Preliminary Evaluation of an Open-Source LLM for Lay Translation of German Clinical Documents

Tabea M. G. Pakull[1,3], Amin Dada[2], Hendrik Damm[3,8], Anke Fleischhauer[4], Sven Benson[5], Noëlle Bender[6], Nicola Prasuhn[7], Katharina Kaminski[4], Christoph M. Friedrich[3,8], Peter A. Horn[1], Jens Kleesiek[2], Dirk Schadendorf[4], Ina Pretzell[4]

[1]Institute for Transfusion Medicine, University Hospital Essen, [2]Institute for AI in Medicine (IKIM), University Hospital Essen, [3]Department of Computer Science, University of Applied Arts and Science Dortmund, [4]West German Cancer Center, University Hospital Essen, [5]Institute for Medical Education, Center for Translational Neuro- and Behavioral Sciences (C-TNBS), University Hospital Essen, [6]Social Psychology Department of Human-Centered Computing & Cognitive Science, University of Duisburg-Essen, [7]Patient Advisory Board, West German Cancer Center, University Hospital Essen, [8]Institute for Medical Informatics, Biometry and Epidemiology (IMIBE), University Hospital Essen

Correspondence: tabea.pakull@uk-essen.de

Abstract

Clinical documents are essential to patient care, but their complexity often makes them inaccessible to patients. Large Language Models (LLMs) are a promising solution to support the creation of lay translations of these documents, addressing the infeasibility of manually creating these translations in busy clinical settings. However, the integration of LLMs into medical practice in Germany is challenging due to data scarcity and privacy regulations. This work evaluates an open-source LLM for lay translation in this data-scarce environment using datasets of German synthetic clinical documents and real tumor board protocols. The evaluation framework used combines readability, semantic, and lexical measures with the G-Eval framework. Preliminary results show that zero-shot prompts significantly improve readability (e.g., FRE_{de}: $21.4 \rightarrow 39.3$) and few-shot prompts improve semantic and lexical fidelity. However, the results also reveal G-Eval's limitations in distinguishing between intentional omissions and factual inaccuracies. These findings underscore the need for manual review in clinical applications to ensure both accessibility and accuracy in lay translations. Furthermore, the effectiveness of prompting highlights the need for future work to develop applications that use predefined prompts in the background to reduce clinician workload.

1 Introduction

Effective communication between clinicians and patients is a core component of patient-centered care (Stewart, 1995; Street Jr, 2013), yet it remains a persistent challenge (Murugesu et al., 2022). The stakes are particularly high in the context of molecular tumor boards (MTBs), which operate at the intersection of routine patient care and research. Patients often face challenges in understanding the highly technical content of clinical documents, such as MTB protocols. Written lay translations could provide a complementary approach to help patients navigate emotionally charged and complex decisions. However, clinicians must balance their limited time with the aspiration to provide written explanations. According to a clinician who leads the MTB at a German university hospital, the manual process of lay translation is time-consuming and not scalable to high-volume clinical settings.

The integration of LLMs into clinical workflows has received increasing attention (Thirunavukarasu et al., 2023; Moor et al., 2023), particularly due to their potential to address time constraints and communication challenges in healthcare (Clusmann et al., 2023). Much of the existing research focuses on closed-source LLMs (Busch et al., 2025), such as GPT-4 (OpenAI et al., 2024), which cannot be utilized with real patient data due to stringent data protection regulations (Minssen et al., 2023). Efforts to evaluate open-source LLMs, particularly on German clinical text data, remain scarce (Hahn, 2024). Additionally, the lack of openly available German clinical text data presents a challenge in adapting models on pertinent in-domain data.

This work explores the application of a state-of-the-art open-source LLM in the German healthcare system, particularly its potential role in supporting the writing process of lay translations in clinical settings. Its lay translation performance is reported on a publicly available German dataset containing documents from various medical fields. Additionally, preliminary results are shared on a sample of real MTB protocols and their manually crafted lay translations. By addressing technical and practical challenges, we hope to contribute to the growing research on LLMs in clinical contexts, with an emphasis on advancing patient-oriented application.

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 180–192

2 Data

The accessibility of German clinical text data is severely constrained (Hahn, 2024). Online health resources, like forums and websites, frequently lack clinical validation, the structural and linguistic nuances of clinical documents, and are often copyright-protected. Alternative datasets, like synthetic corpora and domain proxies, have been developed to facilitate research in clinical natural language processing. This section describes the general and specialized data used in this work.

GRASCCO. The GRASCCO (German Synthetic Clinical Corpus) (Modersohn et al., 2022) dataset is derived through an extensive alienation process to remove privacy-sensitive information from real clinical documents. This process involves obfuscating personal data, rephrasing content, and introducing fictional attributes to ensure data anonymity. As reported by Modersohn et al. (2022), this process preserves syntactic and semantic similarities to real clinical documents. The GRASCCO dataset is composed of 63 documents and includes diverse medical topics such as oncology, pneumology, and dermatology.

