the average ROUGE scores of the best performing model.

The manual evaluation was performed by a now retired senior nurse and clinical coder, and Swedish native speaker with several years of working experience (in this paper referred to as *the evaluator*). The evaluator manually reviewed ten randomly selected discharge summaries, generated by the model with the highest performance in the automatic evaluation, according to 12 criteria. Out of these 12 criteria, ten originated from a previous work where a similar manual evaluation was performed (Moen et al., 2016). Two criteria (criterion 9 and criterion 10) were added in order to evaluate potential hallucinations present in the generated discharge summaries. The criteria used for the manual evaluation are available in Table 1.

## 5. Results

### 5.1. Fine-tuning Set and Subsets

The pairing of care episodes to discharge summaries resulted in a fine-tuning set consisting of 20 345 care episode-discharge summary pairs. From this fine-tuning set, four subsets were derived, consisting of 20 345, 12 494, 2 575, and 7 722 care episode-discharge summary pairs, respectively.

### 5.2. Evaluation of the Model Performance

In the automatic evaluation, the mean ROUGE scores of the highest performing model was compared against two benchmarks. The results of this evaluation is available in Table 2. Based on these results, the model based on the *FILT2* subset performs best out of all the fine-tuned models. Thus, this model was further evaluated in the manual evaluation.

The implications of the results of the manual evaluation are described in section 6.2.

## 6. Discussion

### 6.1. Automatic Evaluation

The results from the automatic evaluation shows that model fine-tuned on the *FILT2* subset outperforms the other models in terms of all measured ROUGE score metrics.

Furthermore, the *FILT2* model outperformed the Random benchmark as well, which is positive as this implies that the *FILT2* model is better than randomness. However, since the margin with which the *FILT2* model outperformed the Random benchmark is not very significant, this implies that this model is not substantially better than randomness. Moreover, the *FILT2* model was outperformed by the Oracle benchmark in terms of all measured ROUGE score metrics, but even here it should be noted that the margin with which Oracle outperforms the *FILT2* model is not very significant. This implies that the *FILT2* model approaches the upper bounds of what is achievable on the *FILT2* subset test set (Moen et al., 2016).

#### 6.1.1. Comparison to Previous Research

In terms of the ROUGE scores, the results of the *FILT2* model is generally lower than the results reported in similar previous research in almost all cases for all metrics measured in this paper, see Table 3.

### 6.2. Manual Evaluation

Based on the results from the manual evaluation it can be stated that the *FILT2* model is prone to not include clinically important information in the discharge summaries that it generates when this information is available in the care episode that is being summarized. Based on the manual evaluation, the likelihood of including clinically important information differs depending on the type of information being summarized, with reason for admission and long-term diagnosis being least, and most, likely to be included in a generated discharge summary, respectively.

Furthermore, the discharge summaries generated by the *FILT2* model are also prone to include hallucinations of a severe nature in the discharge summaries that it generates. Moreover, based on the manual evaluation, discharge summaries generated by the *FILT2* model are lacking when it comes to readability, as the flow and overall content of the text was deemed to be very poor by the evaluator.

### 6.3. Aptitude of Data Set

As previously stated, the data set used in this paper consisted of only six columns, had no explicit

grouping of patient notes into care episodes or connection between care patient notes and discharge summaries. Furthermore, basic information such as the when, where, and by whom the patient notes were written was not present in the data set.

As a result, the task of grouping the patient notes in the data set together into care episodes, and then pairing those care episodes together with the correct discharge summaries was largely performed on the basis of assumptions. Thus, there is no guarantee that all, or even a majority of, care episodes in the data set have been correctly established and/or paired with their respective discharge summary.

## 7. Conclusions

### 7.1. Performance of the Fine-Tuned Model

In this paper, four instances of a BART model pre-trained on Swedish text were fine-tuned on four variations of a data set consisting of care episode-discharge summary pairs written in Swedish, extracted from the Health Bank research infrastructure, for the task of summarizing patient notes into discharge summaries. Based on an automatic evaluation, as well as a manual evaluation performed by a senior nurse and clinical coder, it can be concluded that the best performing fine-tuned BART model resulting from the work in this paper produces discharge summaries with severe shortcomings. Thus, this model is far from ready to be used in any real-life clinical setting.

### 7.2. Future Work

Since this is the first work to be conducted on the topic of automatically summarizing Swedish patient notes into discharge summaries, there are many possible directions for future work.

Firstly, while some efforts are made in this paper to correctly group patient notes together into care episodes, and then pair these with the correct discharge summary, future work should explore further efforts to more accurately perform this task. This task could either be performed with the current *Corpus II* data set, or aim to extract more data in the form of additional columns from Health Bank in order to alleviate this task.

Secondly, future work should explore the possibilities of using other text summarization methods with the *Corpus II* data set, such as Extractive Text Summarization, ETS or Hybrid Text Summarization, HTS. Perhaps of particular interest is HTS as it has shown good results in previous research on this topic (Shing et al., 2021).

Finally, future work should explore the possibility of replacing the *KB-BART* model used in this work with a similar model pre-trained on Swedish clinical text, rather than the "regular" Swedish text that *KB-BART* is pre-trained on. This is relevant as



ID	Question	Rating
1	Sender	Yes= 1, No= 0
2	Reason for admission	Yes= 1, No= 0
3	Long-term diagnosis	Yes= 1, No= 0
4	Procedures	Yes= 1, No= 0
5	Tests	Yes= 1, No= 0
6	Medication	Yes= 1, No= 0
7	Health status at discharge	Yes= 1, No= 0
8	Plans for the future	Yes= 1, No= 0
9	Does the summary contain information that cannot be traced back to the source notes?	Yes= 1, No= 0
10	If the question above is true, how serious is the incorrect information contained in the summary?	0.0 – 1.0 Trivial= 0.0, Severe= 1.0
11	Readability: how good is the flow of the text?	0.0 – 1.0 Bad= 0.0, Excellent= 1.0
12	Readability: how good is the content of the summary?	0.0 – 1.0 Bad= 0.0, Excellent= 1.0

Table 1: Evaluation criteria for manual evaluation. Partially adopted from (Moen et al., 2016, p. 8).

Model	$n$	R1	R2	RL	RS
<b>FULL</b>	1 491	0.197	0.042	0.099	0.034
<b>FILT1</b>	1 170	0.202	0.041	0.099	0.036
<b>FILT2</b>	227	<i>0.280</i>	<i>0.057</i>	<i>0.122</i>	<i>0.068</i>
<b>METR</b>	554	0.195	0.043	0.096	0.033
<b>Oracle</b>	227	<b>0.300</b>	<b>0.090</b>	<b>0.128</b>	<b>0.074</b>
<b>Random</b>	227	0.260	0.045	0.110	0.058

Table 2: Performance of the models on respective test set, along with benchmark performance on *FILT2* test set. Best score per metric among fine-tuned models in italic. Best score per metric overall in bold. All values rounded to three decimals.  $R1$ =ROUGE-1,  $R2$ =ROUGE-2,  $RL$ =ROUGE-L,  $RS$ =ROUGE-S.

previous research has shown that this approach can increase performance in downstream tasks (Jerdhaf et al., 2022). In doing this, the potential increase in the fine-tuned model’s performance on data from the Health Bank can be explored.

One interesting observation is the easiness to generate clinical language using the *KB-BART* model. This could prove to be an entrance point to generating large amounts of clinical text for use as training data, and should be explored further. However, the ethical issues in regards to the risk of generating text containing personal information in violation of the General Data Protection Regulation (GDPR) must be considered.

The data used for this article is available from Health Bank for academic use after registration by the user. The full work behind this paper is detailed

Work	R1	R2	RL
<b>Abstractive</b>			
(Diaz et al., 2020)	0.950	0.940	0.950
Hartman and Campion*	0.395	0.105	0.184
(Zhu et al., 2023)**	0.362	0.202	0.358
(Pal, 2022)***	0.383	0.238	0.349
(Ando et al., 2022)	0.153	0.196	0.121
This paper	0.280	0.057	0.122
<b>Extractive</b>			
(Moen et al., 2016)	0.382	0.184	0.367
(Xiong et al., 2019)	-	-	0.629
<b>Hybrid</b>			
(Shing et al., 2021)****	0.524	0.409	0.511

Table 3: Comparison with previous research. \* (Hartman and Campion, 2022). Results from the so-called "truncation approach", as this is the most approach most comparable to the approach in this paper. \*\* Results from the so-called *DISCHARGE* set, as this is the most approach most comparable to the approach in this paper. \*\*\* Results from the so-called *Setup 1* set, as this is the most approach most comparable to the approach in this paper, as well as one of the highest performing. \*\*\*\* Averaged results across sections from *RNN+RL ext + BART* model, to give a comparison across all sections of the EHR, as this paper does, for the best performing model.

in the first author’s master’s thesis (Berg, 2023).

## 8. Bibliographical References

- Ayham Alomari, Norisma Idris, Aznul Qalid Md Sabri, and Izzat Alsmadi. 2022. [Deep reinforcement and transfer learning for abstractive text summarization: A review](#). *Computer Speech & Language*, 71:101276.
- Kenichiro Ando, Mamoru Komachi, Takashi Okumura, Hiromasa Horiguchi, and Yuji Matsumoto. 2022. [Is In-hospital Meta-information Useful for Abstractive Discharge Summary Generation?](#) In *2022 International Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, pages 143–148, Tainan, Taiwan. IEEE.
- Nils Berg. 2023. Fine-tuning and evaluating a Swedish language model for automatic discharge summary generation from Swedish clinical notes. Master's thesis, Karolinska Institutet, Stockholm University.
- Rishi Bommasani and Claire Cardie. 2020. [Intrinsic Evaluation of Summarization Datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8075–8096, Online. Association for Computational Linguistics.
- Anja Braet, Caroline Weltens, Luk Bruyneel, and Walter Sermeus. 2016. [The quality of transitions from hospital to home: A hospital-based cohort study of patient groups with high and low readmission rates](#). *International Journal of Care Coordination*, 19(1-2):29–41.
- Joanne Callen, Melanie Alderton, and Jean McIntosh. 2008. [Evaluation of electronic discharge summaries: A comparison of documentation in electronic and handwritten discharge summaries](#). *International Journal of Medical Informatics*, 77(9):613–620.
- Hercules Dalianis, Aron Henriksson, Maria Kvist, Sumithra Velupillai, and Rebecka Weegar. 2015. [HEALTH BANK – A Workbench for Data Science Applications in Healthcare](#). *Proceedings of the CAiSE-2015 Industry Track co-located with 27th Conference on Advanced Information Systems Engineering*, Vol-1381:1–18.
- Diana Diaz, Celia Cintas, William Ogallo, and Aisha Walcott-Bryant. 2020. [Towards Automatic Generation of Context-Based Abstractive Discharge Summaries for Supporting Transition of Care](#).
- Vince Hartman and Thomas R. Champion. 2022. [A Day-to-Day Approach for Automating the Hospital Course Section of the Discharge Summary](#). *AMIA Annual Symposium proceedings. AMIA Symposium*, 2022:216–225.
- Leora I. Horwitz, Grace Y. Jenq, Ursula C. Brewster, Christine Chen, Sandhya Kanade, Peter H. Van Ness, Katy L. B. Araujo, Boback Ziaeeian, John P. Moriarty, Robert L. Fogerty, and Harlan M. Krumholz. 2013. [Comprehensive quality of discharge summaries at an academic medical center](#). *Journal of Hospital Medicine*, 8(8):436–443.
- Oskar Jerdhaf, Marina Santini, Peter Lundberg, Tomas Bjerner, Yosef Al-Abasse, Arne Jonsson, and Thomas Vakili. 2022. [Evaluating Pre-Trained Language Models for Focused Terminology Extraction from Swedish Medical Records](#). In *Proceedings of the Workshop on Terminology in the 21st century: many faces, many places*, pages 30–32, Marseille, France. European Language Resources Association.
- Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific Data*, 3(1):160035.
- Sunil Kripalani, Frank LeFevre, Christopher O. Phillips, Mark V. Williams, Preetha Basaviah, and David W. Baker. 2007. [Deficits in communication and information transfer between hospital-based and primary care physicians: implications for patient safety and continuity of care](#). *JAMA*, 297(8):831–841.
- Anastasios Lamproudis, Therese Olsen Svenning, Torbjørn Torsvik, Taridzo Chomutare, Andrius Budrionis, Phuong Dinh Ngo, Thomas Vakili, and Hercules Dalianis. 2023. [Using a large open clinical corpus for improved ICD-10 diagnosis coding](#). In *To appear in AMIA Annual Symposium Proceedings*. American Medical Informatics Association.
- Chin-Yew Lin. 2004. [ROUGE: A Package for Automatic Evaluation of Summaries](#). In *Proceedings of Workshop on Text Summarization of ACL, Spain*, pages 74–81.
- Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. 2016. [Comparison of automatic summarisation methods for clinical free text notes](#). *Artificial Intelligence in Medicine*, 67:25–37.
- Kevin J. O'Leary, David M. Liebovitz, Joseph Feinglass, David T. Liss, and David W. Baker. 2006. [Outpatient physicians' satisfaction with discharge summaries and perceived need for an electronic discharge summary](#). *Journal of Hospital Medicine*, 1(5):317–320.

Koyena Pal. 2022. Summarization and Generation of Discharge Summary Medical Report, [https://cs.brown.edu/media/filer\\_public/91/33/913389ac-49a0-4056-a886-424499c6e511/palkoyena.pdf](https://cs.brown.edu/media/filer_public/91/33/913389ac-49a0-4056-a886-424499c6e511/palkoyena.pdf), Visit date 2023-07-06.

Phoebe Scarfield, Thomas David Shepherd, Caitriona Stapleton, Alexandra Starks, Ellen Benn, Sara Khalid, Bryony Dayment, Alex Moate, Sandra Mohamed, and Jasmine Lee. 2022. Improving the quality and content of discharge summaries on acute medicine wards: a quality improvement project. *BMJ Open Quality*, 11(2):e001780.

Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. 2021. Towards Clinical Encounter Summarization: Learning to Compose Discharge Summaries from Prior Notes. ArXiv:2104.13498 [cs].

Markus Unnewehr, Bernhard Schaaf, Rusi Marev, Jason Fitch, and Hendrik Friederichs. 2015. Optimizing the quality of hospital discharge summaries – a systematic review and practical tools. *Postgraduate Medicine*, 127(6):630–639.

Ying Xiong, Buzhou Tang, Qingcai Chen, Xiaolong Wang, and Jun Yan. 2019. A Study on Automatic Generation of Chinese Discharge Summary. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1681–1687, San Diego, CA, USA. IEEE.

Rowan Yemm, Debi Bhattacharya, and David Wright. 2014. What constitutes a high quality discharge summary? A comparison between the views of secondary and primary care doctors. *International Journal of Medical Education*, 5:125–131.

Yunqi Zhu, Xuebing Yang, Yuanyuan Wu, and Wensheng Zhang. 2023. Leveraging Summary Guidance on Medical Report Summarization, Visit date 2023-07-06. <https://arxiv.org/abs/2302.04001>.

## 9. Language Resource References

The text set *Stockholm EPR Gastro ICD-10 Pseudo Corpus II* is available from the research infrastructure Swedish Health Record Research Bank<sup>4</sup>, at Stockholm University.

