

DIMMI - Drug InforMation Mining in Italian: A CALAMITA Challenge

Raffaele Manna^{1,†}, Maria Pia di Buono^{1,*} and Luca Giordano^{1,†}

¹University of Naples "L'Orientale", via Duomo 219, 80139 Napoli, Italy

Abstract

Patients' knowledge about drugs and medications is crucial as it allows them to administer them safely. This knowledge frequently comes from written prescriptions, patient information leaflets (PILs), or from reading drug Web pages. DIMMI (Drug InforMation Mining in Italian) is a challenge aiming at evaluating the proficiency of Large Language Models in extracting drug-specific information from PILs. The challenge seeks to advance the understanding of effectiveness in processing complex medical information in Italian, and to enhance drug information extraction and pharmacovigilance efforts. Participants are provided with a dataset of 600 Italian PILs and the objective is to develop models capable of accurately answering specific questions related to drug dosage, usage, side effects, drug-drug interactions. The challenge should be approached as an information extraction task through a zero-shot mode, purely based on the model pre-existing knowledge and understanding or through in-context learning (Retrieval-Augmented Generation (RAG) or few-shot mode). The answers generated by the models will be compared against the gold standard (GS), created to establish a reliable, accurate, and a comprehensive set of answers against which participant submissions can be evaluated. For each drug and each information category, the GS contains the correct information extracted from the leaflets through a manual annotation.

Keywords

Patient information leaflets, Information extraction, Large Language Models, Italian

1. Introduction and Motivation

Patients' knowledge about drugs and medications is crucial as it allows them to administer them safely. This knowledge frequently comes from written prescriptions, patient information leaflets (PILs), or from reading drug Web pages. Nevertheless, this information has been described as often inconsistent, incomplete, and difficult for patients to read and understand [1]. Despite the fact that in 2009 the European Commission issued guidelines¹ to recommend the publication of patient information leaflets with accessible and understandable information for patients, several scholars [2, 3, 4] 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, there has been a growing interest in the utilization of Large Language Models (LLMs) within the medical field to improve various aspects of healthcare, including medical education and clinical decision-making support [5]. Several specialized medical LLMs have been developed through novel pre-training methodologies or enhancements of existing models. Moreover, several evaluation campaigns have been undertaken to evaluate the efficacy of natural language processing models in facilitating knowledge retrieval for clinicians and patients alike. Examples of such campaigns are the 1) *Medical Question Answering Task* at *TREC-2017 LiveQA* [6] and subsequent studies [7], which led to two datasets, *LiveQA* and *MedicationQA*; 2) the tasks on *Medical Consumer Question Answering* proposed by Nguyen et al. [8] based on their dataset *MedRedQA*. Both campaigns have contributed significantly to bridging the gap between consumers' medication questions and trusted answers, and, more generally, to the development of resources tailored to healthcare information retrieval. For a thorough survey of evaluation campaigns on clinical natural language processing refer to Filannino and Uzuner [9].

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 [10], as any concealed inaccuracies in diagnoses and health advice could lead to severe outcomes [11]. For these reasons, in the evolving landscape of Artificial Intelligence (AI) applications in

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 – 06, 2024, Pisa, Italy

*Corresponding author.


†These authors contributed equally.

✉ raffaele.manna@unior.it (R. Manna); mpdibuono@unior.it

(M. P. d. Buono); giordanoluca.uni@gmail.com (L. Giordano)

ORCID 0009-0006-6285-8557 (R. Manna); 0009-0009-9372-3323

(M. P. d. Buono); 0009-0002-3048-4408 (L. Giordano)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License

Attribution 4.0 International (CC BY 4.0).

¹GUIDELINE ON THE READABILITY OF THE LABELLING AND PACKAGE LEAFLET OF MEDICINAL PRODUCTS FOR HUMAN USE - European Commission (2009).

https://health.ec.europa.eu/document/download/d8612682-ad17-40e3-8130-23395ec80380_en

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 [12].

To delve deeper into this topic, within the CALAMITA campaign [13], we introduce DIMMI (Drug InforMation Mining in Italian), a challenge centered on evaluating the proficiency of LLMs in extracting drug-specific information from Italian PILs.

By this, the task aims at contributing to the development of AI systems for enhancing drug information extraction and pharmacovigilance efforts, specifically for the Italian language.

2. DIMMI

As DIMMI seeks to advance the understanding of LLM effectiveness in processing complex medical information in Italian, participants are provided with the complete leaflets for each drug and the objective is to develop models capable of accurately answering specific questions related to a drug, such as its dosage, usage, etc.