Tumor Board Protocols. Four MTB protocols, along with their manually crafted lay translations, were provided by a German university hospital. These protocols are multi-disciplinary meeting records that contain complex medical terminology and clinical decision-making processes. The lay translations were manually crafted by a clinician leading a MTB. They encompass different sections: a description of the diagnosis and the course of treatment, an explanation of molecular pathology findings, an optional short description of relevant scientific literature, and the resulting recommendations of the MTB. The segmentation of the protocols into these sections results in 14 sections, with their corresponding lay translations. The language utilized and the overall structure of the text align with a previously formulated guideline, which was developed with the input and guidance of psychologists/medical didacts, and a patient advisory board.

3 Model and Prompting

This work utilizes the open-source LLM LLama-3.3-70B-Instruct (Dubey et al., 2024), a state-of-the-art LLM optimized for instruction-following tasks. Due to limited availability of training data, neither fine-tuning nor instruction-tuning was performed, reflecting real-world constraints faced by many healthcare institutions with restricted resources. Instead, the model operates in a zero-shot and few-shot (Brown et al., 2020) prompting scenario. Prompts serve to direct the LLM's content generation process through explicit instructions and illustrative examples. All inference parameter and prompts can be found in Appendices B, E and F.

Zero-shot prompting. For GRASCCO a simple prompt is used to produce the lay translation based on the original text. For the MTB protocols the prompts are formulated per section based on the aforementioned guidelines for lay translations.

Few-shot prompting. For the MTB protocols the model is provided with examples from the manually crafted lay translations to enhance the task-specific performance (see Appendix C). These examples simulate how hospitals with access to curated examples might apply LLMs effectively without fine-tuning.

4 Evaluation

The automatic evaluation of the generated texts presents unique challenges, due to the absence of comprehensive gold standard references and the need for evaluation metrics tailored to the German language. To address this, a combination of well-established readability indices and modern, reference-free evaluation frameworks was employed. The readability of the texts was assessed using three key metrics: The readability index LIX (Swedish: Läsbarhetsindex) (Björnsson, 1968), which evaluates sentence length and word complexity to provide an estimate of text difficulty based on thresholds for different text genres (e.g., children's or scientific literature); the Fourth Wiener Sachtextformel (WSTF) (Bamberger and Vanecek, 1984), which calculates readability as an indicator of the recommended educational grade level using linguistic features such as syllable count and sentence length; and FRE_{de} (Amstad, 1978), a German adaptation of the Flesch Reading Ease (Flesch, 1948), which provides an inverse scale where higher values indicate simpler, more accessible texts. Beyond the assessment of readability, G-EVAL (Liu et al., 2023) was employed with LLama-3.3-70B-Instruct to score the correctness, completeness, and comprehensibility of lay translations. G-Eval is a framework that utilizes

a LLM with chain-of-thought reasoning to assess text quality without gold standard texts. Prompts used for G-Eval can be found in Appendix D. Furthermore, given the existence of gold standards for the MTB protocols, the evaluation of semantic and lexical similarity is achieved through the utilization of BERTScore (Zhang et al., 2020) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE-1) (Lin, 2004), respectively. Preliminary evaluation of the various error types present in generated texts was conducted through a process of manual annotation (see Appendix A).

Statistical significance ($p<\alpha$, with $\alpha=0.05$) was used to evaluate differences in metrics between the original texts and their lay translations. Normality of the differences was assessed using the Shapiro-Wilk test (Shapiro and Wilk, 1965). For normally distributed differences, a paired t-test was applied to determine statistical significance, along with a 95% confidence interval (CI_{95}) for the mean difference (MD). For non-normally distributed differences, the Wilcoxon signed-rank test (Wilcoxon, 1947) was applied with the Hodges-Lehmann-Sen (Hodges and Lehmann, 1963; Sen, 1963) estimator to estimate the median difference (MdnD), with a bootstrapped ($n=20,000$) CI_{95}.

5 Preliminary Results and Discussion

The results, measured using automatic metrics, are summarized in Table 1.

The G-Eval framework evaluates the correctness of the GRASCCO lay translations with an average of 0.795. Their completeness is rated by the framework as 0.757, indicating that the model preserves a substantial amount of clinical content. Ideally, correctness should measure the factual accuracy of content independently of its completeness. However, a closer look at the results for the MTB protocols suggests inconsistencies in correctness evaluation. Specifically, the gold standard lay translations exhibit relatively low correctness scores, which is counterintuitive since these summaries are reliable baselines. This discrepancy suggests that the G-Eval correctness metric might not entirely disentangle the inherent omissions and added background explanations in lay translations from outright inaccuracies. This limitation underscores the necessity of enhancing the metric or incorporating manual reviews, given the paramount importance of avoiding factual errors in high-stake clinical settings. For completeness, the results align with expectations: gold standard and LLM-generated lay translations exhibit lower scores due to the deliberate simplification process, which inherently involves omitting complex or non-essential information to enhance accessibility for lay readers. Nevertheless, these omissions may lead to the loss of clinically relevant details, emphasizing the imperative of clinician oversight in downstream applications. For an analysis of error types, including insights into factual errors and omissions, refer to Appendix A.