---

<sup>4</sup>Health Bank, <http://www.dsv.su.se/healthbank>

# Biomedical Entity Linking for Dutch: Fine-tuning a Self-alignment BERT Model on an Automatically Generated Wikipedia Corpus

Fons Hartendorp<sup>1</sup>, Tom Seinen<sup>2</sup>, Erik van Mulligen<sup>2</sup>, Suzan Verberne<sup>1</sup>

<sup>1</sup>Leiden Institute of Advanced Computer Science, Leiden University, the Netherlands

<sup>2</sup>Dept of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands

t.seinen@erasmusmc.nl, e.vanmulligen@erasmusmc.nl, s.verberne@liacs.leidenuniv.nl

## Abstract

Biomedical entity linking, a main component in automatic information extraction from health-related texts, plays a pivotal role in connecting textual entities (such as diseases, drugs and body parts mentioned by patients) to their corresponding concepts in a structured biomedical knowledge base. The task remains challenging despite recent developments in natural language processing. This paper presents the first evaluated biomedical entity linking model for the Dutch language. We use MedRoBERTa.nl as base model and perform second-phase pretraining through self-alignment on a Dutch biomedical ontology extracted from the UMLS and Dutch SNOMED. We derive a corpus from Wikipedia of ontology-linked Dutch biomedical entities in context and fine-tune our model on this dataset. We evaluate our model on the Dutch portion of the Mantra GSC-corpus and achieve 54.7% classification accuracy and 69.8% 1-distance accuracy. We then perform a case study on a collection of unlabeled, patient-support forum data and show that our model is hampered by the limited quality of the preceding entity recognition step. Manual evaluation of small sample indicates that of the correctly extracted entities, around 65% is linked to the correct concept in the ontology. Our results indicate that biomedical entity linking in a language other than English remains challenging, but our Dutch model can be used to for high-level analysis of patient-generated text.

**Keywords:** Biomedical Entity Linking, Dutch, Data and evaluation

## 1. Introduction

Biomedical entity linking (BEL) is the task of linking mentions of biomedical entities in free text to their corresponding canonical form in a knowledge base (Garda et al., 2023) (Figure 1). Entity linking is a commonly used step after entity extraction to enable normalization and aggregation of entity mentions. Applications include automatically categorizing and improving search in medical scientific literature and information extraction from clinical notes and patient forums (Lee et al., 2016). In the analysis of patient experiences and patient-doctor communication, BEL can identify common concepts and aggregate free-text mentions from different authors and contexts. For example, a patient on an online support forum might mention that they have trouble with sleeping after taking medication. A BEL model would be able to link the mention “trouble with sleeping” to the medical concept *insomnia* in a medical ontology and thereby aggregate all the mentions of insomnia from all patients.

Initial text pattern-based attempts to entity linking date back to the early 2000s, while modern models incorporate machine-learning algorithms (French and McInnes, 2022). The task remains challenging for four reasons: 1) The high diversity in surface form of identical biomedical terms. For example, *MI* and *hartaanval* (heart attack) both belong to the same canonical concept form *myocard infarct*

(myocardial infarction). 2) The similarity in surface form of different biomedical terms: *candida* and *cardia* refer to a yeast and the heart respectively, while their Levenshtein distance is only two. 3) Free text generated by patients and medical professionals is often noisy, including spelling errors and (personal) abbreviations. 4) The number of entities in the biomedical domain is very large. The Unified Medical Language System (UMLS), the largest biomedical ontology and composed of various medical vocabularies, contains more than 3.3 million unique concepts (Bodenreider, 2004; Vashishth et al., 2021).

Labeled biomedical entity linking datasets are limited, particularly in languages other than English.<sup>1</sup> In this paper, we present WALVIS, a weakly labeled Dutch biomedical entity linking dataset that was automatically generated using Wikidata and Wikipedia. We evaluate the quality of the WALVIS dataset and its effectiveness for training BEL models. Specifically, we train a BEL model for Dutch using self-alignment pretraining of BERT (sapBERT) (Liu et al., 2021) on a cleaned Dutch sample of the UMLS. We further fine-tune this model in a supervised setting on WALVIS and evaluate it on the Dutch subset of the Mantra GSC corpus (Kors et al., 2015). Additionally, we perform a case study on a collection of unlabeled, patient-support forum

<sup>1</sup><https://paperswithcode.com/datasets?mod=texts&task=entity-linking>

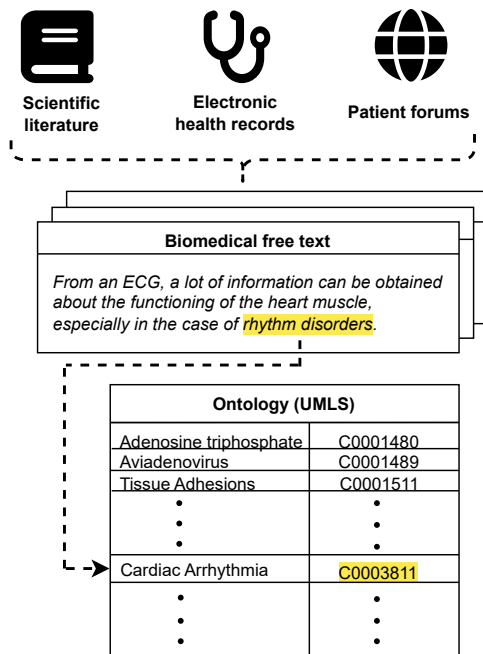


Figure 1: The task of biomedical entity linking. An entity recognition model identifies entities in free text that are then passed to the biomedical entity linking (BEL) model. The BEL model associates the new, unseen mention with its corresponding concept from an ontology.

data to give an indication of the effectiveness of our Dutch SapBERT on patient-written texts.

The contributions of this paper are as follows: 1) introduction of a method for automatically generating a weakly labeled BEL-dataset in any preferred target language, by combining the UMLS, Wikidata and Wikipedia and thereby obviating the need for manual labelling by a domain expert. 2) introduction of the first evaluated BEL model trained on the Dutch language. 3) Evaluation of the model's performance and generalizability on the Dutch portion of the Mantra GSC dataset. 4) An analysis of the model's performance on patient-generated text through a case study on an online patient-support forum. We release our code and data on github.<sup>2</sup>

## 2. Related Work

The goal of BEL is to associate an entity mention in a text with its corresponding concept in a medical ontology, usually the UMLS. BEL models are commonly part of pipelines including biomedical named entity recognition (NER), followed by BEL and finally relation extraction (French and McInnes, 2022). Some dedicated entity-linking corpora such as ShARe/CLEF and the NCBI dataset have been

<sup>2</sup>[https://anonymous.4open.science/r/biomedical\\_entity\\_linking-FCB4](https://anonymous.4open.science/r/biomedical_entity_linking-FCB4)

published (Pradhan et al., 2013), encouraging the development and evaluation of pure BEL-models without possible propagation of errors from the entity recognition module. A BEL-model typically involves a candidate generation step followed by candidate ranking (McInnes et al., 2009; D'Souza and Ng, 2015). In machine learning approaches to BEL, it is considered a *mapping* problem. However, learning the mapping function is complicated by the lack of large, labeled datasets for training and the huge amount of classes (Loureiro and Jorge, 2020).

With representation learning, the need for a labeled dataset can be obviated by leveraging the incorporated knowledge of a medical ontology. Since 2019, several BERT models (Devlin et al., 2019) for the biomedical domain have been released for English (Lee et al., 2020; Gu et al., 2021). For BEL, the entity embeddings are then further improved in a second-phase pretraining step by using information from the ontology (Sung et al., 2020; Liu et al., 2021). At inference, a similarity search is performed between the embedding of the new, unseen mention and the precomputed embeddings of all the terms from the ontology. The mention is then linked to the most similar term from the ontology. Self-alignment pretraining BERT (SapBERT) is a current state-of-the-art model that achieves 81.1% accuracy on the COMETA corpus and 52.2% on the MedMentions corpus (Liu et al., 2021; Basaldella et al., 2020; Loureiro and Jorge, 2020). Improvements have been attempted by incorporating context in the second-phase pretraining step or by using cluster-based inference (Zhang et al., 2022; Ujii et al., 2021; Angell et al., 2021). In the past years, generative language models have also been explored for the task of BEL (Yuan et al., 2022).

There is limited prior work on BEL for Dutch. There are two public medical annotation tools that includes Dutch and BEL: the rule-based Dutch implementation of MedSpaCy's QuickUMLS (Seinen et al., 2023), and the Dutch model pack for the MedCat library (Kraljevic et al., 2021). MedCat's linking module consists of two steps. First, dictionary matching is used for linking unambiguous terms (e.g. unique terms in the ontology, linked to one concept). Second, ambiguous terms are linked based on context embedding similarity. The context embeddings are trained on unambiguous terms and their context, in this case from the Dutch medical Wikipedia articles. Although MedCat uses similarity search with Word2Vec or BERT embeddings, the embeddings are not refined by leveraging knowledge incorporated in the ontology.

### 3. Preliminaries

BEL is the task of mapping entity mentions in text documents to canonical concepts in a given ontology. A mention is a string that describes an entity in natural language. A concept is semantic unit that is clearly defined in the ontology and has a unique identifier. Mentions (words/phrases from text) can refer to either real world entities or abstract concepts from the ontology. We formally define the task of BEL as follows:

**Problem definition** Given a biomedical ontology  $O$  consisting of  $n$  concepts  $O = \{c_1, c_2, \dots, c_n\}$ , a document  $D$  that contains a set  $M$  of  $p$  biomedical mentions  $M = \{m_1, m_2, \dots, m_p\}$ , the task of BEL is to learn a mapping  $M \rightarrow O$  that maps the mention  $m_j \in M$  to the corresponding concept  $c_i \in O$  that it refers to.

#### 3.1. Unified Medical Language System

The Unified Medical Language System (UMLS) is a large and comprehensive biomedical ontology created and maintained by the US National Library of Medicine. It is a collection of over 160 vocabularies, containing more than 15 million entries in 27 different languages. It maps entries from different databases and terminologies to around 3.3 million unique concepts, that are identified by their Concept Unique Identifier (CUI). The Dutch portion contains around 290,000 terms. The UMLS also contains data on 54 types of semantic relations between concepts, both hierarchical (e.g. ‘is a’) and non-hierarchical (e.g. ‘is conceptually related to’).<sup>3</sup>

#### 3.2. Self-Alignment Pretraining BERT

The main challenge of BEL in a representation learning setting is the quality of the entity embeddings (Basaldella et al., 2020). Self-supervised learning with masked language modelling on medical data has improved BEL, but does not lead to a well separated representation space (Liu et al., 2021).

Self-alignment pretraining (sap) (Liu et al., 2021) improves the embeddings of a pretrained BERT model, by self-aligning synonymous entries from a biomedical ontology. Formally, the goal of self-alignment is to learn a function  $f(\cdot; \theta) : O \rightarrow E$  that is parameterized by  $\theta$  and where  $O$  represents the set of terms in an ontology and  $E$  the corresponding embedding representations with  $\forall e \in E, e \in \mathbb{R}^d$ . In sapBERT,  $f$  is modelled by a BERT model with the output [CLS] token as embedding representation of the input term  $c$ . The similarity between two

<sup>3</sup>[https://www.nlm.nih.gov/research/umls/knowledge\\_sources/metathesaurus/release/statistics.html](https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html)

terms,  $\langle f(c_i), f(c_j) \rangle$  can be estimated by taking the cosine similarity. During the training procedure, *online hard triplet mining* is used for generating informative pairs that are used for contrastive learning. From each mini-batch, a random anchor term  $c_a$  is drawn. Together with a positive match – or synonym –  $c_p$  and a negative match  $c_n$ , the triplet  $(c_a, c_p, c_n)$  is formed. Informative triplets are generated by selecting positive matches (synonyms in the ontology) that get very dissimilar embeddings and, conversely, negative matches with embeddings that are nearly similar. Formally, triplets are selected that violate the following condition:

$$\|f(c_a) - f(c_p)\|_2 < \|f(c_a) - f(c_n)\|_2 + \lambda \quad (1)$$

where  $\lambda$  is a pre-set margin. That is, we only select those triplets where the distance between the anchor and the positive term is larger than the distance between the anchor and the negative term plus margin  $\lambda$ . The mining of informative triplets only is useful for improving the embeddings, since otherwise non-informative triplets would dominate the training process due to the enormous size of the ontology (Liu et al., 2021). The Multi-Similarity loss function is used for pulling the embeddings of positive pairs closer and pushing the embeddings of negative pairs further apart (Wang et al., 2019). This process leads to a better separated representation space by leveraging the semantic biases of synonymy relations in the ontology.

## 4. Methods

Due to the need for expensive, manually labelling by domain experts, BEL datasets are not broadly available, especially in languages other than English. We introduce a method for automatically generating a weakly labeled BEL dataset in any given target language, by combining the structured knowledge source Wikidata, the UMLS and inter-article hyperlinks on Wikipedia. We implement the pipeline for Dutch. We first clean and enhance the Dutch subset of the UMLS and generate a Dutch biomedical ontology specifically tailored for BEL tasks.

### 4.1. Enhancing the UMLS

Roughly 1.7% of the UMLS 2022AB release, comprising 290,056 terms, is in Dutch.<sup>4</sup> However, there is variability in the quality of the records. By following the same steps as the Dutch medical concepts project<sup>5</sup>, we created a cleaned, UMLS-based

<sup>4</sup>[https://www.nlm.nih.gov/research/umls/knowledge\\_sources/metathesaurus/release/index.html](https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/index.html)

<sup>5</sup><https://github.com/umcu/dutch-medical-concepts>

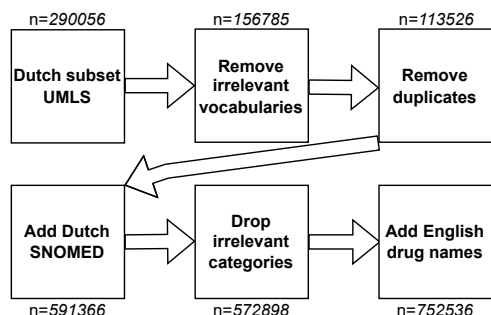


Figure 2: Flow diagram of ontology enhancement. The remaining number of entries are denoted in *italic*.

Dutch biomedical ontology in several filtering and expansion steps.<sup>6</sup> An overview is provided in Figure 2. We exported, using the MetaMorphosys-tool, the terms of all seven Dutch vocabularies that are included in the release. From the 290,056 terms, we removed the vocabularies LNC-NL-NL and ICPC2ICD10DUT that contain composed terms that are non-informative for the task of BEL such as `report:finding:date:polyclinical:document:endocrinology`.<sup>7</sup> From the vocabularies ICD10DUT and MDRDUT we removed descriptive subterms, such as non-specified, as they are usually not found in free text. Also, duplicate entries were dropped, irrespective of capitalization. We added the Dutch SNOMED vocabulary, as this is not included in the UMLS. Since the US SNOMED is included in the UMLS, we matched Dutch to English terms on their SNOMED ID, and subsequently assign them their corresponding UMLS IDs (CUIs), dropping ambiguous terms. Entries linked to one of 26 semantic types that we considered non-relevant for BEL, such as *Birds* and *Geographic areas* were also removed. Finally, we added English drug names from the ATC, DRUGBANK and RXNORM vocabularies, since they are occasionally used in Dutch (Miller and Britt, 1995; Wishart et al., 2018; Nelson et al., 2011).

The newly generated Dutch biomedical ontology contains 752,536 terms sourced from 11 vocabularies, all linked to one of the 366,071 distinct concepts. On average, each term is associated with one synonym, but the distribution is heavily right skewed (25% percentile is 0 synonyms per term / 75% percentile is 2 synonyms per term). Table 1 shows the semantic group distribution of the ontology. The semantic groups are not classes in our entity linking problem, but rather a categorization of the classes.