The challenge should be approached as an information extraction task through a zero-shot mode, purely based on the model pre-existing knowledge and understanding or through in-context learning (Retrieval-Augmented Generation (RAG) or few-shot mode). The answers generated by the models will be compared against the gold standard (GS), created to establish a reliable, accurate, and comprehensive set of answers against which participant submissions can be evaluated. For each drug and each information category (e.g., dosage, usage, side effects, drug-drug interactions), the GS contains the correct information extracted from the leaflets, manually annotated according to some categories described in Section 4.1.

3. Data description

3.1. Origin of data

The challenge dataset is derived from the D-LeafIT Corpus [14], available on GitHub², made up of 1819 Italian drug package leaflets. The corpus has been created extracting PILs available on the Italian Agency for Medications (Agenzia Italiana del Farmaco - AIFA) website³, among which 1439 refer to generic drugs and 380 to class A drugs.

In the original corpus, 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. The DIMMI dataset

represents a subset of 600 entries randomly selected from the D-LeafIT corpus.

It is worth stressing that the information is extracted from pdf files and converted into texts, this means that some errors and typos may occur. Furthermore, the original D-LeafIT presents some data noise, e.g., the presence of paratext, and wrong encoding from pdf files. To fix these issues, we perform a cleaning procedure as a pre-processing phase, to obtain the final dataset. The procedure is mainly automatic and based on recurrent patterns, so that some of the aforementioned issues could be still present. The dataset pre-processing phase can be summarized in two main steps, that are:

- Correcting the separation of each leaflets by identifying regular patterns which indicate the beginning/end of a unique leaflet.
- Removing additional information about the issue date, the pharmaceutical company, and the marketing authorization.

Additionally, we notice the presence of several cases of duplicate entries, due to different reasons, as described below:

1. Same drug name, same dosage form, same ingredient amount, **different issue dates** → These cases indicate that the leaflet has been updated and all the versions are recorded into the AIFA repository. In such cases, on the basis of their ID, **the less recent leaflet has been removed**.
2. Same drug name, same dosage form, **different ingredient amount** → These cases may present, or not, the same information leaflets. **We do not remove the duplicate entries**, even though they present the same information about the classes we are interested in.
3. Same drug name, **different dosage form**, same ingredient amount → **These duplicates are not removed** as dosage information can be differentiated on the basis of the drug form.
4. Same drug name, same dosage form, same ingredient amount, **different pharmaceutical company** - These duplicates are removed and just one entry is kept. We usually prefer **keeping the one reporting in the name 'DOC generici'**. If this is not possible, we keep the first occurrence.

3.2. Data format

The whole leaflets are provided in the dataset, so that the context is available. Additionally we provide the drug name for each leaflet. The final dataset, released⁴ as a .tsv (tab-separated values) format, contains four columns. For

²<https://github.com/unior-nlp-research-group/D-LeafIT>

³<https://www.aifa.gov.it/en/home>

⁴<https://huggingface.co/datasets/RafaMann/DIMMI>

ID	ID_Loc	Drug_Name	Text
119	119_276	BOTAM	BOTAM 0,4 mg capsule (...) Tamsulosina cloridrato Medicinale equivalente (...).

Table 1

Example of a DIMMI entry

each entry we present an ID, an ID_LOC which indicates the id location in the original corpus, the drug name (without any reference to the ingredient amount, the dosage form, and the pharmaceutical company), and the leaflet text (Table 1).

Participants in the DIMMI challenge are required to use LLMs to extract the following information from the PILs text: 'Molecule', 'Usage', 'Dosage', 'Drug Interaction', and 'Side Effect'. These information must be provided as output in a structured format such as TSV or JSON, with reference to each ID and drug name contained in the evaluation dataset. The information extracted for each ID and drug name with reference to 'Molecule', 'Usage', 'Dosage', 'Drug Interaction', and 'Side Effect' must be represented in the form of a list of strings (see Section 4.2).

The evaluation dataset for the DIMMI corpus contains columns for the following entity types: 'Molecule', 'Usage', 'Dosage', 'Drug Interaction', and 'Side Effect'. For each instance (drug leaflet) in the DIMMI corpus, these entity-specific columns are populated with a list of strings, representing the annotated entities of the corresponding type.

The 'Molecule' column will contain a list of the unique molecular entities mentioned in the text, while the 'Usage' column will include a list of the specific uses or indications for the drug. The 'Dosage' column will hold a list of the textual spans describing the dosage, administration, or regimen information. The 'Drug Interaction' column will contain a list of the potential interactions with other drugs, and the 'Side Effect' column will include a list of the adverse effects associated with the drug.