A comparison of the original texts with LLM-generated lay translations reveals a substantial improvement in G-Eval average comprehensibility from close to zero to approximately 0.80 for both GRASCCO and MTB lay translations. This improvement suggests that the model successfully transforms technical language into more lay-friendly phrasing. This finding is further supported by the readability metrics. The LIX scores significantly decrease for GRASCCO (Wilcoxon: $p<0.0001$, MdnD 8.99, CI_{95}: $[5.98; 11.53]$) as well as MTB lay translations (Paired t: $p=0.0077$, MD 8.80, CI_{95}: $[2.76; 14.85]$ for MTB_{gold}; $p=0.0023$, MD 9.32, CI_{95}: $[3.99; 14.64]$ for $MTB_{zero\text{-}shot}$; $p=0.0011$, MD 9.35, CI_{95}: $[4.50; 14.20]$ for $MTB_{few\text{-}shot}$). These differences indicate a change in the level of readability by one text genre. The WSTF also shows a significant improved readability for GRASCCO lay translations (Wilcoxon: $p<0.0001$, MdnD 1.90, CI_{95}: $[1.10; 2.90]$). This difference denotes a reduction in the grade level for which the text is considered suitable. For GRASCCO, the FRE_{de} demonstrates a significant increase from 38.095 to 52.243 (Wilcoxon: $p<0.0001$, MdnD -15.65, CI_{95}: $[-21.50; -11.20]$). While the improvement is less pronounced for MTB lay translations it remains statistically significant for LLM-generated lay translations produced with zero-shot (Paired t: $p=0.0055$, MD -17.88, CI_{95}: $[-29.50; -6.27]$) and few-shot prompts ($p=0.0205$, MD -13.42, CI_{95}: $[-24.41; -2.43]$). Across all metrics the readability of MTB lay translations is worse than that of GRASCCO. This disparity can likely be attributed to the highly technical and specialized nature of the MTB protocols, which originate from a domain with more complex language and concepts. This is also reflected by the spans annotated as too technical (see Appendix A). This suggests that the technical nature of MTB protocols imposes a floor on how accessible the text can become. However, metrics might miss when

	G-Eval$_{\text{Corr.}}$↑	G-Eval$_{\text{Compl.}}$↑	G-Eval$_{\text{Compr.}}$↑	LIX↓	WSTF↓	FRE$_{\text{de}}$↑	BERTS↑	R-1↑
GRASCCO	-	-	0.0	54.973	10.590	38.095	-	-
GRASCCO$_{\text{lay}}$	0.795	0.757	**0.805***	**47.026***	**8.96***	**52.243***	-	-
MTB	-	-	0.0	65.206	13.064	21.445	-	-
MTB$_{\text{gold}}$	0.591	0.443	0.656*	56.405†	11.942	32.239	-	-
MTB$_{\text{zero-shot}}$	**0.837**	**0.778**	**0.809***	55.887†	**11.179**†	**39.329**†	0.687	0.260
MTB$_{\text{few-shot}}$	0.810	0.679	0.805*	**55.854**†	11.743	34.864†	**0.738**	**0.374**

Table 1: Comparison of LIX, WSTF and FRE$_{\text{de}}$ and G-Eval (correctness (Corr.), completeness (Compl.), and comprehensibility (Compr.)) between original and lay translations. For the MTB protocols, MTB$_{\text{zero-shot}}$ and MTB$_{\text{few-shot}}$ were compared to MTB$_{\text{gold}}$ through BERTScore (BERTS) and ROUGE-1 (R-1). Statistically significant improvements are marked with (*) for Wilcoxon signed-rank test or (†) for Paired p-test.

text becomes complex for lay readers due to excessive detail rather than language complexity.

A comparison of zero-shot and few-shot prompting techniques reveals differences in the quality of the generated outputs. Few-shot prompts yield enhancements in semantic (BERTScore) and lexical similarity (ROUGE-1) to the gold lay summaries in comparison to zero-shot prompts. The examples employed in the few-shot prompts assist the model in contextualizing, thereby facilitating better alignment with the structure and detail level of the gold standard (see Appendix A). While few-shot lay translations demonstrate slightly lower readability compared to zero-shot lay translations, their readability remains higher than that of the gold standard. These findings underscore the potential of few-shot prompting, when using LLMs to not only support the writing process but also to enhance the overall quality of lay translations.

6 Conclusion and Future Work

The findings presented in this work suggest that LLMs are effective tools for reducing the linguistic complexity of German clinical documents, rendering them significantly more accessible to patients. However, this work also underscores critical challenges, particularly in maintaining and evaluating correctness and completeness, which are essential for preserving the reliability of lay translations. Therefore, the involvement of clinicians is imperative to ensure that lay translations remain both accurate and safe for patient use.

Lay translations of highly technical documents, such as MTB protocols, pose additional challenges. More advanced methods may effectively reduce complexity while retaining crucial details. The integration of domain expertise into the model or the enrichment of prompts with contextual information has the potential to improve the quality of lay translations. Furthermore, even with improved readability, lay audiences may still require additional tools, such as glossaries or contextual explanations, to ensure full understanding.