<sup>6</sup>Note that we cannot re-use their data because the UMLS is licensed and cannot be re-shared.

<sup>7</sup>Translated from Dutch for the readers' convenience

```

SELECT ?concept ?conceptLabel ?cui ?article
WHERE {
  ?concept wdt:P2892 ?cui .
  ?article schema:about ?concept .
  ?article schema:isPartOf
    <https://nl.wikipedia.org/>.

SERVICE wikibase:label {
  bd:serviceParam wikibase:language "nl"
}
  
```

Listing 1: SPARQL query for retrieving all Wikidata entities that contain a UMLS CUI and where there exists an article about the entity that is part of the Dutch Wikipedia

The four largest groups – disorders (DISO), chemicals & drugs (CHEM), procedures (PROC) and anatomy (ANAT) – make up for 97% of the terms in the ontology.

Since the UMLS and SNOMED are licensed, we cannot distribute the ontology. However, comprehensive details of all steps are provided in a Python Notebook in the project's Github repository. The ontology can be reproduced after requesting a UMLS and SNOMED license.

## 4.2. Corpus Compilation

For the automatic generation of our weakly labeled dataset WALVIS, we combine our enhanced ontology with textual data from Wikipedia and structured data from Wikidata. An overview is provided in Figure 3. Wikidata is a collaboratively edited multilingual knowledge graph that acts as central storage for structured data of its Wikimedia sister projects including Wikipedia (Vrandečić and Krötzsch, 2014). Relevant data can be obtained from Wikidata through SPARQL queries. We retrieve all 4,519 Wikidata entities that have been assigned a UMLS CUI and are linked to a Wikipedia page that is part of the Dutch Wikipedia collection. The SPARQL query is provided in Listing 1.

We process all pages from the Dutch Wikipedia dump of March 2023<sup>8</sup> using the SpaCy sentence splitter with the Dutch `nl_core_news_sm` pipeline. We then collect all 51,693 sentences that contain a hyperlink to one of the 4,519 Dutch Wikipedia articles that on their turn are linked to a Wikidata entity with a UMLS CUI property. The anchor texts of the hyperlinks are considered biomedical entity mentions. On average, a sentence contains 18 ( $\pm 9$ ) tokens and 53,960 (0.06%) of the tokens in the

<sup>8</sup><https://dumps.wikimedia.org>

Table 1: Semantic group distributions of the ontology, train- and validation set of the no-duplicates, ontology-filtered subsets (\*) from the WALVIS corpus (WALVIS\*) and Mantra GSC (Mantra\*). DISO: *disorders*, CHEM: *chemicals & drugs*, PROC: *procedures*, ANAT: *anatomy*, LIVG: *living beings*, PHEN: *phenomena*, DEVI: *devices*, PHYS: *physiology* and ACTI: *activities & behaviors*, OBJC: *Objects*, GENE: *genes & molecular sequences*, OCCU: *occupations*, CONC: *concepts & ideas*. 1559 terms in the ontology are not assigned a semantic group (other).

GROUP	EXAMPLE	ONTOLOGY		WALVIS* TRA.		WALVIS* VAL.		MANTRA*	
		COUNT	PERC.	COUNT	PERC.	COUNT	PERC.	COUNT	PERC.
DISO	MS ( <i>MULTIPLE SCLEROSIS</i> )	310057	41.3	957	49.8	224	46.7	149	39.3
CHEM	Neupro	185096	24.6	402	20.9	108	22.5	66	17.4
PROC	Dialyse ( <i>DIALYSIS</i> )	124345	16.6	90	4.7	20	4.2	68	17.9
ANAT	Heup ( <i>HIP</i> )	108622	14.5	391	20.4	105	21.9	33	17.4
LIVB	Patiënt ( <i>PATIENT</i> )	7586	1.0	14	0.7	6	1.2	29	7.7
PHEN	Licht ( <i>LIGHT</i> )	5997	0.8	4	0.2	1	0.2	7	1.8
DEVI	IUD's	3153	0.4	3	0.2	0	0.0	5	1.3
PHYS	Groei ( <i>GROWTH</i> )	3125	0.4	33	1.7	11	2.3	19	5.0
ACTI	Macht ( <i>POWER</i> )	1053	0.1	0	0.0	0	0.0	1	0.3
OBJC	Stof ( <i>FABRIC</i> )	678	0.1	13	0.7	3	0.6	2	0.5
GENE	Codon	497	0.1	3	0.2	2	0.4	0	0.0
OCCU	Genomics	464	0.1	10	0.5	0	0.0	0	0.0
CONC	Retentie ( <i>RETENTION</i> )	304	0.0	0	0.0	0	0.0	0	0.0
OTH.		1559	0.2	0	0.0	0	0.0	0	0.0
<b>TOTAL</b>		<b>752536</b>		<b>1920</b>		<b>480</b>		<b>379</b>	

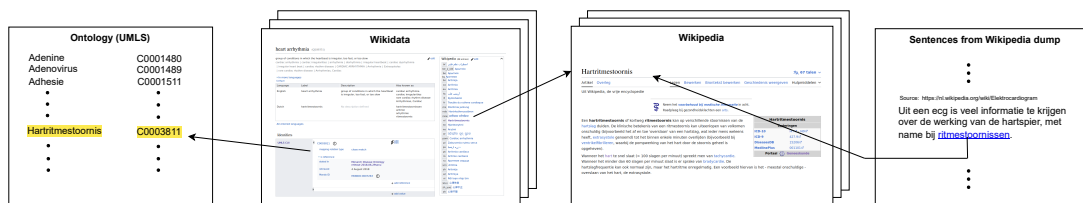


Figure 3: WALVIS corpus compilation. All Wikidata entries with a linked Dutch Wikipedia page and a UMLS CUI that are in the ontology are retrieved using SPARQL. Then, all sentences from the Wikipedia dump are parsed and selected if they contain a hyperlink to one of the collected Wikipedia pages.

collection are biomedical entity mentions that are linked to a UMLS CUI (Table 2).

For the WALVIS\* subset, we kept each first unique mention and dropped their duplicates. Also, only mentions that link to a CUI that is present in the ontology are included. The WALVIS\* corpus contains 2,400 unique mentions from 2,307 sentences. 1,751 mentions are unseen by our model in the sapBERT training phase as they are not present in the ontology. The mentions map to 1,086 unique CUI's that are all included in the ontology.

In Table 1, we see that the distribution of mentions over the semantic groups in the train- and validation set of WALVIS\* is relatively similar to the distribution of terms in the ontology over the semantic groups, except for procedures (PROC). Procedures are possibly terms more commonly used by medical experts only, compared to disorders, chemicals & drugs and anatomical terms, which could explain their lower prevalence on Wikipedia.

The code for parsing the Wikipedia dump and creating the corpus is available on Github, and the the WALVIS corpus and WALVIS\* subset are available for download in XML format.

### 4.3. Self-Alignment Pre-training

We use the RoBERTa-derived language model MedRoBERTa.nl as base model. MedRoBERTa.nl was pretrained on nearly 10 million anonymized hospital notes obtained from the Amsterdam University Medical Centres (Verkijk and Vossen, 2021). The model is distributed with uninitialized head layers, allowing for fine-tuning on specific tasks.

We generate the training data for the self alignment pretraining from the cleaned Dutch medical ontology (Section 4.1). We generate a text file with positive pairs in the form of: CUI || term 1 || term 2, where term 1 and term 2 are synonyms, so associated to the same CUI in the ontology. If more than 2



Table 2: Corpora statistics. The WALVIS corpus contains many duplicate mentions that occur in different contexts. The WALVIS\* subset (Wal.\*) contains no duplicate mentions and only links to CUI's that have an entry in the ontology. We created a similar subset of the Mantra GSC corpus (Man.\*).

	WALVIS	WAL.*	MAN.*
SENTENCES	51515	2307	166
AVG. #TOK/SENT	18	20	17
MENTIONS	53781	2400	379
UNIQUE MENTIONS	3201	2400	379
UNSEEN MENTIONS	49497	1751	214
CUI's	56141	2758	402
UNIQUE CUI's	1334	1086	359
UNLINKABLE CUI's	47548	0	0

terms are associated to the same CUI, all pairwise combinations are traversed and added. We sample from the pool of positive pairs during the contrastive learning step for improving the pretrained BERT embeddings. Negative pairs are sampled online by randomly drawing a term from the ontology that is not linked to the same CUI. Both the negative and positive pairs must violate the minimum margin condition in Equation 1.

We use Multi-Similarity loss for re-aligning of the embeddings with parameters set to the same values as in Liu et al. (2021). We use a learning rate of 0.0001 with a weight decay of 0.01 for  $\{0, 1, 3, 10\}$  epoch(s) with a batch size of 512. The similarity margin  $\lambda$  is set to 0.2. The [CLS] token is used as representation of the input term. The model is built in Pytorch 2.1.0, mostly based on code from Liu et al. (2021).<sup>9</sup>

#### 4.4. Fine-tuning

Fine-tuning on the WALVIS corpus is performed in a similar manner. Now, the positive pairs are generated by combining mentions, linked terms and their corresponding CUI from the labeled dataset: CUI||mention||linked term. The hyperparameters are set to the same values as in Section 4.3. We fine-tune for  $\{0, 1, 3, 10\}$  epoch(s), building on the pretrained models from the previous step.

#### 4.5. Inference

All terms from the ontology are fed to the trained model, generating a set of precomputed embeddings. At inference, a new, unseen mention is also fed to the trained model and a nearest neighbour search can be performed with the precomputed embeddings. The new mention is assigned the CUI of the most similar embedding from the ontology.

<sup>9</sup><https://github.com/cambridge/ttl/sapbert>

Since a nearest neighbour search on 752,536 items is computationally expensive, we built a FAISS index from the precomputed embeddings. FAISS is a library for approximate nearest neighbour search of dense vectors.<sup>10</sup> For memory purposes, the precomputed embeddings are first compressed by using only their first 256 principal components.

#### 4.6. Evaluation Data and Metrics

We evaluated our method on the Dutch subset of the Mantra GSC corpus. The Mantra GSC corpus is a hand-labeled corpus annotated by domain-experts that was originally created for biomedical concept recognition in languages other than English (Kors et al., 2015). The texts are sourced from MEDLINE titles and drug labels. The biomedical entities are also annotated with a UMLS CUI, that we use as gold labels for our linking model. Since the ontology does not contain all UMLS CUIs, we use the WALVIS\* and Mantra\* subsets that contain only mentions that link to a CUI that is included in the ontology. In both corpora, duplicate mentions were also removed since our model is not context-aware. Table 2 shows the corpora statistics of WALVIS\* and Mantra\*. The Mantra\* subset contains 379 mentions from 166 sentences. The sentences are slightly shorter than the WALVIS\* sentences, on average 17 tokens per sentence, and have more entity mentions per sentence.

For finding the optimal number of sapBERT- and fine-tune epochs, we performed a hyperparameter optimization on the train set of WALVIS\* and validated on its validation set. In the evaluation phase, we fine-tuned our optimal model on the full WALVIS\* subset and evaluated on the Mantra\* corpus.

In addition to our primary metric classification accuracy, we also look at the 1-distance accuracy. For this metric, predictions are scored correct if they are any kind of 1-distance UMLS relation away from the gold label. For example, the prediction *cystopyelonephritis* for a term with gold label *pyelonephritis* would be correct since the UMLS contains a 'classified as'-relation between the two.

#### 4.7. Case Study on Patient-support Forum

The Dutch online patient-support forum <https://www.kanker.nl/> donated anonymized textual data from between 2013 and 2016 in the form of blog posts, discussions and question-answering threads. The data does not have any manual annotations. We split the data in sentences using pySBD's sentence splitter, finding 123,338 sentences and 2,191,424 tokens. Before being able to apply BEL,

<sup>10</sup><https://github.com/facebookresearch/faiss/wiki>

we need entity extraction. To that end, we finetune MedRoBERTa.nl for NER on the machine translated MedMentions dataset in Dutch (Seinen et al., 2024).<sup>11</sup> This gives us 368,840 medical named entities. We run both the base model and our finetuned BEL model on these entities to link them to the Dutch UMLS and analyze the results.

## 5. Results and Analysis

We first assess the quality of our automatically generated WALVIS-corpus. We then turn to an evaluation of the optimal model on Mantra\* and perform a brief error analysis. Finally, we explore its performance on entities sourced from the patient-support forum.

### 5.1. Quality of WALVIS-Corpus

We randomly sample 100 mentions from the WALVIS-corpus and manually evaluate the correctness of their label (Wikidata–UMLS link). The grading was performed by the first author using a tool that was developed for this purpose, for easy comparison of UMLS entries.<sup>12</sup> 28 mentions were found to be linked to a concept that is related but not the same. For example, the mention *kerndelingen* (divisions of the nucleus) on the Dutch Wikipedia page about asexual reproduction, is linked to *cell nucleus*, which is related but not the same. The remaining 72 mentions seem to be labeled correctly. The label quality score indicates that the quality of the automatically generated corpus is suboptimal and that the data is not suited for evaluation purposes. The 100 samples and their grading can be found in our Github repository.

### 5.2. Main Results

In hyperparameter tuning, the model with 3 sapBERT epochs and 10 fine-tune epochs performed optimal with a classification accuracy of 30.5% and a 1-distance accuracy of 49.8% on the validation set of WALVIS. Table 3 shows the results on the Mantra\* corpus. All results are averaged over 5 runs with different random seeds. The model (3S10FT) achieves a classification accuracy of 54.7% and a 1-distance accuracy of 69.8%. That is a 10.1% point and 13.1% point improvement respectively, compared to the base model (BM).

In Table 3, the results grouped by semantic group are separately shown. We do not see a clear relation between the size of the semantic groups in

Table 3: Evaluation results on the Mantra\* corpus for the base model (BM) and our optimal model (trained for 3 self-alignment epochs + 10 fine-tune epochs). The semantic groups are not classes themselves, but rather a categorization of the classes. DISO: *disorders*, PROC: *procedures*, CHEM: *chemicals & drugs*, ANAT: *anatomy*, LIVG: *living beings*, PHYS: *physiology*, PHEN: *phenomena*, DEVI: *devices*, OBJC: *Objects*, ACTI: *activities & behaviors*. The total micro-average is shown for all 379 mentions averaged over 5 experiment runs with different random seeds.

GROUP	#	ACCURACY		1-DIST ACC.	
		BM	3S10FT	BM	3S10FT
DISO	149	49.3	59.6	63.0	77.0
PROC	68	29.7	39.5	41.5	56.1
CHEM	66	48.2	57.6	58.5	67.3
ANAT	33	57.6	66.7	66.7	78.2
LIVB	29	33.8	48.3	48.3	61.4
PHYS	19	56.8	58.9	66.3	71.6
PHEN	7	57.1	76.2	71.2	82.4
DEVI	5	20.0	20.0	20.0	28.0
OBJC	2	0.0	0.0	50.0	70.0
ACTI	1	0.0	20.0	0.0	100.0
<b>TOTAL</b>	<b>379</b>	<b>44.6</b>	<b>54.7</b>	<b>56.7</b>	<b>69.8</b>

the training data (the ontology and WALVIS), and their evaluation performance on Mantra\*. On all four largest groups in the training data, an average improvement of around 10% point in classification accuracy is observed, even though the largest group – disorders (DISO) – is with 149 samples four and a half times larger than the fourth-largest group ANAT (anatomy). We note that the numbers of mentions per semantic group in Mantra\* are too small to derive clear conclusions.