3.3. Prompting

For each drug in the dataset, we evaluate the results from two types of zero-shot prompts in Italian, i.e., specific task-focused prompts and structured prompts.

The former type is composed of five questions for each of the information type we want to extract, as reported below⁵:

1. *Qual è la molecola di {drug_name}?* (What is the molecule of {drug_name}?) - to extract the molecule

⁵It is worth stressing that in the prompt examples {drug_name} is not a masked word, it represents a placeholder to indicate one of the entries from the column drug_name in DIMMI dataset.

2. *Per cosa si usa {drug_name}?* (What is {drug_name} used for?) - to extract the usage
3. *Qual è la posologia raccomandata per {drug_name}?* (What is the recommended dosage for {drug_name}?) - to extract the dosage
4. *Quali sono gli effetti collaterali di {drug_name}?* (What are the potential side effects of taking {drug_name}?) - to extract side effects
5. *Con quali medicinali interagisce {drug_name}?* (What are the drug interaction of {drug_name}?) - to extract the interaction with other drugs

The latter type of prompt aims at extracting all the relevant information with a specific instruction to help the model understand the expected output structure and facilitates extraction as it follows:

- *Fornisci le seguenti informazioni su {drug_name}:
Molecola:
Uso:
Posologia:
Effetti collaterali:
Interazioni con altri medicinali:
(Provide the following information about {drug_name}:
Molecule:
Usage:
Dosage:
Side Effects:
Drug interaction:)*

3.4. Dataset statistics

As mentioned before, the final dataset is composed by 600 unique PILs in Italian, providing a comprehensive dataset for the challenge. The documents in the DIMMI dataset exhibit a wide range of lengths (Table 2), with the shortest document containing 363 tokens and the longest extending to 11,730 tokens. This range in token count directly corresponds to the word count, indicating that each word is treated as a single token in this analysis. On average, each document contains approximately 2,520 words, with a standard deviation of 848 words, indicating moderate variability in document length. The distribution of document lengths is further characterized by the 25th, 50th (median), and 75th percentiles, which are 1,960, 2,448, and 2,980.75 words, respectively.

In total, the corpus contains 1,511,724 words and tokens. The lexical diversity of the corpus is reflected in the

DIMMI Statistics	
num_documents	600
mean_length	2519.54
min_length	363
max_length	11730
std_length	848.41
percentiles	.25:1960, .5: 2448, .75: 2980
total_words	1511724
mean_words_per_doc	2519.54
total_tokens	1511724
min_tokens	363
max_tokens	11730
unique_tokens	58901
type_token_ratio	.038

Table 2
DIMMI statistics

58,901 unique tokens identified, resulting in a type-token ratio (TTR) of 0.0390. This relatively low TTR suggests a high degree of repetition within the text, which is typical for technical and regulatory documents such as drug package leaflets. Importantly, there are no empty documents in the corpus, ensuring that all entries contribute meaningful content to the dataset.

4. Evaluation metrics

We will evaluate the results using accuracy, precision, recall and F-1 score using a gold standard as benchmark (see Section 4.1).

The details for each metric are provided below:

- **Precision metric:** For example: Dosage: If the model extracts "200mg-400mg every 4-6 hours" and this is correct, the precision for dosage is 100%; Side Effects: If the model extracts "Stomach upset, nausea" and this is partially correct (missing other side effects), the precision for side effects might be 50% (depending on how many side effects are correctly identified);
- **Recall metric:** For example: Dosage: If the correct dosage is "200mg-400mg every 4-6 hours" and the model extracts only "200mg-400mg," the recall for dosage is 50%. Side Effects: If the correct side effects are "Stomach upset, nausea, dizziness, headache" and the model extracts "Stomach upset, nausea," the recall for side effects is 50%.
- **F1-score metric:** A balanced measure of precision and recall. A higher F1-score indicates better performance.
- **Accuracy:** The overall percentage of correct extractions across all classes. As far as this metric is concerned, we also evaluate the class-Level Accuracy, as the accuracy for each specific class separately.

4.1. Gold Standard Creation

In order to evaluate the system results, we created a gold standard (GS), manually annotating the following categories: i) molecule; ii) dosage; iii) drug interaction; iv) usage; v) side effect. For each of the aforementioned classes we define some guidelines and specifications for the annotation, as summarised in the following paragraphs.