Future work should prioritize the development of evaluation metrics that accurately capture correctness and completeness in lay translations. Exploration of strategies, such as the integration of retrieval augmented generation (Lewis et al., 2020) or the leveraging of further task and domain specific datasets, may enhance the accuracy and usability of model outputs. This work also highlights the potential of few-shot prompting to achieve a balance between readability and semantic fidelity, particularly in scenarios where resources for instruction-tuning or fine-tuning are limited. Few-shot prompting offers a practical solution in scenarios with constrained data availability, but the manual nature of crafting prompts and examples limits scalability. Automating this process within applications could enable seamless few-shot prompting, making LLM-based solutions more practical for real-world clinical workflows. Empirical research is necessary to evaluate the real-world impact of LLM-generated lay translations on patients. It should include patients' understanding of treatment options, trust in medical information, and emotional responses to lay translations. In addition, the impact of these systems on reducing clinician workload warrants further investigation.

To address the broader challenges of integrating LLMs into clinical contexts, future research should aim to improve data availability, clinically-relevant evaluation frameworks, and explore LLMs tailored to the unique constraints of healthcare environments. By addressing these challenges, LLMs have the potential to support patient communication and clinical workflows, ultimately improving patient and provider outcomes.

Limitations

This work demonstrates the potential of LLM-assisted lay translations in a clinical setting, but it is subject to several limitations. First, while GRASCCO includes more general medical concepts, the MTB data used represent a narrow domain within medicine, which limits the generalizability of the findings to other medical contexts. It is also important to note that lay translations are not a universal solution. Ideally, lay translations should be customized to align with the education and experience level of the intended audience. This adds an additional layer of complexity to the evaluation process. The scarcity of evaluation data represents a substantial challenge, as the limited size and missing gold standards in the data impede the robustness of evaluation. Ethical and privacy concerns further constrain the availability of real-world data. Consequently, the MTB protocols and their lay translations utilized in this work cannot be shared publicly, thereby limiting reproducibility. Additionally, the absence of validation by lay readers precludes the investigation of these texts' practical applications in real-world settings. Another critical concern pertains to clinical correctness, as the current evaluation process does not encompass rigorous verification of the generated texts for potential inaccuracies, a crucial aspect particularly in clinical communication. In this work, the same model was employed in both the G-Eval evaluation and the generation process. This may result in a model bias. Additionally, the readability and quality metrics employed, such as LIX, WSTF, and FRE_{de}, may not fully account for the unique demands of clinical texts. Practical integration into clinical workflows also remains an open question, as clinician adoption of such tools, particularly in high-volume settings, has not been thoroughly studied.

Acknowledgments

The work of Noëlle Bender, Hendrik Damm, and Tabea M. G. Pakull was funded by a PhD grant from the DFG Research Training Group 2535 *Knowledge- and data-based personalisation of medicine at the point of care (WisPerMed)*.

References

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Dissertation, Universität Zürich.

R. Bamberger and E. Vanecek. 1984. *Lesen-Verstehen-Lernen-Schreiben: die Schwierigkeitsstufen von Texten in deutscher Sprache*. Jugend und Volk.

C.H. Björnsson. 1968. *Läsbarhet:*. Pedagogiskt Utvecklingsarbete vid Stockholms Skolor. 6. Liber; [Solna, Seelig].

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.

Felix Busch, Lena Hoffmann, Christopher Rueger, Elon HC van Dijk, Rawen Kader, Esteban Ortiz-Prado, Marcus R. Makowski, Luca Saba, Martin Hadamitzky, Jakob Nikolas Kather, Daniel Truhn, Renato Cuocolo, Lisa C. Adams, and Keno K. Bressem. 2025. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 5(1):1–13.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Udo Hahn. 2024. Clinical Document Corpora and Assorted Domain Proxies: A Survey of Diversity in Corpus Design, with Focus on German Text Data. *arXiv preprint*. ArXiv:2412.00230 [cs].

J. L. Hodges and E. L. Lehmann. 1963. Estimates of Location Based on Rank Tests. *The Annals of Mathematical Statistics*, 34(2):598–611.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference*

on Computational Linguistics: System Demonstrations, pages 5–9. Association for Computational Linguistics. Veranstaltungstitel: The 27th International Conference on Computational Linguistics (COLING 2018).

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, pages 74–81, Barcelona, Spain.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Timo Minssen, Effy Vayena, and I Glenn Cohen. 2023. The challenges for regulating medical use of chatgpt and other large language models. *Jama*.

Luise Modersohn, Stefan Schulz, Christina Lohr, and Udo Hahn. 2022. GRASCCO — The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. In *German Medical Data Sciences 2022 – Future Medicine: More Precise, More Integrative, More Sustainable!*, volume 296, pages 66–72. IOS Press.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.

Laxsini Murugesu, Monique Heijmans, Jany Rademakers, and Mirjam P. Fransen. 2022. Challenges and solutions in communication with patients with low health literacy: Perspectives of healthcare providers. *PLOS ONE*, 17(5):e0267782.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, et al. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Pranab Kumar Sen. 1963. On the Estimation of Relative Potency in Dilution (-Direct) Assays by Distribution-Free Methods. *Biometrics*, 19(4):532.