### 5.3. Error Analysis

We manually reviewed the mispredictions made by our optimal model. Due to the sometimes noisy and at some points extremely branched structure of the UMLS, seemingly small differences between prediction and gold label are scored incorrect. For example, the mention *advies* (advice) is linked to *voorlichting en advies* (counseling-C0010210) by our optimal model. However, in Mantra GSC, its gold label is given as *adviseren* (advice-C0150600). The prediction is called correct by the 1-distance metric, since a RN (‘Relation Narrow’) exists between the two concepts in the UMLS.

Sometimes, a mention is linked to an on surface form-level similar but semantically slightly different concept from the ontology. For example, mention *cannabis* is linked to the plant genus *cannabis* (C0936079), while its gold label in Mantra GSC is

<sup>11</sup><https://github.com/mi-erasmusmc/DutchClinicalCorpora>

<sup>12</sup>[https://anonymous.4open.science/r/biomedical\\_entity\\_linking-FCB4/ontology-browser/](https://anonymous.4open.science/r/biomedical_entity_linking-FCB4/ontology-browser/)

Table 4: Top 5 most found named entities and corresponding linked concepts from the patient-support forum data from [kanker.nl](https://www.kanker.nl). Translated from Dutch for the reader’s convenience.

#	NAMED ENTITY	LINKED CONCEPT
4356	CANCER	PRIMARY MALIGNANT NEOPLASM
4240	CHEMO	CHEMO-IMMUNOTHERAPY
3043	SURGERY	OPERATIVE SURGICAL PROCEDURES
3034	THERAPY	MILIEU THERAPY
2287	TUMOR	NEOPLASMS

the drug *cannabis* (C0678449). Also, the mention *pijnlijke rug* (sore back) is linked to *pijnlijke rug* (sore back-C0863105), but labeled as *rugpijn* (back pain-C0004604). This indicates that our accuracies are perhaps an under-estimation of the actual effectiveness of the entity linking – a finding that was also observed in previous work (Dirkson et al., 2023).

We further observe a high focus on surface form by our model. For example, mention *oren* (ears) is linked to *ren* (running-C0022646) instead of gold label *oor* (ear-C0013443). Moreover, mentions in all capitals, are often linked to a concept in all capitals, sometimes to a concept that is on surface form and semantical meaning very different. For example, mention *SOMATOTYPE* is linked to *DOPAMINERGIC AGENTS* (C0013036), while the surface form of its gold label is exactly similar to the mention but lower cased: *somatotype* (C0037669). Lower casing all terms in the ontology and newly seen mentions could help, but by doing so some information is lost, for example in abbreviations (‘pos’ is commonly used for ‘positive’, whereas ‘POS’ could mean ‘Polycystic Ovary Syndrome’). While a context-aware model like KRISBERT could reduce reliance on surface form, we did not implement such a model due to the lack of large, publicly available Dutch medical literature that is required for training (Zhang et al., 2022).

#### 5.4. Case Study

On the unlabeled data from <https://www.kanker.nl>, our finetuned BEL-model disagreed with the base model in the linking of 77.7% (286,654) of the found named entities. This indicates that self-alignment pre-training and finetuning has a substantial effect on the model behaviour. To get an impression of the model quality, we manually graded 100 randomly sampled mentions. We found that 42 mentions were errors in the named entity recognition step. Of the 58 correct entity mentions, 20 entities (34%) are linked to a wrong concept, another 20 mentions (34%) are linked to a related

concept and the remaining 18 (31%) are linked correctly. The grading can be found on Github.

The most commonly found named entities, such as *kanker* (cancer), *chemo* (chemo), and *operatie* (surgery), seem likely to appear frequently on a support forum for cancer patients (Table 4). While the named entities are simpler terms compared to their linked concepts, for the majority they seem to be words that are also practiced by medical professionals. The main difference between layman talk and medical jargon is probably to be found in the context the words are used in. If we look at the semantic groups of the linked entities, we see that Disorders – the largest group in the ontology – is also the most used semantic group on the forum, but with 34% much less proportionally than on Wikipedia (50% in WALVIS) and medical literature (39% in Mantra). Procedures on the other hand, are much more prevalent on the forum (23%) than on Wikipedia (5%) according to our model.

## 6. Conclusion

To the best of our knowledge, our work is the first to introduce an evaluated biomedical entity linking model in the Dutch language. We also present a method for automatically generating a weakly labeled biomedical entity linking dataset in any preferred target language, by combining the data from a biomedical ontology, Wikidata and Wikipedia pages. Using this method, we introduce the first – weakly labeled – Dutch biomedical entity linking corpus: WALVIS. We trained a BEL model by self alignment pretraining on the MedRoBERTa.nl, followed by fine-tuning on WALVIS\*. With around 70% 1-distance accuracy on the external evaluation set Mantra\*, we achieve a substantial improvement over the base model. This was achieved with a relatively small fine-tuning corpus. A case study on a collection of patient-written texts showed that the main source of error remains to be the named entity recognition step. Manual evaluation of small sample indicates that of the correctly extracted entities, our model links around 65% to a correct or closely related concept in the ontology.

In our evaluation on Mantra, we observe that our model relies heavily on surface form, which is for example observed by the erroneous linking of upper case mentions to upper case concepts that are otherwise very dissimilar. A context-aware model could further improve performance.

In conclusion, our biomedical entity linking model can be used for higher-level analysis of patient-oriented text data in Dutch. In future work, a larger corpus for fine-tuning could further improve the model’s performance. A larger corpus could be created by automatically translating the English Wikipedia pages, which are not only larger in num-

ber, but also contain more words per article.

## 7. Bibliographical References

- Rico Angell, Nicholas Monath, Sunil Mohan, Nishant Yadav, and Andrew McCallum. 2021. [Clustering-based Inference for Biomedical Entity Linking](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2598–2608, Online. Association for Computational Linguistics.
- Marco Basaldella, Fangyu Liu, Ehsan Shareghi, and Nigel Collier. 2020. [COMETA: A Corpus for Medical Entity Linking in the Social Media](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3122–3137, Online. Association for Computational Linguistics.
- Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl\_1):D267–D270.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT Pre-training of Deep Bidirectional Transformers for Language Understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Anne Dirkson, Suzan Verberne, Gerard Van Oortmerssen, Hans Gelderblom, and Wessel Kraaij. 2023. How do others cope? Extracting coping strategies for adverse drug events from social media. *Journal of Biomedical Informatics*, 139:104228.
- Jennifer D’Souza and Vincent Ng. 2015. Sieve-based entity linking for the biomedical domain. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 297–302.
- Evan French and Bridget T McInnes. 2022. An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics*, page 104252.
- Samuele Garda, Leon Weber-Genzel, Robert Martin, and Ulf Leser. 2023. BELB: a biomedical entity linking benchmark. *Bioinformatics*, 39(11):btad698.

- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M Van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, 22(5):948–956.
- Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, et al. 2021. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. *Artificial intelligence in medicine*, 117:102083.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.
- Sunwon Lee, Donghyeon Kim, Kyubum Lee, Jaehoon Choi, Seongsoon Kim, Minji Jeon, Sangrak Lim, Donghee Choi, Sunkyu Kim, Aik-Choon Tan, et al. 2016. BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PloS one*, 11(10):e0164680.
- Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. [Self-Alignment Pretraining for Biomedical Entity Representations](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.
- Daniel Loureiro and Alípio Mário Jorge. 2020. Medlinker: Medical entity linking with neural representations and dictionary matching. In *European Conference on Information Retrieval*, pages 230–237. Springer.
- Bridget T McInnes, Ted Pedersen, and Serguei VS Pakhomov. 2009. UMLS-Interface and UMLS-Similarity: open source software for measuring paths and semantic similarity. In *AMIA annual symposium proceedings*, volume 2009, page 431. American Medical Informatics Association.
- GC Miller and H Britt. 1995. A new drug classification for computer systems: the ATC extension code. *International journal of bio-medical computing*, 40(2):121–124.
- Stuart J Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. 2011. Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4):441–448.
- Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Lee M Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana K Savova. 2013. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. *CLEF (working notes)*, 1179.
- Tom M Seinen, Jan A Kors, Erik M van Mulligen, Egill Fridgeirsson, and Peter R Rijnbeek. 2023. [The added value of text from Dutch general practitioner notes in predictive modeling](#). *Journal of the American Medical Informatics Association*, 30(12):1973–1984.
- Tom M Seinen, Jan A Kors, Erik M van Mulligen, and Peter R Rijnbeek. 2024. Annotation-preserving machine translation of English corpora to validate Dutch clinical concept extraction tools. *medRxiv*, pages 2024–03.
- Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. [Biomedical Entity Representations with Synonym Marginalization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650, Online. Association for Computational Linguistics.
- Shogo Ujiie, Hayate Iso, and Eiji Aramaki. 2021. Biomedical Entity Linking with Contrastive Context Matching. *arXiv preprint arXiv:2106.07583*.
- Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P Rosé. 2021. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *Journal of biomedical informatics*, 121:103880.
- Stella Verkijk and Piek Vossen. 2021. Medroberta.nl: a language model for Dutch electronic health records. *Computational Linguistics in the Netherlands Journal*, 11:141–159.
- Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledge base. *Communications of the ACM*, 57(10):78–85.
- Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. 2019. Multi-similarity loss with general pair weighting for

deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5022–5030.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Hongyi Yuan, Zheng Yuan, and Sheng Yu. 2022. Generative biomedical entity linking via knowledge base-guided pre-training and synonyms-aware fine-tuning. *arXiv preprint arXiv:2204.05164*.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. [Knowledge-Rich Self-Supervision for Biomedical Entity Linking](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 868–880, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

# Unveiling Voices: Identification of Concerns in a Social Media Breast Cancer Cohort via Natural Language Processing

Swati Rajwal<sup>1\*</sup>, Avinash Kumar Pandey<sup>2\*</sup>, Zhishuo Han<sup>2</sup>, Abeed Sarker<sup>1</sup>

<sup>1</sup>Dept. of Biomedical Informatics, Emory University, <sup>2</sup>Goizueta Business School, Emory University  
{swati.rajwal, avinash.kumar.pandey, zhishuo.han, abeed.sarker}@emory.edu

\* Shared First Author.

## Abstract

We leveraged a dataset of ~1.5 million Twitter (now X) posts to develop a framework for analyzing breast cancer (BC) patients' concerns and possible reasons for treatment discontinuation. Our primary objectives were threefold: (1) to curate and collect data from a BC cohort; (2) to identify topics related to uncertainty/concerns in BC-related posts; and (3) to conduct a sentiment intensity analysis of posts to identify and analyze negatively polarized posts. RoBERTa outperformed other models with a micro-averaged  $F_1$  score of 0.894 and a macro-averaged  $F_1$  score of 0.853 for (1). For (2), we used GPT-4 and BERTopic, and qualitatively analyzed posts under relevant topics. For (3), sentiment intensity analysis of posts followed by qualitative analyses shed light on potential reasons behind treatment discontinuation. Our work demonstrates the utility of social media mining to discover BC patient concerns. Information derived from the cohort data may help design strategies in the future for increasing treatment compliance.

**Keywords:** Natural Language Processing, Breast Cancer, Social Media, Concerns, Topic Modeling

## 1. Introduction

### 1.1. Background

In 2020, there were 1,603,844 new cases of breast cancer (BC) and 602,347 died of BC in the United States as per the [Centers for Disease Control and Prevention \(2021\)](#). While advances in treatment have improved survival rates, a critical challenge persists in the continuity of long-term therapies which are crucial for reducing the risk of cancer recurrence. Unfortunately, many patients discontinue their treatments prematurely which is often linked to adverse patient-centered outcomes (PCOs). PCOs encompass a range of patient-specific experiences and even extend to broader socio-economic concerns related to treatments that are inherently difficult to measure and are underrepresented in electronic health records (EHRs). Motivated by this, our objective is to verify if social media BC cohort contains information about patient concerns, sentiments, and potential reasons for treatment non-compliance or discontinuation.

### 1.2. Related Work

Social media data has long been used for a range of tasks such as sentiment and opinion mining ([Pak and Paroubek, 2010](#); [Ananth et al., 2017](#)), medication-/drug-related information analysis ([Klein et al., 2024](#); [Nikfarjam et al., 2019](#); [Weissenbacher et al., 2021](#)), mental health-related research ([Amir et al., 2019](#); [Le Glaz et al., 2021](#)), substance use and recovery ([Kepner et al., 2022](#); [Balsamo et al., 2023](#); [Yang et al., 2023](#)), public health ([Antonius and Rich, 2013](#)) and many others. Additionally, researchers are actively using

social media platforms to discover targeted cohorts ([Krauss et al., 2015](#); [Sarker et al., 2017](#); [Al-Garadi et al., 2020](#)). Studies have shown that people often use social media to share their health-related experiences, including for BC ([Attai et al., 2015](#); [Nzali et al., 2017](#); [Sarker, 2017](#)). Thus, social media is a promising resource for capturing patient experiences and sentiments, or PCOs, provided the target cohort is accurately identified. In this work, we aim to utilize a dataset collected in prior work by [Al-Garadi et al. \(2020\)](#). The dataset contains  $N = 1,454,638$  tweets from 583,962 unique users. We developed a BC self-report identification system utilizing supervised machine learning models and RoBERTa ([Liu et al., 2019](#)). We used data from the cohort to identify BC-related uncertainty or concern topics. We performed a sentiment intensity analysis of patients who voiced dissatisfaction and identification of treatment discontinuation in the self-reported posts category.

## 2. Methodology

### 2.1. Self-Report Classifiers

We used manually-annotated data (Train=3513, Dev=302, Test=1204) to compare decision tree, logistic regression, random forest, naïve Bayes, and RoBERTa ([Liu et al., 2019](#)) for the task of BC self-report classification.

### 2.2. GPT-4 & Topic Modelling Framework

We used GPT-4 to generate an initial list of 20 seed words ([A.1](#)) related to uncertainty/fear. GPT-4's suggestions were often formal and not popu-

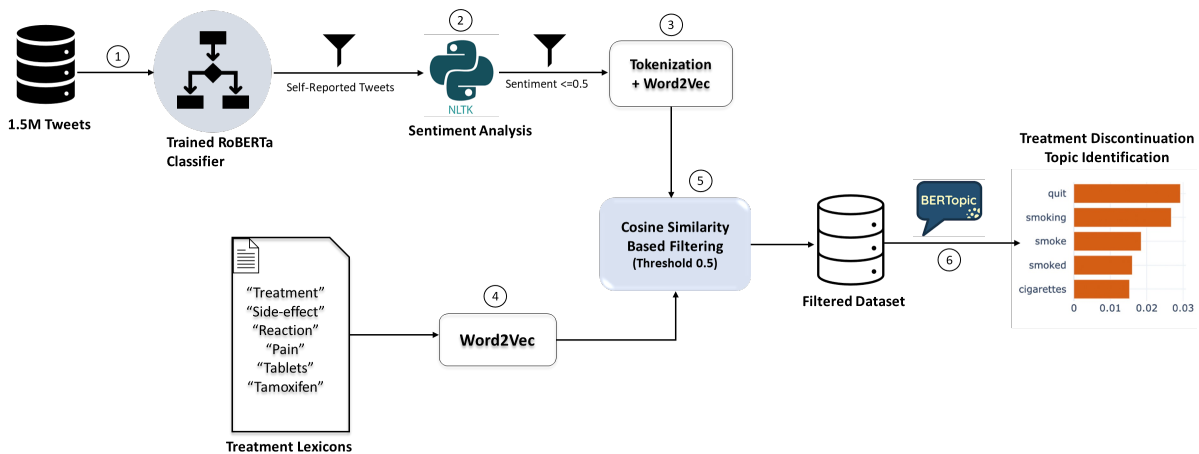


Figure 1: NLP pipeline for treatment discontinuation topic identification in a breast cancer Twitter cohort.

larly used by people to express their fear. Therefore, we utilized the Word2Vec word embedding model (Mikolov et al., 2013), particularly leveraging the GoogleNews-vectors-negative300 pretrained model<sup>1</sup> to identify similar keywords. The Word2Vec model’s output, after applying cosine similarity, helped us in curating our final list of keywords (A.2). Finally, we applied BERT topic modeling (Grootendorst, 2022) to all posts containing these terms in order to identify the key issues that are of concern to BC patients.

