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[1] 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

---

[1] GUIDELINE ON THE READABILITY OF THE LABELLING AND PACKAGE LEAFLET OF MEDICINAL PRODUCTS FOR HUMAN USE - European Commission, 2009

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 54–63
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

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[2], 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"[3], i.e., comprehensive, electron-

---

[2] AIFA - Homepage

[3] 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[4] 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[5], while Class A drugs in Italy are defined by AIFA as essential medicines and medicines for chronic diseases.[6] 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*[7], *requests*[8] and *selenium*.[9] 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*.[10] 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, 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[11], 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.

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

Table 1: Corpus Description

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[12] for future use and further exploration by the research community.

---

[4] AIFA - Banca Dati Farmaci

[5] EMA - Glossary:Generic medicine

[6] Ministry of Health - Drug classes

[7] pandas

[8] requests

[9] selenium

[10] PyPDF2

[11] See the CMD(h) ANNOTATED QRD TEMPLATE FOR MR/DC PROCEDURES

[12] 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.[13] 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,

---

[13] 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.[14] 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) DURATION, (vii) DOSAGGIO (dosage), (viii) FORM (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 ρ and Kendall's τ 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 Gulpease 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.

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[15] (Galatolo and Cimino, 2023). Specifically, we exploit cerbero-7b-openchat-gguf[16]. 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.[17]

Enviromental 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[18] 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*[19]. 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.

---

[14] DeepMount00/Italian_NER_XXL on HuggingFace

[15] Cerbero-7B on GitHub

[16] galatolo/cerbero-7b-openchat-gguf on HuggingFace

[17] HuggingFace documentation - Quantization

[18] llama-cpp-python

[19] "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 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.

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

The first type of results is considered correct, while 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.

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

Table 3: Result types

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.*[20]

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 consulare sempre il medico prima di assumere questo prodotto.*[21]

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:

---

[20] 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.

[21] 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.*[22]

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[23]. 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 instrinsic 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.

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 ρ and Kendall's τ, 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

---

[22] 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.

[23] 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.

	ZSP1	ZSP2	ZSP1	ZSP2
	Spearman's ρ	**Spearman's ρ**	**Kendall's τ**	**Kendall's τ**
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 ρ and Kendall's τ correlation coefficients between cosine similarities and NER F1 scores of answers to ZSP1 and ZSP2

Prompt	GT	ZSP1	ZSP2
Q1	43.30 (*s* 8)	48.80 (*s* 7)	49.40 (*s* 7.90)
Q2	41.50 (*s* 5.80)	48.40 (*s* 4.40)	48.80 (*s* 6.60)
Q3	52.20 (*s* 6.20)	55.60 (*s* 5.40)	55.10 (*s* 4.80)
Q4	37 (*s* 9.40)	43.20 (*s* 10.30)	47.50 (*s* 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 "Ricercatori 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`. 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 Woldesenbet, 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`. 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 leggibilita 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*.