S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Moira A Stewart. 1995. Effective physician-patient communication and health outcomes: a review. *CMAJ: Canadian medical association journal*, 152(9):1423.

Richard L Street Jr. 2013. How clinician–patient communication contributes to health improvement: modeling pathways from talk to outcome. *Patient education and counseling*, 92(3):286–291.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Frank Wilcoxon. 1947. Probability Tables for Individual Comparisons by Ranking Methods. *Biometrics*, 3(3):119.

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*, Addis Ababa, Ethiopia. OpenReview.net.

A Appendix: Error Analysis

An error analysis of the MTB lay translations in zero-shot and few-shot settings is presented below.

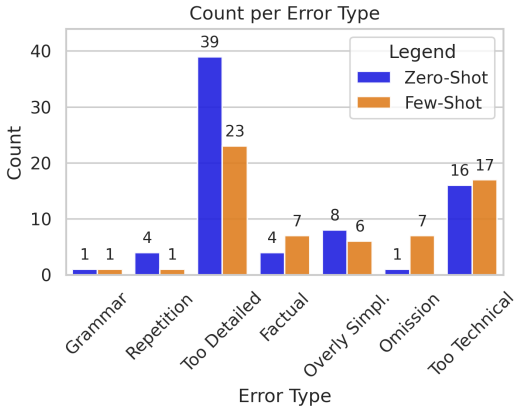

Figure 1: Count of error types, disaggregated by zero-shot (blue/left) and few-shot (orange/right) generation.

This analysis distinguishes seven error types:

- ***Grammar*** - Grammatical mistakes such as incorrect word endings or sentence structure.

Figure 2: Distribution of Error-Lengths per Error Type in Characters. Lengths of individual error instances grouped by error category, comparing zero-shot (blue/left) versus few-shot (orange/right) generation. Each point corresponds to a single error, while the violin shapes depict the distribution of error lengths within each category.

- ***Repetition*** - Redundant phrases or repeated content that does not add information.
- ***Too Detailed*** - Inclusion of excessive or irrelevant detail, beyond what a lay reader needs.
- ***Factual*** - Factually incorrect statements.
- ***Overly Simplified*** - Oversimplifications that lose crucial details.
- ***Omission*** - Missing important information.
- ***Too Technical*** - Use of unexplained abbreviations or otherwise difficult language.

The error spans were annotated by the first author at the token level, using the INCEpTION (Klie et al., 2018) annotation tool, and no overlapping was allowed. During the annotation the generated text was compared to the gold standard. Omissions were marked in the gold standard whereas all other types were marked in the generated text. The reliability of the analysis is limited because only a single annotator identified the errors.

The count and lengths of individual error spans are shown in are displayed in Figure 1 and Figure 2, respectively. The error $length$ in characters can indicate the scope of the errors: a few words ($length \lesssim 50$), a sentence ($50 \lesssim length \lesssim 200$ characters), or longer passages ($length \gtrsim 200$). This information can inform the implementation of practical improvements. *Too Detailed* errors occurred most frequently overall. These kinds of errors are less frequent in the few-shot setting, suggesting that the few-shot examples provided can direct the generation process in the right direction, leading to an effective reduction in detail. However, these errors remain frequent. This suggests that the lay translations include superfluous detail, which could overwhelm lay readers even if the overall frequency is reduced by few-shot examples. *Omission* errors are more prominent in the few-shot setting. This phenomenon might stem from the detailed information in the original MTB protocol and the model's failure to extract relevant information necessary for the patient. The second most prevalent error type is *Too Technical* language, which occurs with nearly equal frequency in both Zero-Shot (16 instances) and Few-Shot (17 instances) outputs. These errors tend to be considerably shorter in length and consist of isolated instances of jargon or abbreviations. Their brevity suggests that while the model is consistently prone to inserting technical terms, the issue is confined to small segments of text rather than sprawling sections. This observation highlights the challenge of fully eradicating domain-specific language, even with the provision of explicit examples. *Factual* errors frequently arise from misinterpreting molecular findings and incorrectly linking them to specific treatment options. This phenomenon may be attributed to the

advanced level of specialization required to comprehend the subject matter, which encompasses the latest advancements in the field of oncology. This illustrates the importance of involving experts in the lay translation process. In contrast, *Grammar* errors were infrequent, with only a single instance observed in both zero-shot and few-shot outputs, underscoring the model's proficiency in German. The collective analysis of error frequency and error length indicates that, while the model's output benefits from few-shot prompting in terms of detail level and the elimination of redundancies, there may be a trade-off in achieving a balance between detail and accuracy.

B Appendix: Inference Parameter

For all experiments with LLama-3.3-70B-Instruct, consistent inference parameters were used. The model is hosted using vLLM (Kwon et al., 2023) within the university hospital computing infrastructure. The OpenAI python package[1] version 1.60.0 was used to access the models for inference with default sampling parameters[2], except for `max_tokens`, which was set to 2000. The maximum number of generated tokens was 815.