### 2.3. Sentiment Analysis Framework

We calculated the sentiment score for each Tweet using NLTK’s Sentiment Analyzer.<sup>2</sup> Then we applied BERTopic to a corpus of posts that had negative sentiment (score  $\leq 0.5$ ) and contained specific terms indicative of treatments, side effects, and symptoms associated with BC (A.3). The lexicon was derived from a PubMed paper on BC symptoms (Koo et al., 2017).

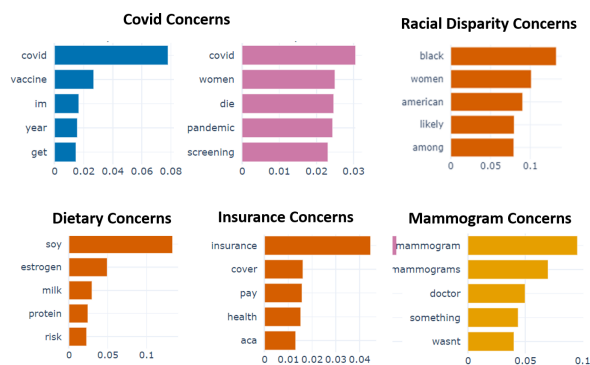


Figure 2: Key concerns among breast cancer patients discovered via topic modeling.

<sup>1</sup> [huggingface.co/fse/word2vec-google-news-300](https://huggingface.co/fse/word2vec-google-news-300)

<sup>2</sup> [nltk.org/howto/sentiment.html](https://nltk.org/howto/sentiment.html)

## 3. Results and Discussion

### 3.1. Classifier Evaluation & Results

RoBERTa significantly outperformed all models with a micro-averaged  $F_1$  score of 0.894, macro-averaged  $F_1$  score of 0.853 and, lowest log loss of 0.332 (Table 1, Appx. A.4). In the  $\sim 1.5$  million dataset, the best-performing (RoBERTa) classifier identified 154,571 posts as self-reported BC. Such posts were further used to identify topics for treatment discontinuation as shown in Figure 1.

### 3.2. Breast Cancer Concerns

BERTopic revealed insightful topics associated with the public’s concerns. In Fig. 2, the Y-axis shows topics (cluster of terms) that have been identified in the text. X-axis represents the scores/weights of each term within a specific topic. Each bar’s length indicates the strength of association between the term and its respective topic. The different colors for each topic are for visual clarity only. Below are the key issues associated with each topic:

- **Impact of COVID on Screenings:** A major topic cluster was around the impact of COVID-19 on BC patients. Concerns were raised about seeking medical help during the pandemic which could lead to delayed diagnoses and an increase in mortality rates. These findings are in line with a report by the CDC (2023).
- **Mammogram Anxiety:** Anxiety associated with mammogram screening was another notable topic. Discussions revealed anxiety towards mammogram screenings, exacerbated by the pandemic and fear of discovering BC.
- **Insurance Issues:** Posts (Table A.1) reflected frustration over the lack of coverage and the financial burden placed on patients, highlighting the systemic barriers to accessing care.



Model	Hyperparameter	F <sub>1</sub> micro	F <sub>1</sub> macro	F <sub>2</sub> micro	F <sub>2</sub> macro	Log loss
Decision Tree	criterion='gini', max_depth=10	0.778	0.608	0.778	0.596	0.734
Logistic Reg.	C=10, penalty='l2'	0.772	0.576	0.772	0.570	0.464
Naïve Bayes	alpha=0.1	0.745	0.427	0.745	0.468	0.568
Random forest	n_estimators=50	0.752	0.476	0.752	0.498	0.652
<b>RoBERTa</b>	epochs=20, batch_size=16	<b>0.894</b>	<b>0.853</b>	<b>0.894</b>	<b>0.841</b>	<b>0.332</b>

Table 1: Self-reported breast cancer tweet classification results across multiple evaluation metrics.

- **Soy Consumption:** Dietary habits and its link to BC is discussed. The chatter pointed to confusion regarding the consumption of soy and its alleged estrogenic effects.
- **Racial Disparity in Mortality:** The conversations highlighted concerns around the disproportionate impact of BC on African-American women including higher mortality rates.

### 3.3. Sentiment Analysis and Treatment Discontinuation

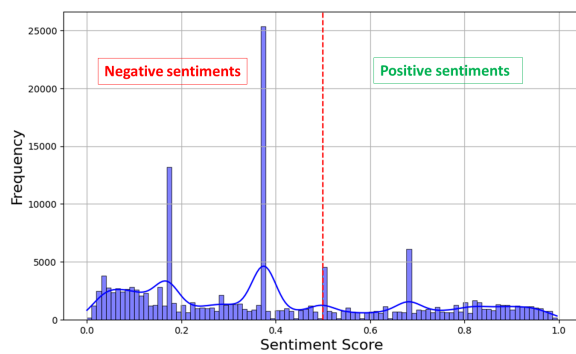


Figure 3: Sentiment Analysis of Breast Cancer Tweets via RoBERTa reveals more negative posts.

Sentiment analysis of self-reported BC posts revealed a pronounced skew towards negative sentiments, with the majority of sentiment scores falling at or below the 0.5 threshold (Figure 3). This trend suggests that the conversations are predominantly centered around the challenges faced by individuals, potentially reflecting the adverse effects of BC treatments. The BERTopic analysis of negative sentiment posts using keywords related to the side effects and treatment experiences of BC patients highlighted following main themes potentially linked to treatment discontinuation (Figure 4):

- **Hair/Baldness:** Concerns over hair loss reveal the psychological impact and social implications of chemotherapy-induced alopecia.
- **Smoking Restrictions:** Discussion around "quit," "smoking," suggests that patients might face difficulty adhering to treatment protocols that necessitate smoking cessation.

- **Insurance Issues:** High frequency of words like "insurance," "aca," "cost," and "bill" indicate that financial burdens and insurance coverage limitations are significant barriers.
- **Bowel/Colon Issues:** References to terms such as "colon," "colonoscopy," "bowel," "colorectal," & "polyps" suggest gastrointestinal side effects or procedures related to treatment may be intolerable for some patients.
- **Sexual Abuse Concerns:** Presence of terms like "sexual," "abuse," "rape," "assault," and "abuser" may reflect traumatic personal histories that intersect with treatment experiences.
- **Mental/Emotional Health:** A significant number of mentions of "cry," "crying," "emotional," "emotions," and "cried" highlight the emotional and psychological toll of BC and treatment.

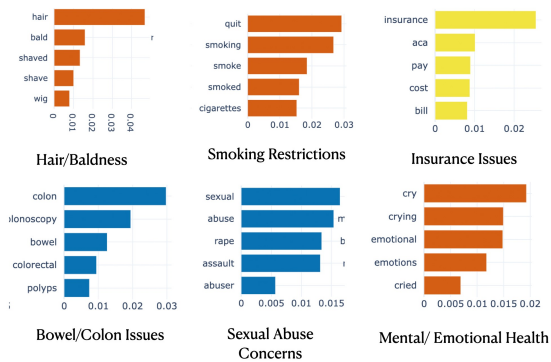


Figure 4: Treatment discontinuation topics from negative sentiment self-reported posts.

### 3.4. Post Classification Analyses

#### 3.4.1. Error Analysis

Lack of context, ambiguous references, and the use of informal language were the primary reasons for classification errors. For instance, Tweet 1 in A.6 expresses fear and mentions a family history of BC without directly stating a diagnosis. This poses challenges for the model to accurately classify such posts as self-reports. Moreover, the context in which BC is discussed varies widely. For

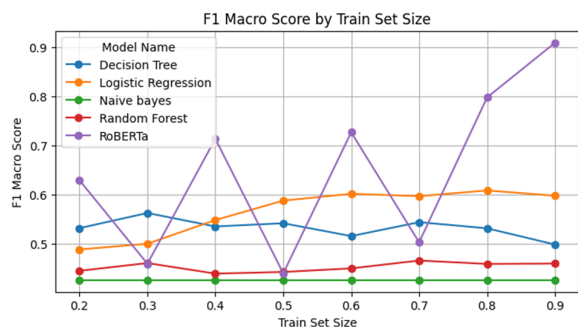


Figure 5:  $F_1$ -score at different training set sizes for different classifiers.

example, Tweet 2 (A.6) combines personal success with a history of overcoming cancer, and Tweet 3 (A.6) talks about a writing project, neither of which may align with the classifier’s training on more direct reports of BC. In case of false positives, we found that the model may be overly sensitive to certain keywords typically associated with personal experiences of BC, such as "battle with BC," "treatment," "therapy," and mentions of specific treatments like "Trastuzumab" or "radio therapy." The model also often misinterprets the sharing of news articles (Tweets in A.5) as personal reports of BC. Lastly, the model sometimes misclassifies supportive messages and discussions around BC as self reports. These findings show areas of possible improvement for the cohort creation step.

### 3.4.2. Classification Performance at Different Training Data Sizes

To assess if the cohort creation process (i.e., self-report detection for BC) can be improved in the future by increasing the training set size, we conducted training set size vs. performance experiments. As depicted in Fig. 5 and 6, RoBERTa shows a continuously increasing trend in  $F_1$  score and decreasing log loss as training size increases. This suggests that further improvement in the cohort creation process is possible by annotating more data, which can enable us to more accurately detect the BC cohort in the future. Though small, the logistic regression model also shows gradual improvement with an increase in training size.

### 3.5. Limitations and Future Directions

Ideally, for comprehensive longitudinal sentiment analysis, a sufficient number of posts for each unique user ID is required. However, due to the scarcity of posts per user in our dataset, we performed a consolidated sentiment analysis. Additionally, in topic modeling, we identified discussions of other health-related experiences (such as COVID-19) with BC scenarios. This overlap affected clas-

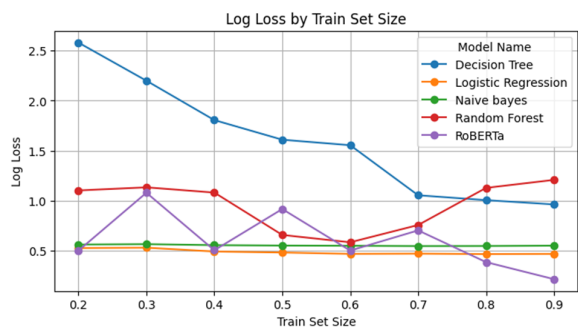


Figure 6: Log loss at different training set sizes for different classifiers.

sifier’s precision in isolating BC-specific conversations from other health concerns. These issues could be potentially mitigated by conducting experiments on a larger dataset containing posts from multiple social media sources rather than one. In the future, it will be interesting to study how current framework performs similar tasks while using data from various other social media platforms such as Reddit and Facebook, and from other time periods. It is also important to collect long-term data from this cohort to potentially discover temporal trends.

## 4. Conclusion

We investigated the potential use of Twitter (X) dataset for unveiling concerns among a cohort of BC patients. Specifically, we focused on (1) utilizing ML and transformer-based models for training an automatic classifier, (2) identifying BC-related concerns, and (3) performing sentiment analysis on self-reported posts and identifying potential reasons for treatment discontinuation of patients. Our experimental results highlight RoBERTa as the best-performing model for cohort identification. Our topic modeling framework of BC patient discussions reveals that concerns extend beyond traditional PCOs based solely on treatment side effects. Patients express a range of issues, including anxiety related to diagnostic mammogram procedures, barriers to screening due to COVID-19 pandemic, challenges with insurance coverage, and significant socio-emotional distress, such as racial discrimination and concerns about sexual abuse. Our findings reveal broader systemic, medical, and social challenges that need to be addressed for targeted public health messaging and inclusive community support aimed at alleviating fears and ensuring equitable access to healthcare resources.

## 5. Data and Code Availability

Code is openly available here: [github.com/swati-rajwal/BreastCancer\\_tweets\\_project](https://github.com/swati-rajwal/BreastCancer_tweets_project).

## 6. Bibliographical References

- Mohammed Ali Al-Garadi, Yuan-Chi Yang, Sahithi Lakamana, Jie Lin, Sabrina Li, Angel Xie, Whitney Hogg-Bremer, Mylin Torres, Imon Banerjee, and Abeed Sarker. 2020. Automatic breast cancer cohort detection from social media for studying factors affecting patient-centered outcomes. In *Artificial Intelligence in Medicine*, pages 100–110, Cham. Springer International Publishing.
- Silvio Amir, Mark Dredze, and John W. Ayers. 2019. [Mental health surveillance over social media with digital cohorts](#).
- SVSolai Ananth, Chandu Pmss, PG Scholar, and Assistant Professor. 2017. Live twitter knowledge as a corpus for sentiment analysis and opinion mining. *Global Journal of Pure and Applied Mathematics*, 13.
- Nicky Antonius and L. Rich. 2013. [Discovering collection and analysis techniques for social media to improve public safety](#). *The International Technology Management Review*, 3.
- Deanna J. Attai, Michael S. Cowher, Mohammed Al-Hamadani, Jody M. Schoger, Alicia C. Staley, and Jeffrey Landercasper. 2015. [Twitter social media is an effective tool for breast cancer patient education and support: patient-reported outcomes by survey](#).
- Duilio Balsamo, Paolo Bajardi, Gianmarco De Francisci Morales, Corrado Monti, and Rossano Schifanella. 2023. [The pursuit of peer support for opioid use recovery on reddit](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 17.
- CDC. Cancer and COVID-19 | CDC — cdc.gov. <https://www.cdc.gov/cancer/dcpc/about/covid-19.htm>. [Accessed 03-11-2024].
- Centers for Disease Control and Prevention. 2021. [Uscs data visualizations](#).
- Maarten Grootendorst. 2022. [Bertopic: Neural topic modeling with a class-based tf-idf procedure](#). *arXiv preprint arXiv:2203.05794*.
- Wayne Kepner, Meredith C. Meacham, and Alicia L. Nobles. 2022. [Types and sources of stigma on opioid use treatment and recovery communities on reddit](#). *Substance Use and Misuse*, 57.
- Ari Z Klein, Juan M Banda, Yuting Guo, Ana Lucia Schmidt, Dongfang Xu, Ivan Flores Amaro, Raul Rodriguez-Esteban, Abeed Sarker, and Graciela Gonzalez-Hernandez. 2024. [Overview of the 8th Social Media Mining for Health Applications \(SMM4H\) shared tasks at the AMIA 2023 Annual Symposium](#). *Journal of the American Medical Informatics Association*, page ocae010.
- Minjoung Monica Koo, Christian von Wagner, Gary A. Abel, Sean McPhail, Greg P. Rubin, and Georgios Lyratzopoulos. 2017. [Typical and atypical presenting symptoms of breast cancer and their associations with diagnostic intervals: Evidence from a national audit of cancer diagnosis](#). *Cancer Epidemiology*, 48:140–146.
- Melissa J. Krauss, Shaina J. Sowles, Megan Moreno, Kidist Zewdie, Richard A. Grucza, Laura J. Bierut, and Patricia A. Cavazos-Rehg. 2015. [Hookah-related twitter chatter: A content analysis](#). *Preventing Chronic Disease*, 12.
- Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan DeVyllder, Michel Walter, Sofian Berrouguet, and Christophe Lemey. 2021. [Machine learning and natural language processing in mental health: Systematic review](#). *J Med Internet Res*, 23(5):e15708.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#).
- Azadeh Nikfarjam, Julia D Ransohoff, Alison Callahan, Erik Jones, Brian Loew, Bernice Y Kwong, Kavita Y Sarin, and Nigam H Shah. 2019. [Early detection of adverse drug reactions in social health networks: A natural language processing pipeline for signal detection](#). *JMIR Public Health Surveill*, 5(2):e11264.
- Mike Donald Tapi Nzali, Sandra Bringay, Christian Lavergne, Caroline Mollevi, and Thomas Opitez. 2017. What patients can tell us: topic analysis for social media on breast cancer. *JMIR medical informatics*, 5(3):e7779.
- Alexander Pak and Patrick Paroubek. 2010. [Twitter as a corpus for sentiment analysis and opinion mining](#).
- Abeed Sarker, Pramod Chandrashekar, Arjun Magge, Haitao Cai, Ari Klein, and Graciela Gonzalez. 2017. [Discovering cohorts of pregnant](#)

women from social media for safety surveillance and analysis.