Molecule The category is used to identify the main ingredient(s) of the drug. In some cases, the bulking agent(s) may be reported together with the molecule(s). These are not included in the `molecule` class.

Dosage information This class refers to the recommended dosage for drug administration. We do not annotate the treatment duration neither the maximum dosage in the dosage information.

For dosage information we distinguish between dosage for children and adults. We do not distinguish dosage for infants or elders (the former is annotated as dosage information for children, the latter as dosage information for adults, as reported below).

When the same dosage can be used for both adults and children, the general dosage information category is applied.

Example:

10 mg una volta al giorno negli adulti e nei bambini di età uguale o superiore ai 10 anni
(10 mg once a day in adults and children aged 10 years or older)

Furthermore, dosage information could be differentiated on the basis of age/weight. In such cases, unless dosages for adults and children are explicitly differentiated, we always use the general category dosage.

Example:

Adulti, anziani e bambini di età pari o superiore a 12 anni con un peso corporeo pari o superiore a 50 chilogrammi (kg): • da 1 a 2 g una volta al giorno a seconda della gravità e del tipo di infezione
(Adults, elderly, and children aged 12 years and older with a body weight of 50 kilograms (kg) or more: • 1 to 2 g once a day, depending on the severity and type of infection.)

Dosage for infants can be expressed through a co-reference to some other dosage, e.g., for adults or children, sometimes with a different time schedule, as in *lo stesso dosaggio sopra descritto ma somministrato una volta ogni due giorni* (The same dosage as described

above, but administered once every two days.). Unless the dosage is explicitly mentioned, we do not annotate these spans, as the information is context-dependent.

The treatment of specific diseases might require different dosages for the same drug. When they are reported in the leaflet, following a minimum span principle, we annotate all the dosages without any specification about the disease. Due to the aforementioned annotation choice, the annotation results will be a set of dosage information, as in the following example (annotated spans reported in bold face).

Example:

Aspergillosi: - 2 capsule una volta al giorno per un periodo di 2-5 mesi; (...) Candidosi: 1-2 capsule 1 volta al giorno per un periodo da 3 settimane a 7 mesi (...) Criptococcosi non meningea: 2 capsule una volta al giorno per un periodo dai 2 mesi ad 1 anno (...)

When the same dosage can be applied in more than one cases, span duplicates may be present (e.g., *2 capsule una volta al giorno*). In the final GS, these are removed so that only one span for each type is kept.

Some drugs must be administered according to a schedule that spans different time periods, with or without dosage variations. In such cases, we annotate only the initial recommended dosage.

In some cases, the posology section does not provide specific dosage information and instead includes a general recommendation to consult a doctor. In these instances, we consider the information to be missing and do not annotate the general statement.

Drug interaction As for drug interactions, we annotate the name of molecules and drugs when they are available. In some cases, the information about drug interaction is reported as a general reference to the use of some drugs (e.g., *medicinali per abbassare la pressione* - medicines to lower blood pressure). In such instances, as we cannot identify the specific molecule or drug, we annotate the general reference. Information about drug interactions may also appear as a reference to certain types of relationships with other molecules, as in *derivati della fenotiazina* (phenothiazine derivatives). For our annotations, we omit additional information and select the minimal span, in the aforementioned example, *fenotiazina* (phenothiazine).

Similarly, when the information pertains to the drug class instead of reporting the molecule, e.g., *lassativi* (laxatives), we annotate the minimal span, even though in some cases the drug use is specified, e.g., *medicinali usati per trattare la stipsi* (medicines used to treat constipation).

We apply a hierarchical priority to identify and annotate

the minimal span that conveys the information about drug interactions, as follows:

1. Molecule
2. Drug class
3. Drug use

The aforementioned hierarchy helps us identify the span to be annotated. When included, drug names are always annotated.

When the interaction information is reported with the specific pharmaceutical form (e.g., *eritromicina iniettabile*), only the minimal possible span is annotated, i.e., *eritromicina*.

In some cases, examples of interacting molecules or drug names are provided alongside the drug class or use (e.g., *medicinali usati per il trattamento dell'HIV/AIDS, per esempio ketoconazolo e itraconazolo* - medicines used for the treatment of HIV/AIDS, for example, ketoconazole and itraconazole). In these instances, we annotate both, as the list of drugs and molecules may not be exhaustive. If the list is exhaustive, we do not annotate the general reference to the drug use; we only annotate the drug molecules or names.

Interactions with some other molecules can be conditioned by the taken amount, e.g. *cimetidina, preso in dosi giornaliere superiori a 800mg* (cimetidine, taken in daily doses greater than 800 mg). Also in these cases the molecule name is the only span annotated.