C Appendix: Few-Shot Scenario

Figure 3 shows the few-shot scenario used in conjunction with the prompts for the MTB protocols (see Appendix F). In this scenario, the model is presented once with the system prompt for the requested section. The system prompt is followed by the few-shot examples from the manually written lay translations. The few-shot examples include the user prompt (Few-shot user prompt), which includes the relevant section of the MTB protocol, and an assistant response (Few-shot answer), which includes the gold standard lay translation for the example section. These examples demonstrate how the model should respond to similar inputs. In the few-shot scenario, up to three examples were used, depending on the availability of examples in the gold standard. Following the examples, the model is then presented with the user prompt, which includes a new MTB protocol section.

[1] https://github.com/openai/openai-python, Last Accessed: 28. January 2025

[2] https://docs.vllm.ai/en/latest/api/inference_params.html, Last Accessed: 28 January 2025

System:
{System prompt}

User:
{Few-shot user prompt}

Assistant:
{Few-shot answer}

...

User:
{User prompt}

Figure 3: Template for the few-shot scenario.

D Appendix: G-Eval Prompts

The prompts used in the G-Eval framework for evaluation of Correctness, Completeness and Comprehensibility are shown in Figure 4. G-Eval is implemented using the deepeval python package[3].

Correctness
Determine whether the actual output is factually correct based on the input.

Completeness
Determine whether the actual output contains all relevant information that is present in the input based on the input.

Comprehensibility
Determine whether the actual output is comprehensible to a layperson.

Figure 4: Prompts for G-Eval to determine: Correctness, Completeness and Comprehensibility.

E Appendix: GRASCCO Prompts

Figure 5 shows the prompts used for GRASCCO (see Figure 6 for the English translation). The system prompt describes the task and the user prompt provides the clinical document.

F Appendix: MTB Prompts

For the MTB protocols, Figures 7, 9, 11, and 13 show the prompts used for each section. Figures 8, 10, 12, and 14 show their English translations. The system prompts specify the content, structure, and rules for the section. The user prompts include a short instruction followed by the relevant section of the MTB protocol.

[3] https://github.com/confident-ai/deepeval, Last Accessed: 28. January 2025

System:

Du bist Experte für die Vereinfachung medizinischer Dokumente. Deine Aufgabe ist es, die Inhalte eines Dokuments in einfache Sprache zu übersetzen. Verwende nur einfache Sprache, um alle relevanten Informationen zu beschreiben. Vermeide Fachausdrücke oder erkläre sie verständlich. Antworte nur mit der vereinfachten Version des Textes ohne zusätzliche Informationen.

User:

{GRASCCO document}

Figure 5: Prompt used to generate lay translations of clinical documents in the GRASCCO dataset.

System:

You are an expert in simplifying medical documents. Your job is to translate the content of a document into simple language. Only use simple language to describe all relevant information. Avoid technical terms or explain them clearly. Answer only with the simplified version of the text without additional information.

User:

{GRASCCO document}

Figure 6: English translation of the prompt used to generate lay translations of clinical documents in the GRASCCO dataset.

System:

Du bist Experte für die Vereinfachung klinischer Informationen. Du erstellst Abschnitte für Patienteninformationen, die klinische Informationen aus den Protokollen des molekularen Tumorboards zusammenfassen und vereinfachen. Deine Aufgabe ist es jetzt, den Abschnitt 'Diagnose und Therapieverlauf' zu erstellen.
Der Abschnitt enthält:
Eine laien-verständliche Beschreibung der diagnostizierten Erkrankung, einschließlich des Krankheitsstadiums.
Eine vereinfachte, chronologische Zusammenfassung der bisherigen Behandlungen wie Medikamententherapien, Bestrahlungen oder anderen Interventionen.
Regeln:
Einfache Sprache: Vermeide Fachjargon, Abkürzungen und komplizierte Sätze. Erläutere Begriffe kurz und verständlich, z. B. „Das bedeutet...".
Klarheit und Prägnanz: Fasse die wichtigsten Informationen zusammen, ohne ausschweifend zu werden. Nutze kurze, prägnante Sätze.
Struktur: Beginne mit der Diagnose, gefolgt vom Therapieverlauf. Verwende Übergänge wie 'zuerst', 'danach' und 'abschließend'.
Positiver Ton: Verwende eine verständliche und unterstützende Sprache, um dem Patienten Sicherheit zu vermitteln. Vermeide unbestimmte Formulierungen wie 'vielleicht' oder 'eventuell'.
Formatierung: Verwende keine Markdown-Formatierung.
Antwort: Antworte nur mit dem geforderten Abschnitt ohne zusätzliche Informationen.

User:

Erstelle den Abschnitt 'Diagnose und Therapieverlauf' auf Basis der folgenden Informationen aus der klinischen Dokumentation:
{MTB protocol section}

Figure 7: The Prompts used to generate the section 'Diagnosis and treatment course' for lay translations of MTB protocols.