Graciela Gonzalez-Hernandez; Karen O'Connor; Guergana Savova; Abeed Sarker. 2017. [Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text](#). *Yearbook of Medical Informatics*, 26(01):214–227. Publisher: Georg Thieme Verlag KG.

Davy Weissenbacher, Suyu Ge, Ari Klein, Karen O'Connor, Robert Gross, Sean Hennessy, and Graciela Gonzalez-Hernandez. 2021. [Active neural networks to detect mentions of changes to medication treatment in social media](#). *Journal of the American Medical Informatics Association*, 28(12):2551–2561.

Yuan-Chi Yang, Mohammed Ali Al-Garadi, Jennifer S. Love, Hannah L. F. Cooper, Jeanmarie Perrone, and Abeed Sarker. 2023. [Can accurate demographic information about people who use prescription medications nonmedically be derived from twitter?](#) *Proceedings of the National Academy of Sciences*, 120(8):e2207391120.

## A. Appendices

### A.1. GPT-4 Generated Seed Words

"uncertain", "doubtful", "ambiguous", "hesitant", "perplexed", "confused", "questionable", "unsure", "indecisive", "bewildered", "vague", "puzzled", "skeptical", "inconclusive", "wavering", "distrustful", "baffled", "ambivalent", "hazy", "tentative"

### A.2. Final List of Fear-related Keywords

"scared", "nervous", "worried", "suspicion", "uncertain", "reluctant", "confused", "doubtful", "unsure", "skeptical", "awkward", "insurance", "soy", "covid", "black", "bra", "concerned"

### A.3. List of Treatment Keywords

"treatment", "medication", "medicine", "med", "tablets", "side effect", "reaction", "drug", "tamoxifen", "chemo", "mental", "emotion", "lump", "pain", "sleep", "docetaxel", "oncologist", "doc", "stop"

### A.4. Hyperparameter search space

For the *Naïve Bayes classifier*, 'alpha' range = [0.01, 0.1, 1, 10, 100]. For *Decision Tree model*, we tested the 'max\_depth' = [None, 3, 5, 10, 15, 20]. For the *Random Forest model*, 'n\_estimators' = [10, 50, 100] and 'max\_depth' = [None, 5, 10] were tested. For the *Logistic Regression model*, the 'C' (inverse

of regularization strength) range = [0.01, 0.1, 1, 10], and the penalties = ['l1', 'l2'].

### A.5. False positive Examples

**Tweet 1:** "Cam's Corner her Battle with Stage 3 Breast Cancer"

**Tweet 2:** "Daily Mirror - Breast Cancer: Difficult Treatment Decisions."

**Tweet 3:** "Effect at One Year of Adjuvant Trastuzumab for HER2+ Breast Cancer Combined with Radiation or an Anthracycline on Left Ventricular Ejection Fraction"

**Tweet 4:** "@user No. Breast cancer reconstruction."

**Tweet 5:** "@user Breast cancer: hope the best of the best, and hope you can feel better"

**Tweet 6:** "<number> and held years of breast cancer for my mother in france . several surgeries, chemo and radio therapy. <repeat> cost her..."

**Tweet 7:** "they pulled out of utah now <user> is back. tonight on <user> at <number>– why a breast cancer survivor feels let down and why the charity says they never left. pic.twitter.com / <number>lkl<number>gj"

### A.6. False Negatives Examples

**Tweet 1:** "I just found a rather large, hurtful lump in my breast. As a woman, I dont care if its nothing at all, the thought still terrifies me. Breast cancer is prominent in my family. Iam only 24 yo"

**Tweet 2:** "I'd be honored to tell you amazing women more about my story of being the first black woman to own a tequila brand and overcoming pancreatic and breast cancer."

**Tweet 3:** "If you're wondering what I've been doing writing-wise lately: I've been creating a breast cancer journal."

**Tweet 4:** "I should NOT have to tell people, wear a mask because I have metastatic breast cancer. Give a damn about people like me who have immune issues. #resiliencechat"

**Tweet 5:** "@user I have medical tattoos, 2 tiny blues when i had radiation for breast cancer."

**Tweet 6:**"b<user> selfish stockpiling - i order shopping every week. order weds for delivery fri or sat. yesterday no slots available. we are pensioners, i am just finishing breast cancer treatment so not easy to get out for "big" shop..."

Topic	Tweet Examples
Covid-19 affected Breast Cancer Patients	<p>“my <b>fear of being locked up again</b> is one of two contributors to my refusal to seek medical help for possible <b>breast cancer</b>.”</p> <p>“<b>pandemic leads women to delay mammograms</b>; experts fear future rise in breast cancer deaths &lt;url&gt; via &lt;user&gt;</p> <p>“&lt;user&gt; always had a normal <b>mammogram</b>. this year she was <b>scared to go in due to covid</b>. she pushed it back a few months. when doctors found stage &lt;number&gt; breast cancer.”</p>
Anxiety associated with Mammogram Screening	<p>“thousands of women under &lt;age&gt; with breast cancer showing up on their <b>mammograms</b>. i think another <b>huge issue is lumps being fobbed off as hormonal changes</b> - natasha was only referred because she put her foot down”</p> <p>“i had this conversation with my mom in the summer when she was <b>scared to get her mammogram</b>. she went and <b>she did have breast cancer</b> that fortunately was caught early.”</p> <p>“same with <b>mammograms</b>. I am <b>so scared to get them</b> and i have &lt;allcaps&gt; to because my <b>family has a history of breast cancer</b>”</p> <p>“also, this isn’t doctors, it’s the insurance, but i’m pissed that i <b>can’t get a mammogram</b> when my granny literally has breast cancer. they’ll do an ultrasound <b>but not a mammogram</b>. the f**k is that?”</p>
Concerns regarding Insurance Policies	<p>“the <b>last thing a breastcancer patient needs</b> to worry about is <b>if their insurance will cover treatment</b>. unfortunately, this is the reality for most facing this disease. help us eliminate barriers to care”</p> <p>“I am &lt;number&gt; just found a lump in my breast, <b>got a referral for a mammogram</b> from my pcp, got a <b>letter from insurance</b> promptly &lt;allcaps&gt; <b>telling me ,not all of this will be covered</b> by us just fyi, now i have to worry. . . &amp; maybe have to pay out of pocket to find out”</p> <p>“so i,am mad because <b>my insurance doesn’t cover</b> a well women, <b>No exam nor a mammogram</b> till i,am &lt;number&gt; i mean <b>breast cancer can affect women of all ages</b> you don’t think late teens deserve to check on their health.”</p> <p>“im <b>supposed to have mammograms on the reg</b> bc of family history but <b>insurance barely or just doesnt pay for them</b>. im lucky to have insurance at all but I m terrified my broke ass is gonna get breast cancer; not be able to do shit about it”</p>
Confusions regarding dietary habits	<p>“&lt;user&gt; &lt;user&gt; i’d like to switch to <b>soy milk</b>. i’m worried that it’s been <b>linked to breast cancer</b> due to the <b>presence of estrogen</b>”</p> <p>“proponents say that <b>soy can help prevent</b> heart disease, <b>breast cancer</b>, and more. detractors worry that <b>soy might interfere with thyroid function</b> and block nutrient absorption. who’s right?”</p>
Racial Disparity Concerns	<p>“i visited the cdc website to learn about cancer disparities between black and white american women, and i was extremely shocked to learn about the development of <b>breast cancer</b> so present in <b>young black women</b> when it compares to <b>young white women</b>.”</p> <p>“<b>black women</b> have a &lt;number&gt;% <b>higher death rate</b> from <b>breast cancer</b>. we shouldn’t be surprised that communities with unequal health outcomes have unequal trust in vaccines.”</p> <p>“black women, specifically <b>young black women</b>, are more <b>susceptible to breast cancer</b>. Don’t be afraid to self-exam often!”</p>

Table A.1: Table showing sample tweets with different topics of concerns/uncertainty amongst BC patients

# Intent Detection and Entity Extraction from Biomedical Literature

Ankan Mullick<sup>1\*</sup>, Mukur Gupta<sup>2\*</sup>, Pawan Goyal<sup>1</sup>

<sup>1</sup>Computer Science and Engineering Department, IIT Kharagpur, India

<sup>2</sup>Computer Science Department, Columbia University, USA

ankanm@kgpian.iitkgp.ac.in, mukur.gupta@columbia.edu, pawang@cse.iitkgp.ac.in

## Abstract

Biomedical queries have become increasingly prevalent in web searches, reflecting the growing interest in accessing biomedical literature. Despite recent research on large-language models (LLMs) motivated by endeavors to attain generalized intelligence, their efficacy in replacing task and domain-specific natural language understanding approaches remains questionable. In this paper, we address this question by conducting a comprehensive empirical evaluation of intent detection and named entity recognition (NER) tasks from biomedical text. We show that Supervised Fine Tuned approaches are still relevant and more effective than general-purpose LLMs. Biomedical transformer models such as PubMedBERT can surpass ChatGPT on NER task with only 5 supervised examples.

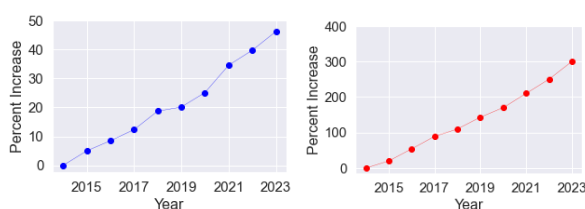
## 1. Introduction

Research on large-language models has skyrocketed in the post-ChatGPT era. Researchers are now aiming for generalized intelligence by increasing model size (Brown et al., 2020; Chowdhery et al., 2022; Hoffmann et al., 2022), expanding & rearranging pretraining data (Touvron et al., 2023a,b; Sarkar and Gupta, 2021) and incorporating human feedback (Ouyang et al., 2022; Dubois et al., 2023). It is shown that the adoption of GPT-4 (OpenAI, 2023) can potentially affect up to 80% of the U.S. workforce (Eloundou et al., 2023). These generalization reasoning demonstrations raise an important question for the research community - does this mark an end to the task and domain-specific natural language understanding approaches? While some research places LLMs as “General Purpose Technologies” (Eloundou et al., 2023; Zhang et al., 2023a) for solving a range of complicated tasks, we show that these models struggle to perform well on domain-specific complex tasks and specialized Supervised Fine-tuned (SFT) models are still needed to solve language understanding use-cases.

Over the past two decades, web searches have evolved dramatically transitioning from generic interfaces to more intent-specific and entity-aware systems capable of immediately displaying diverse multi-modal responses. Particularly, biomedical inquiries, spanning topics such as medical treatment, medical diagnosis, disease, etc. have seen a surge in popularity across search engines. Fig. 1 shows the increase in the percentage of Biomedical queries on Bing search and Google trends<sup>1</sup>.

\*Authors contributed equally

<sup>1</sup>Google trends data of last 10 years on five topics (Health, Medical Treatment, Medical diagnosis, Disease, Pharmaceutical drug) was gathered from Google Trends (<https://trends.google.com/trends/>)



(a) Google yearly Trend (b) Bing Query yearly

Figure 1: Biomedical query search Statistics

As large volumes of biomedical data continue to be generated every second on various online platforms the role of information retrieval systems in processing domain-specific texts becomes increasingly important. However, handling biomedical text data presents unique challenges, as the medical queries on search engines and online medical forums are often incomplete, do not follow a specific structure, and contain hard-to-interpret context-specific medical terminologies, as shown in Table 1. While recent research is centered around the development of general-purpose LLMs, that are shown to exhibit exceptional Common Sense Reasoning capabilities (Touvron et al., 2023b), we show that these models face challenges in transferring their performance to intricate biomedical domains. To this end, we focus on two crucial natural language understanding tasks of intent detection and named entity recognition from biomedical text.

For the past two decades, different directions of intent detection and corresponding entity extraction have been explored. (Sun et al., 2016; Wang et al., 2020; Mu et al., 2017b,a) demonstrate intent detection in the form of out-of-domain data detection. Other research works explore methods like few shot (Xia et al., 2021), zero-shot (Xia et al., 2018), and clustering frameworks (Mullick et al., 2022b). (Yani et al., 2022; Zhao et al., 2021; Fetahu et al.,

Biomedical Text	Intent
Pharmacokinetic properties of <b>abacavir</b> were not altered by the addition of either <b>lamivudine</b> or <b>zidovudine</b> or the combination of <b>lamivudine</b> and <b>zidovudine</b> .	<b>Drug</b>
<b>Canavan disease</b> , or <b>spongy degeneration of the brain</b> , is a severe <b>leukodystrophy</b> caused by <b>deficiency of aspartoacylase (ASPA)</b> .	<b>Disease</b>

Table 1: Intent & corresponding Entity (highlighted) examples from DDI and NCBI Disease datasets.

2022) explore entity recognition task in various settings. In the medical domain, Zhou et al. (2021) focuses on smart healthcare and (Galea et al., 2018; Giorgi and Bader, 2019; Lee et al., 2019) inspect transformer based models for biomedical literature. Mullick et al. (2023, 2022a); Mullick (2023b,a) aims at intent detection and entity extraction and Zhang et al. (2017) explore medical query intents by applying graph-based frameworks. (Mullick et al., 2024; Guha et al., 2021) work on domain specific entity and corresponding relation extraction. (Mullick et al., 2017b, 2016, 2018a,b, 2019, 2017a) aim at opinion-fact entity extraction.