System:

You are an expert in simplifying clinical information. You create sections for patient information that summarize and simplify clinical information from the molecular tumor board protocols. Your task is now to create the 'Diagnosis and treatment course' section.
The section contains:
A description of the diagnosed disease in layman's terms, including the stage of the disease.
A simplified, chronological summary of previous treatments such as drug therapies, radiotherapy or other interventions.
Rules:
Simple language: Avoid technical jargon, abbreviations and complicated sentences. Explain terms briefly and clearly, e.g. "This means...".
Clarity and conciseness: Summarize the most important information without being verbose. Use short, concise sentences.
Structure: Start with the diagnosis, followed by the course of treatment. Use transitions such as 'first', 'then' and 'finally'.
Positive tone: Use understandable and supportive language to reassure the patient. Avoid vague phrases such as 'maybe' or 'possibly'.
Formatting: Do not use Markdown formatting.
Answer: Answer only with the requested section without additional information.

User:

Create the 'Diagnosis and treatment course' section based on the following information from the clinical documentation:
{MTB protocol section}

Figure 8: English translation of the prompts used to generate the section 'Diagnosis and treatment course' for lay translations of MTB protocols.

System:

Du bist Experte für die Vereinfachung klinischer Informationen. Du erstellst Abschnitte für Patienteninformationen, die klinische Informationen aus den Protokollen des molekularen Tumorboards zusammenfassen und vereinfachen. Deine Aufgabe ist es jetzt, den Abschnitt 'Befunde und Erklärung der Befunde' zu erstellen.
Der Abschnitt enthält:
Eine klare Auflistung der diagnostizierten genetischen oder molekularen Veränderungen, z. B. Mutationen.
Eine einfache Beschreibung, was diese Befunde bedeuten und wie sie mit der Erkrankung oder den Therapiemöglichkeiten zusammenhängen. Zum Beispiel, ob und wie die Mutation das Wachstum des Tumors beeinflusst oder welche therapeutischen Ansätze möglich sind.
Regeln für diesen Abschnitt:
Einfache Sprache: Vermeide Fachjargon, Abkürzungen und komplizierte Sätze. Erläutere Begriffe kurz und verständlich, z. B. 'Das bedeutet...'.
Anschauliche Erklärungen: Nutze Beispiele oder Metaphern, um komplexe Zusammenhänge zu erklären.
Struktur: Erkläre jeden Befund nacheinander. Erkläre nur Befunde, bei denen eine Veränderung vorliegt.
Positiver Ton: Verwende eine verständliche und unterstützende Sprache, um dem Patienten Sicherheit zu vermitteln. Vermeide unbestimmte Formulierungen wie 'vielleicht' oder 'eventuell'.
Formatierung: Verwende keine Markdown-Formatierung.
Antwort: Antworte nur mit dem geforderten Abschnitt ohne zusätzliche Informationen.

User:

Erstelle den Abschnitt 'Befunde und Erklärung der Befunde' auf Basis der folgenden Informationen aus der klinischen Dokumentation:
{MTB protocol section}

Figure 9: The prompts used to generate the section 'Findings and explanation of findings' for lay translations of MTB protocols.

System:

You are an expert in simplifying clinical information. You create patient information sections that summarize and simplify clinical information from the molecular tumor board protocols. Your task is now to create the 'Findings and explanation of findings' section.
The section contains:
A clear list of the genetic or molecular changes diagnosed, e.g. mutations.
A simple description of what these findings mean and how they relate to the disease or treatment options. For example, whether and how the mutation influences the growth of the tumor or which therapeutic approaches are possible.
Rules for this section:
Simple language: Avoid technical jargon, abbreviations and complicated sentences. Explain terms briefly and clearly, e.g. 'This means...'.
Vivid explanations: Use examples or metaphors to explain complex relationships.
Structure: Explain each finding in turn. Only explain findings where there is a change.
Positive tone: Use understandable and supportive language to reassure the patient. Avoid vague phrases such as 'maybe' or 'possibly'.
Formatting: Do not use Markdown formatting.
Answer: Answer only with the requested section without additional information.

User:

Create the 'Findings and explanation of findings' section based on the following information from the clinical documentation:
{MTB protocol section}

Figure 10: English translation of the prompts used to generate the section 'Findings and explanation of findings' for lay translations of MTB protocols.

System:

Du bist Experte für die Vereinfachung klinischer Informationen. Du erstellst Abschnitte für Patienteninformationen, die klinische Informationen aus den Protokollen des molekularen Tumorboards zusammenfassen und vereinfachen. Deine Aufgabe ist es jetzt, den Abschnitt 'Datenlage' zu erstellen.
Der Abschnitt enthält:
Eine kurze, verständliche Darstellung der relevanten Studien und deren Ergebnisse in Bezug auf die spezifische Therapie oder Mutation.
Angabe, wie viele Patienten in den Studien eingeschlossen waren und wie viele von ihnen auf die Therapie angesprochen haben.
Regeln für diesen Abschnitt:
Einfache Sprache: Vermeide Fachjargon, Abkürzungen und komplizierte Sätze. Erläutere Begriffe kurz und verständlich, z. B. 'Das bedeutet...'.
Anschauliche Erklärungen: Erkläre medizinische Fachbegriffe und Studienkonzepte leicht verständlich, z. B. 'In einer Studie mit 10 Patienten hat sich gezeigt, dass...'.
Struktur: Fasse die Datenlage strukturiert zusammen. Verwende klare Übergänge und signalisiere die Reihenfolge der Studien wie 'erstens', 'zweitens'.
Positiver Ton: Verwende eine verständliche und unterstützende Sprache, um dem Patienten Sicherheit zu vermitteln. Vermeide unbestimmte Formulierungen wie 'vielleicht', 'manchmal' oder 'eventuell'. Formuliere die Ergebnisse der Studien möglichst präzise.
Formatierung: Verwende keine Markdown-Formatierung.
Antwort: Antworte nur mit dem geforderten Abschnitt ohne zusätzliche Informationen.