There is no unified and exhaustive comparison of existing approaches with the recent LLMs for intent detection and entity extraction tasks across various datasets in biomedical literature. Our work differs from the prior research in two ways: we present a thorough empirical evaluation of the intent detection on three datasets and corresponding named entity extraction (NER) approaches on 27 unique entities covered in 5 biomedical datasets spanning across domains like drugs, diseases, chemicals, genetics and, human anatomy. We evaluate various supervised approaches (transformer-based, handcrafted features, etc.) and benchmark them against two widely used large language models in the biomedical domain. Our experiments reveal that the biomedical transformer-based PubMedBERT model outperforms few-shot prompted ChatGPT (Turbo 3.5) on 4 biomedical NER benchmarks with just 5 supervised examples. We make our code publicly available.<sup>2</sup>

## 2. Datasets

We show our comparative study on a variety of datasets, which are widely used as benchmarks in the biomedical domain. We use five different Named Entity Recognition datasets: JNLPBA (Collier and Kim, 2004), DDI (combining DDI-Drugbank and DDI-Medline) (Segura-Bedmar et al., 2013), BC5CDR (Smith et al., 2008), NCBI-Disease (Li

<sup>2</sup>[https://github.com/bioNLU-colling2024/biomed-NER-intent\\_detection](https://github.com/bioNLU-colling2024/biomed-NER-intent_detection)

Dataset	Entity Type	# Entities	#Train	#Test
JNLPBA	Gene & Protein	5	2000	404
DDI	Drug	4	714	112
BC5CDR	Chem & Disease	2	1000	500
NCBI-Disease	Disease	4	693	100
AnatEM	Anatomy	12	300	200

Table 2: Statistics of the NER datasets. We use the pre-defined train-test split as mentioned in the papers.

et al., 2016) and AnatEM (Ohta et al., 2012). Dataset statistics including the entity types, count, and train-test splits are outlined in Table 2. We use the pre-defined train-test divisions from the respective manuscripts.

Along with the two popular intent detection datasets - CMID (Chen et al., 2020) and KUAKE-QIC (part of the CBLUE (Zhang et al., 2021) benchmark), we combine the three of the above five NER datasets (JNLPBA, DDI, and NCBI-Disease) with respective intent labels (DDI for drugs, NCBI-Disease for disease and JNLPBA for Genetics) for intent classification task - termed as "Intent-Merged" dataset. Dataset statistics are summarized in Table 3.

CMID and KUAKE-QIC datasets, which are originally in Chinese, are translated to English using Google Translation API. For translation validation, a random sample of 400 translated (to English) examples of each dataset are validated manually by two Chinese experts (ALA Language Center Company) with HSK Level-3 proficiency. The human-validation shows 91.75% and 97.0% translation accuracy for CMID and KUAKE-QIC respectively. Hence, we use the translated English data along with their pre-defined intent labels for our experiments. The inter-annotator agreement is 0.89.

Dataset	#Train	#Test Size	#Intents
CMID	9558	2696	4
KUAKE-QIC	6931	1955	11
Intent-Merged	3905	909	3

Table 3: Statistics of Intent Detection datasets.

## 3. Experimental Settings

### 3.1. Intent Detection

Intent detection is a multi-class classification task where we evaluate the accuracy of instruction-tuned ChatGPT (gpt-3.5-turbo-instruct) against various SFT models on three English datasets: CMID, KUAKE-QIC, and Intent-Merged.

**1. Large Language Models:** To ensure consistency with prior works, we employ a  $k$ -shot prompt design, wherein  $k$  examples per class from the training set are used in the prompt. Given the

larger text sizes of the Intent-Merged dataset and the limited context window of LLMs, we use  $k = 1$  for all datasets. We note no significant performance improvement with increasing  $k$  for CMID and KUAKE-QIC datasets. Further details on the prompt template are included in the GitHub repository.

**2. Supervised Finetuned Models:** For SFT, we finetune - BERT (bert-base-uncased) (Devlin et al., 2018), RoBERTa (roberta-base) (Liu et al., 2019), PubMedBERT (Gu et al., 2021), FastText (Chen et al., 2020) and TextCNN (Kim, 2014).

The empirical evaluation is shown in Table 4.

Model	CMID	KUAKE-QIC	Intent-Merged	Mean
BERT	72.26	75.91	96.37	81.51
RoBERTa	<b>72.88</b>	<b>78.16</b>	<b>99.11</b>	<b>83.38</b>
PubMedBERT	72.70	76.88	97.90	82.49
Llama-2	51.11	42.50	39.54	44.38
ChatGPT	42.36	44.04	64.44	50.28
Fasttext	68.43	72.48	96.80	79.24
TextCNN	70.69	75.19	96.15	80.68

Table 4: Accuracy (in %) of intent classification tasks on three datasets.

All the SFT approaches consistently outperform the instruction-tuned ChatGPT. The poor performance of LLMs on the Intent-Merged dataset, which is quite easy for all the SFT approaches, reflects their deficiency in domain-specific knowledge within their general-purpose pretraining datasets. This also shows that models like FastText can outperform ChatGPT, given domain-specific finetuning. We note that transformer architectures give better performance on the translated corpus compared to FastText and TextCNN, which are shown to work well on Chinese data (Chen et al., 2020). RoBERTa gives the highest accuracy across overall mean and individual datasets.

## 3.2. Named Entity Recognition

For NER, we apply a strict match between the predicted entity class and the entity word boundaries and report strict F1-score (as in CoNLL shared task (Tjong Kim Sang and De Meulder, 2003)). We run all models 5 times with different random initialization and report micro-average F1-score along with standard deviations. We also report the overall mean for each approach. For a fair comparison, a maximum sequence length of 512 tokens was used for all models, hence the texts larger the token length were further broken into multiple texts.

### 3.2.1. Supervised Fine-Tuned Models

We thoroughly examine five different settings on the five biomedical datasets.

**Setting A:** Fine-tuned BERT and RoBERTa models are used (pre-trained on general English corpus) without domain pretraining.

**Setting B:** Transformer systems with continued pretraining on biomedical text. We fine-tune BioBERT, PubMedBERT, BioMed RoBERTa, and ClinicalBERT.

**Setting C:** LSTM and Convolutional Neural Networks (with/without CRF) are used to generate the word embeddings and softmax classifiers for tag prediction.

**Setting D:** Hand-crafted word level features with ML classifier: (i) POS tag (ii) shallow parsing features like chunk tag (iii) orthographic boolean features like all capital, is alphanumeric, etc. (iv) n-gram features, etc. We use the GENIA tagger<sup>3</sup> for POS and Chunk tag extraction. We apply XGBoost and a multi-label logistic regression model for NER tag prediction.

**Setting E:** We use state-of-the-art NER model BINDER (Zhang et al., 2023b) along with domain-specific (PubMedBERT and BioBERT) and RoBERTa encoders.

### 3.2.2. Large Language Models

We use instruction-tuned ChatGPT (gpt-3.5-turbo-instruct).

**Setting F:** We modify (Wang et al., 2023) for re-conditioning NER as a Tag generation problem. In addition to the prompt design proposed by (Wang et al., 2023), following (Zhang et al., 2023b) we also add a short description for the entity. Each prompt infers only a single entity tag. Hence, each text instance is passed multiple times for tagging all the entity types. We provide two examples from the train set in each prompt. To motivate both high recall and prevent hallucination in Entity identification, we specifically pick examples with the median number of entity tags in the training dataset.

The evaluation outcomes are shown in Table 5. We find:

**a) SFTs outperform LLMs:** We observe that all SFT approaches surpass ChatGPT by a big margin. Further, from Table 6, it's evident that PubMedBERT can easily outperform ChatGPT on most benchmarks with just five supervised examples.

**b) Transformer SFT Models:** i) PubMedBERT learns good embedding vectors due to the largest pretraining corpus. BINDER combined with PubMedBERT gives the best F1 score as it is able to leverage high-quality embeddings along with entity descriptions which pushes the similar entity tokens closer in the embedding space with a contrastive loss objective. (ii) LSTM/CNN-based neural embedding and traditional ML-based models - XGBoost and Logistic Regression perform poorly because they fail to capture contexts and do not leverage domain-specific pretraining.

<sup>3</sup><http://www.nactem.ac.uk/GENIA/tagger/>



Model/Dataset	DDI	JNLPBA	BC5CDR	NCBI Disease	AnatEM	Mean
BERT (A)	83.94 ± 0.17	72.60 ± 0.13	87.13 ± 0.24	77.44 ± 0.75	78.66 ± 0.35	79.95
RoBERTa (A)	87.13 ± 0.44	74.91 ± 0.11	89.50 ± 0.09	81.67 ± 0.51	81.90 ± 0.60	83.06
BioBERT (B)	88.06 ± 0.08	74.02 ± 0.40	90.19 ± 0.13	81.91 ± 0.80	83.43 ± 0.36	83.52
PubMedBERT (B)	88.84 ± 0.12	75.15 ± 0.06	90.77 ± 0.08	82.34 ± 0.16	84.21 ± 0.21	84.26
BioMed RoBERTa (B)	88.76 ± 0.31	75.14 ± 0.25	90.24 ± 0.18	82.13 ± 0.92	82.70 ± 0.17	83.80
Clinical BERT (B)	83.79 ± 0.22	72.54 ± 0.07	87.90 ± 0.17	76.34 ± 0.64	73.47 ± 0.53	78.80
LSTM (C)	73.00 ± 0.01	67.00 ± 0.01	79.01 ± 0.01	70.00 ± 0.01	74.01 ± 0.01	72.60
LSTM + CRF (C)	74.75 ± 0.04	70.67 ± 0.01	80.47 ± 0.02	73.13 ± 0.02	77.39 ± 0.05	75.23
CNN (C)	73.07 ± 0.01	68.04 ± 0.01	80.00 ± 0.01	67.09 ± 0.01	72.08 ± 0.01	72.06
CNN + CRF (C)	73.29 ± 0.08	70.84 ± 0.11	81.27 ± 0.14	73.59 ± 0.09	75.15 ± 0.13	74.83
Logistic Regression (D)	78.63 ± 0.01	57.03 ± 0.01	78.20 ± 0.01	56.36 ± 0.01	65.48 ± 0.01	67.14
XGBoost (D)	73.55 ± 0.01	53.06 ± 0.01	67.86 ± 0.01	52.62 ± 0.01	59.91 ± 0.01	61.40
BINDER-BioBERT (E)	89.01 ± 0.01	76.63 ± 0.19	91.59 ± 0.09	<b>85.47</b> ± 0.36	86.71 ± 0.25	85.88
<b>BINDER-PubMedBERT (E)</b>	<b>89.12</b> ± 0.01	77.01 ± 0.01	<b>91.88</b> ± 0.01	85.25 ± 0.02	<b>86.95</b> ± 0.02	<b>86.04</b>
BINDER-RoBERTa (E)	87.98 ± 0.01	<b>77.08</b> ± 0.01	90.48 ± 0.03	84.62 ± 0.06	83.91 ± 0.05	84.81
ChatGPT (F)	42.94 ± 3.10	24.5 ± 1.89	44.68 ± 2.78	19.65 ± 1.21	2.92 ± 0.07	26.94

Table 5: Experiment results (Macro average F1-Scores and corresponding standard deviations) on different NER systems trained/finetuned and tested on 5 biomedical Datasets.

**c) Feature-based SFT Models:** (i) ML-based model performs better than CNN/LSTM embedding systems on the DDI dataset, implying that it might be possible to beat the performances on other datasets if the right feature set is selected, which is usually an expensive process. (ii) The range of performance (best F1 - worst F1) for NCBI-Disease corpus is highest, showing that there is a big difference between the selected feature set and the features captured by the neural models. (iii) The addition of a CRF prediction layer on CNN/LSTM improves the performance significantly.

**d) Dataset Quality:** In most of the cases, low F1 is observed on the entities having fewer examples in the training set. For example, entities "CompositeMention" and "Disease Class" show poor performance due to less number of samples in training data. We note that the tag generation problem is difficult for instruction-tuned LLMs. We also experiment with Llama-2 (7b) model<sup>4</sup> and observe that vanilla Llama-2-7b does not achieve good results as it was unable to follow the specified output structure and most of times, ended up hallucinating text. So, we omit vanilla Llama-2 results and will explore further in future.

### 3.2.3. Experimental Setup

We experiment on Tesla T4 16GB GPU, 6 Gbps clock cycle and GDDR5 memory. All experiments (entity extraction and intent detection) took ~60 minutes for training. We fine-tune the models for a maximum of 20 epochs with a learning rate of 5e-5 with AdamW optimizer and 10% warm-up steps. The batch size is 16. Additional details are included in the GitHub Repository.

<sup>4</sup><https://ai.meta.com/llama/>

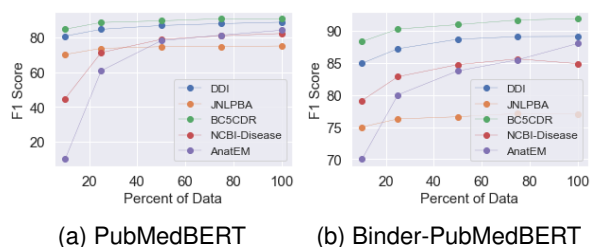


Figure 2: Ablation: Varying Training Size

## 4. Ablations

We study the relationship of SFT models with domain-specific finetuning data:

**Varying Training Data Size:** We vary the size of training data (10%, 25%, 50%, 75% and 100%), while keeping the test set constant and show the performances of PubMedBERT, Binder (PubMedBERT) models in Figure 2. We observe that, unlike raw PubMedBERT, Binder (PubMedBERT) attains a high performance with only 10% of training data. Transformer-based models can learn with very little training data and performance does not decrease much even with 25% training data. Due to domain pre-training, PubMedBERT learns with much fewer samples and saturates faster. This quick-learning behavior seems to be originating from transfer learning. However, LSTM and CNN models suffer from poor performances in low-data settings due to no pretraining (details in GitHub).

# Shots	DDI	JNLPBA	BC5CDR	NCBI	AnatEM
5	2.8 / 60.05	2.0 / 39.09	1.1 / 64.53	3.2 / 27.56	2.6 / 1.5
10	5.8 / 65.54	5.6 / 50.56	2.1 / 69.23	7.05 / 34.32	6.1 / 3.5
30	45.49 / 73.71	61.07 / 58.01	54.07 / 77.97	60.63 / 48.16	21.74 / 10.9
50	81.39 / 76.85	71.04 / 62.23	83.3 / 82.78	75.83 / 49.51	40.69 / 30.96
100	83.56 / 80.94	74.24 / 68.02	88.58 / 85.76	82.38 / 72.78	73.34 / 52.73

Table 6: BINDER-PubMedBERT / PubMedBERT F1-score in K-shot setting

**Few Shot:** In Table 6, we show the performances (F1 Score) of the Binder (PubMedBERT) and PubMedBERT models on a few shot settings with different numbers of training samples. PubMedBERT embeddings perform better in very low-resource setups (5, 10 shots). However, when training examples increase further (30 shots onwards), BINDER (PubMedBERT) outperforms PubMedBERT because of the Bi-Encoder architecture trained on contrastive learning objectives.

Error Type	Entity Text	Label	Prediction
Boundary	3-[(...)]ethynyl] pyridine	B-D,I-D	B-D,B-D
Entity type	heparinase III	B-D,I-D	B-G,I-G
Entity Miss	Hyaluronan lyase	B-D,I-D	O,O

Table 7: Errors by BINDER-PubMedBERT on entity "drug\_n" of DDI dataset. Following abbreviations are used - B-D: B-DRUG\_N, I-D: I-DRUG\_N, B-G: B-Group, I-G: I-GROUP

## 5. Error Analysis

We do a detailed analysis on errors as following:

A) Some errors are due to model failure like RoBERTa's failure to classify 52% of the "other" intents from the KUAKE-QIC dataset. For example, a query such as "I have a *cyst* in the *corner of my right eye* and it grows bigger and bigger." is classified wrongly as "diagnosis" intent but it is of "other" category.

B) Three types of errors are observed for entity extraction (examples from the DDI dataset are shown in Table 7).