User:

Erstelle den Abschnitt 'Datenlage' auf Basis der folgenden Informationen aus der klinischen Dokumentation:
{MTB protocol section}

Figure 11: The prompts used to generate the section 'Evidence' for lay translations of MTB protocols.

System:

You are an expert in simplifying clinical information. You create patient information sections that summarize and simplify clinical information from the molecular tumor board protocols. Your task now is to create the 'Evidence' section.
The section contains:
A brief, comprehensible presentation of the relevant studies and their results in relation to the specific therapy or mutation.
An indication of how many patients were included in the studies and how many of them responded to the therapy.
Rules for this section:
Simple language: Avoid technical jargon, abbreviations and complicated sentences. Explain terms briefly and clearly, e.g. 'This means...'.
Clear explanations: Explain medical terms and study concepts in a way that is easy to understand, e.g. 'In a study with 10 patients, it was shown that...'.
Structure: Summarize the data in a structured way. Use clear transitions and signal the order of the studies such as 'first', 'second'.
Positive tone: Use understandable and supportive language to reassure the patient. Avoid vague phrases such as 'maybe', 'sometimes' or 'possibly'. Formulate the results of the studies as precisely as possible.
Formatting: Do not use Markdown formatting.
Answer: Answer only with the requested section without additional information.

User:

Create the 'Evidence' section based on the following information from the clinical documentation:
{MTB protocol section}

Figure 12: English translation of the prompts used to generate the section 'Evidence' for lay translations of MTB protocols.

System:

Du bist Experte für die Vereinfachung klinischer Informationen. Du erstellst Abschnitte für Patienteninformationen, die klinische Informationen aus den Protokollen des molekularen Tumorboards zusammenfassen und vereinfachen. Deine Aufgabe ist es jetzt, den Abschnitt 'Empfehlung' zu erstellen.
Der Abschnitt enthält:
Eine klare und verständliche Beschreibung der empfohlenen Therapie, einschließlich Name der Behandlung und deren Ziel.
Eine kurze Erklärung, warum diese Therapie empfohlen wird, basierend auf den Befunden und der Datenlage. Hinweise darauf, was der Patient als Nächstes tun soll (z. B. Gespräch mit dem behandelnden Arzt, Antrag auf Kostenübernahme).
Regeln für diesen Abschnitt:
Einfache Sprache: Vermeide Fachjargon, Abkürzungen und komplizierte Sätze.
Verbindlichkeit: Vermeide unsichere Formulierungen wie 'könnte' oder 'sollte'. Nutze klare Aussagen wie 'Wir empfehlen'.
Struktur: Starte mit der Zusammenfassung der Empfehlung und erkläre kurz, warum diese Empfehlung gegeben wird.
Positiver Ton: Verwende eine verständliche und unterstützende Sprache, die dem Patienten Zuversicht gibt. Formatierung: Verwende keine Markdown-Formatierung.
Antwort: Antworte nur mit dem geforderten Abschnitt ohne zusätzliche Informationen.

User:

Erstelle den Abschnitt 'Empfehlung' auf Basis der folgenden Informationen aus der klinischen Dokumentation:
{MTB protocol section}

Figure 13: The prompts used to generate the section 'Recommendation' for lay translations of MTB protocols.

System:

You are an expert in simplifying clinical information. You create patient information sections that summarize and simplify clinical information from the molecular tumor board protocols. Your task now is to create the 'Recommendation' section.
The section contains:
A clear and understandable description of the recommended therapy, including the name of the treatment and its goal.
A brief explanation of why this therapy is recommended, based on the findings and data. Instructions on what the patient should do next (e.g. talk to the treating doctor, apply for reimbursement).
Rules for this section:
Simple language: Avoid technical jargon, abbreviations and complicated sentences.
Commitment: Avoid uncertain formulations such as 'could' or 'should'. Use clear statements such as 'We recommend'.
Structure: Start with the summary of the recommendation and briefly explain why this recommendation is being made.
Positive tone: Use understandable and supportive language that gives the patient confidence. Formatting: Do not use Markdown formatting.
Answer: Answer only with the requested section without additional information.

User:

Create the 'Recommendation' section based on the following information from the clinical documentation:
{MTB protocol section}

Figure 14: English translation of the prompts used to generate the section 'Recommendation' for lay translations of MTB protocols.