C) Some models fail to identify the entity "drug\_n" which represents new or unapproved drugs so a periodic model update is required.

D) Relaxing entity-type error by considering exact F1-score instead of strict F1, we observe an uplift of 4.57% in mean F1.

## 6. Conclusion

The biomedical sector has matured significantly in the past few years. We show instead of relying on general-purpose LLMs, it is important to design an intent detection and entity extraction task for processing domain-specific texts. In this work, we show that fine-tuned RoBERTa and BINDER (PubMedBERT) can work efficiently to detect intents and extract named entities across various benchmark datasets in biomedical literature. In the future, we aim to extract intent and entity jointly as a relation tuple and inspect the performances of various cross-domain scenarios.

## 7. Limitations

Our dataset needs to be scaled up in terms of different languages, sizes, and intent labels which we aim to do in the near future. The approach needs to be updated as a single model for jointly extracting intents and entities for multilingual scenarios which we aim to do as a part of future work.

## 8. Ethical Concerns

We propose to release the algorithmic details and work on public datasets that neither reveal any personal sensitive information nor any toxic statement. So there are no ethical concerns in this work.

## 9. Bibliographical References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#).

Nan Chen, Xiangdong Su, Tongyang Liu, Qizhi Hao, and Ming Wei. 2020. A benchmark dataset and case study for chinese medical question intent classification. *BMC Medical Informatics and Decision Making*, 20.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei

- Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. [Palm: Scaling language modeling with pathways](#).
- Nigel Collier and Jin-Dong Kim. 2004. [Introduction to the bio-entity recognition task at JNLPBA](#). In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 73–78, Geneva, Switzerland. COLING.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. [AlpacaFarm: A simulation framework for methods that learn from human feedback](#).
- Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. [Gpts are gpts: An early look at the labor market impact potential of large language models](#). *arXiv preprint arXiv:2303.10130*.
- Besnik Fetahu, Anjie Fang, Oleg Rokhlenko, and Shervin Malmasi. 2022. [Dynamic gazetteer integration in multilingual models for cross-lingual and cross-domain named entity recognition](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2777–2790, Seattle, United States. Association for Computational Linguistics.
- Dieter Galea, Ivan Laponogov, and Kirill Veselkov. 2018. [Exploiting and assessing multi-source data for supervised biomedical named entity recognition](#). *Bioinformatics*, 34(14):2474–2482.
- John M Giorgi and Gary D Bader. 2019. [Towards reliable named entity recognition in the biomedical domain](#). *Bioinformatics*, 36(1):280–286.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. [Domain-specific language model pretraining for biomedical natural language processing](#). *ACM Transactions on Computing for Healthcare*, 3(1):1–23.
- Souradip Guha, Ankan Mullick, Jatin Agrawal, Swetarekha Ram, Samir Ghui, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2021. [Matscie: An automated tool for the generation of databases of methods and parameters used in the computational materials science literature](#). *Computational Materials Science (Comput. Mater. Sci.)*, 192:110325.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. [Training compute-optimal large language models](#).
- Yoon Kim. 2014. [Convolutional neural networks for sentence classification](#). *arXiv preprint arXiv:1408.5882*.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [BioBERT: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36(4):1234–1240.
- Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wieggers, and Zhiyong Lu. 2016. [BioCreative V CDR task corpus: a resource for chemical disease relation extraction](#). *Database (Oxford)*, 2016:baw068.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *arXiv preprint arXiv:1907.11692*.
- Xin Mu, Kai Ming Ting, and Zhi-Hua Zhou. 2017a. [Classification under streaming emerging new classes: A solution using completely-random trees](#). *IEEE Transactions on Knowledge and Data Engineering*, 29(8):1605–1618.
- Xin Mu, Feida Zhu, Juan Du, Ee-Peng Lim, and Zhi-Hua Zhou. 2017b. [Streaming classification with emerging new class by class matrix sketching](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.
- Ankan Mullick. 2023a. [Exploring multilingual intent dynamics and applications](#). *IJCAI Doctoral Consortium*.
- Ankan Mullick. 2023b. [Novel intent detection and active learning based classification \(student abstract\)](#). *arXiv e-prints*, pages arXiv–2304.

- Ankan Mullick, Akash Ghosh, G Sai Chaitanya, Samir Ghui, Tapas Nayak, Seung-Cheol Lee, Satadeep Bhattacharjee, and Pawan Goyal. 2024. Matscire: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction. *Computational Materials Science*, 233:112659.
- Ankan Mullick, Surjodoy Ghosh D, Shivam Maheshwari, Srotaswini Sahoo, Suman Kalyan Maity, and Pawan Goyal. 2018a. Identifying opinion and fact subcategories from the social web. In *Proceedings of the 2018 ACM International Conference on Supporting Group Work*, pages 145–149.
- Ankan Mullick, Pawan Goyal, and Niloy Ganguly. 2016. A graphical framework to detect and categorize diverse opinions from online news. In *Proceedings of the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 40–49.
- Ankan Mullick, Pawan Goyal, Niloy Ganguly, and Manish Gupta. 2017a. Extracting social lists from twitter. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 391–394.
- Ankan Mullick, Pawan Goyal, Niloy Ganguly, and Manish Gupta. 2018b. Harnessing twitter for answering opinion list queries. *IEEE Transactions on Computational Social Systems*, 5(4):1083–1095.
- Ankan Mullick, Shivam Maheshwari, Pawan Goyal, and Niloy Ganguly. 2017b. A generic opinion-fact classifier with application in understanding opinionatedness in various news section. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 827–828.
- Ankan Mullick, Ishani Mondal, Sourjyadip Ray, R Raghav, G Chaitanya, and Pawan Goyal. 2023. Intent identification and entity extraction for healthcare queries in indic languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836.
- Ankan Mullick, Abhilash Nandy, Manav Nitin Kapadnis, Sohan Patnaik, and R Raghav. 2022a. Fine-grained intent classification in the legal domain. *arXiv preprint arXiv:2205.03509*.
- Ankan Mullick, Sourav Pal, Projjal Chanda, Arijit Panigrahy, Anurag Bharadwaj, Siddhant Singh, and Tanmoy Dam. 2019. D-fj: Deep neural network based factuality judgment. *Technology*, 50:173.
- Ankan Mullick, Sukannya Purkayastha, Pawan Goyal, and Niloy Ganguly. 2022b. A framework to generate high-quality datapoints for multiple novel intent detection. *arXiv preprint arXiv:2205.02005*.
- Tomoko Ohta, Sampo Pyysalo, Jun’ichi Tsujii, and Sophia Ananiadou. 2012. [Open-domain anatomical entity mention detection](#). In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 27–36, Jeju Island, Korea. Association for Computational Linguistics.
- OpenAI. 2023. [Gpt-4 technical report](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#).
- Dipankar Sarkar and Mukur Gupta. 2021. Tso: Curriculum generation using continuous optimization. *CoRR*.
- Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. [SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts \(DDIExtraction 2013\)](#). In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 341–350, Atlanta, Georgia, USA. Association for Computational Linguistics.
- Larry Smith, Lorraine K Tanabe, Rie Johnson Nee Ando, Cheng-Ju Kuo, I-Fang Chung, Chun-Nan Hsu, Yu-Shi Lin, Roman Klinger, Christoph M Friedrich, Kuzman Ganchev, Manabu Torii, Hongfang Liu, Barry Haddow, Craig A Struble, Richard J Povinelli, Andreas Vlachos, William A Baumgartner, Jr, Lawrence Hunter, Bob Carpenter, Richard Tzong-Han Tsai, Hong-Jie Dai, Feng Liu, Yifei Chen, Chengjie Sun, Sophia Katrenko, Pieter Adriaans, Christian Blaschke, Rafael Torres, Mariana Neves, Preslav Nakov, Anna Divoli, Manuel Maña-López, Jacinto Mata, and W John Wilbur. 2008. Overview of BioCreative II gene mention recognition. *Genome Biol.*, 9 Suppl 2(S2):S2.
- Yu Sun, Ke Tang, Leandro L Minku, Shuo Wang, and Xin Yao. 2016. Online ensemble learning of data streams with gradually evolved classes. *IEEE Transactions on Knowledge and Data Engineering*, 28(6):1532–1545.

- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#).
- Min Wang, Ke Fu, Fan Min, and Xiuyi Jia. 2020. Active learning through label error statistical methods. *Knowledge-Based Systems*, 189:105140.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. [Gpt-ner: Named entity recognition via large language models](#).
- Congying Xia, Wenpeng Yin, Yihao Feng, and Philip Yu. 2021. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. *arXiv preprint arXiv:2104.11882*.
- Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and Philip S Yu. 2018. Zero-shot user intent detection via capsule neural networks. *arXiv preprint arXiv:1809.00385*.
- Mohammad Yani, Adila Alfa Krisnadhi, and Indra Budi. 2022. A better entity detection of question for knowledge graph question answering through extracting position-based patterns. *Journal of Big Data*, 9(1):1–26.
- Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, Gyeong-Moon Park, Sung-Ho Bae, Lik-Hang Lee, Pan Hui, In So Kweon, and Choong Seon Hong. 2023a. [One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era](#).
- Chenwei Zhang, Nan Du, Wei Fan, Yaliang Li, Chun-Ta Lu, and S Yu Philip. 2017. Bringing semantic structures to user intent detection in online medical queries. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 1019–1026. IEEE.
- Ningyu Zhang, Moshua Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2021. [Cblue: A chinese biomedical language understanding evaluation benchmark](#).
- Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoi-fung Poon. 2023b. [Optimizing bi-encoder for named entity recognition via contrastive learning](#).
- Lingyun Zhao, Lin Li, Xinhao Zheng, and Jianwei Zhang. 2021. A bert based sentiment analysis and key entity detection approach for online financial texts. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1233–1238. IEEE.
- Binggui Zhou, Guanghua Yang, Zheng Shi, and Shaodan Ma. 2021. Natural language processing for smart healthcare. *arXiv preprint arXiv:2110.15803*.

# Author Index

- Abril, Eulalia P., 93  
Allen-Meares, Paula G., 93  
Andrew, Judith Jeyafreeda, 145  
Angles-Herrero, Emma, 109  
Aramaki, Eiji, 153
- Baker, Rose, 215  
Becart, Joffrey, 26  
Bellan, Patrizio, 124  
Ben-Haddi, Thouria, 189  
Berg, Nils, 246  
Bîndilă, Mihai Bogdan, 204  
Bjertnæs, Øyvind Andresen, 8  
Bobb, Susan C., 215  
Bouillon, Pierrette, 1  
Bourgeois, Florence T., 240  
Boyd, Andrew D., 93  
Briand, Thalassio, 26  
Burgun, Anita, 145
- Colombo, Amos, 137  
Consolandi, Monica, 124  
Coole, Matthew, 184  
Coridon, Clara, 26  
Cornelius, Joseph, 137  
Cuadros, Montse, 163  
Ćulafić, Milica, 104
- Dalianis, Hercules, 246  
Dalmiani, Stefano, 37  
David, Bastien, 1  
De la Clergerie, Éric, 115  
Deilen, Silvana, 44  
Denecke, Kerstin, 20  
di Buono, Maria Pia, 54  
Di Eugenio, Barbara, 93  
Di Scala, Jules, 189  
Dickens, Carolyn A., 93  
Dowson, Dai'Sha, 215  
Dragoni, Mauro, 124
- Eanes, Elisha, 215  
Eaves, Audrey, 215  
Elsweiler, David, 229  
Englebienne, Gwenn, 204
- Fernandez-Pichel, Marcos, 229  
Franchini, Andrea, 137
- Garcelon, Nicolas, 145  
Garcia-Bedoya, Olga, 93  
Gerhard-Krait, Francine, 26  
Gerlach, Johanna, 1  
Ghanbari Haez, Saba, 124  
Giordano, Luca, 54  
Glossop, Zoe, 184  
Goeuriot, Lorraine, 173  
González Torres, Juan José, 204  
Goyal, Pawan, 271  
Grabar, Natalia, 84  
Gracia, Jorge, 109  
Gupta, Mukur, 271
- Han, Zhishuo, 264  
Harris, David A., 240  
Hartendorp, Fons, 253  
Hernández Garrido, Sergio, 44  
Hofstee, Sebastiaan, 204  
Hörner, Julian, 44
- Ieva, Francesca, 37  
Ivchenko, Oksana, 84
- Jelin, Elma, 8
- Klein, Salomé, 26  
Knez, Timotej, 131  
Krasny-Pacini, Agata, 26  
Kuppelin, Marie, 26
- Lal, Daisy Monika, 73, 109  
Lapshinova-Koltunski, Ekaterina, 44  
Laroche, Joé, 26  
Lee, Jooyeon, 220  
Lee, Joseph G., 215  
Lithgow, Oscar William, 137  
Liu, Xingyu, 173  
Liu, Yufeng, 73, 109  
Lobban, Fiona, 184  
Losada, David E., 229  
Lovon-Melgarejo, Jesus, 189

Maaß, Christiane, 44  
Mæhlum, Petter, 8  
Magnolini, Simone, 124  
Mannion, Aidan, 173  
Marshall, Paul, 184  
Mazzucato, Sara, 37  
McNeill, Makyah, 215  
Meoni, Simon, 115  
Meyer, Selina, 229  
Micera, Silvestro, 37  
Miller, Timothy A., 240  
Mitrović, Sandra, 137  
Moccia, Sara, 37  
Moens, Marie-Francine, 197  
Moreno, Jose G., 189  
Mullick, Ankan, 271  
Mutal, Jonathan David, 1

Nguyen, Quang-Hung, 204  
Norman, Rebecka Maria, 8

Øvrelid, Lilja, 8

Pandey, Avinash Kumar, 264  
Paradossi, Umberto, 37  
Passino, Claudio, 37  
Payne, Sheila A., 73  
Petukhova, Volha, 64  
Pham, Luan Huy, 220  
Pitarch, Lucia, 109  
Popeskou, Sotirios Georgios, 137  
Portet, François, 173

Ragsdale, Hannah, 215  
Rajwal, Swati, 264  
Rayson, Paul, 73, 109, 184  
Reichenpfader, Daniel, 20  
Rietjens, Judith, 109  
Rigau, German, 163  
Rinaldi, Fabio, 137  
Rochat, Mathilde, 137  
Rothermich, Kathrin, 215  
Rubino, Raphael, 1  
Ruinelli, Lorenzo, 137  
Ryffel, Théo, 115

Salunke, Devika, 93  
Samuel, David, 8  
Sanna, Leonardo, 124  
Sarker, Abeed, 264  
Schwab, Didier, 173  
Segala, Marina, 124  
Segonne, Vincent, 173

Seinen, Tom, 253  
Shimizu, Seiji, 153  
Stojković, Stefan, 104  
Strasly, Irene, 1  
Szondy, Daniel, 204

Tamine, Lynda, 189  
Tayal, Anuja, 93  
Theel, Vanessa, 44  
Todirascu, Amalia, 26  
Torri, Vittorio, 37

Ulrich, Jean, 26  
Uma, Kanimozhi, 197  
Uzuner, Özlem, 220

van Mulligen, Erik, 253  
Vassiliadou, Hélène, 26  
Velldal, Erik, 8  
Verberne, Suzan, 253  
Vidler, John, 184  
Vincent, Marc, 145

Wakamiya, Shoko, 153  
Wan Ching Ho, Clara, 64  
Wang, Shenghui, 204  
Wunnava, Susmitha, 240

Yada, Shuntaro, 153

Zečević, Anđelka, 104  
Ziemer, Sophie, 44  
Žitnik, Slavko, 131  
Zotova, Elena, 163