Some interacting drugs are reported as the general drug class, together with a plain language explanation and a subclass specification, as in the following example *diuretici (compresse per urinare in particolare quelli chiamati risparmiatori di potassio) (diuretics (tablets for urination, particularly those called potassium-sparing))*. As the molecule is not noted, we do annotate both the general class and subclass (both in bold face in the previous excerpt).

Additionally, also food and beverage can interact with drugs, e.g., *pompelmo, alcol* (grapefruit, alcohol). We opt not to include these substances within the drug interaction class, as we want to focus only on the pharmaceutical drug interaction.

Drug interaction information are considered missing when there is only a general sentence to the fact that the use of any further drug should be reported.

Usage With respect to usage, we consider the minimal possible span, which indicates the disease treated by the specific drug. Thus, for instance, in the sentence {drug_name} è usato nel trattamento della gotta ({drug_name} is used in the treatment of gout), we annotate only *gotta* (gout).

In other cases, some examples of usage may be reported as in *traumi (ad esempio causati dallo sport)* (injuries (for example, those caused by sports)). As those cases are not

representative enough of usages, we do not include them in the annotation, so in the previous excerpt we annotate just *traumi* (injuries).

Within the usage section, sometimes the use of plain text is reported together with reference to the specific disease, e.g., *meningite cirptococcica - un'infezione micotica del cervello* (...). We always annotate the specific term for the disease and discard the plain text description.

When the generic disease class is presented, e.g., *infezioni cutanee* (skin infections), followed by a non exhaustive list of examples, we annotate just the generic use.

Side effects This class indicates all the possible side effects caused by the drug consumption. In PILs, this type of information is generally grouped on the basis of the number of people affected by the side effects to identify different diffusion levels, e.g., very common side effects, very rare side effects. We do not differentiate among the diffusion levels and consider all the side effects belonging to the same class *side_effect*. In some cases, side effects affecting other subjects than the person consuming the drug are reported. For instance, some drugs can affect the fetus as in the following excerpt.

Example:

(...) *Se assume Ricap durante le ultime fasi della gravidanza, il suo bambino potrebbe manifestare i seguenti sintomi: problemi a respirare, colorito bluastrò o violaceo della pelle, convulsioni* (...).

[(...) If you take Ricap during the later stages of pregnancy, your baby may experience the following symptoms: breathing problems, bluish or purplish skin discoloration, seizures (...)]

We do not annotate these secondary side effects and the ones derived from drug overdose.

When the side effect type is reported together with its symptoms we do include those within the class of side effects. For instance, in some cases a list of symptoms *difficoltà respiratoria, riduzione della pressione sanguigna* is combined with the general side effect *reazioni allergiche*. Each of them is annotated separately and included into the list of side effects.

Similarly, we annotate both the plain language side effect and the term, as in *problemi del flusso della bile (colestasi)* (bile flow problems (cholestasis)).

When the side effects are reported as worsening of an already existing disease, e.g., *umentata perdita di capelli*, we annotate the minimum possible span, i.e., *perdita di capelli*.

For drugs containing more than one molecule, side effects are reported along with the side effects for each individual molecule. We annotate all of them.

Side effects can be reported with reference to some

patient/disease type, e.g., *Se è HIV positivo può mostrare effetti indesiderati* (If you are HIV positive, you may experience side effects). In such cases, symptoms are annotated without any further specification.

If duplicates are presented, those are not annotated or removed in the post-processing phase, so that just one entry for symptom type is recorded in the GS.

Sometimes, side effects are grouped by indicating the general area (e.g., organ or functionality) affected, e.g., nervous system disorders. The information might be followed by a list of specific side effects. When this is the case, we discard the general information in favor of the most specific one.

It is worth stressing that other information may be presented in PILs, for instance Precautions for use. As we are not interested in this type of information, we do not annotate such sections.

Inter-Annotator Agreement The annotation has been performed by three people with computational linguistic backgrounds and different levels of expertise. An initial inter-annotator agreement has been evaluated after the first draft of guidelines has been created. Borderline cases and issues have been collected by each of the annotators and subsequently discussed and solved. The guidelines have been updated accordingly and a second round of annotation has been performed in order to compute the final inter-annotator agreement.

The annotation round for evaluating the final inter-annotator agreement has been performed on a subset of 60 leaflets.

The results, calculated before the post-processing phase, show a complete agreement on the molecule class among all the annotators, while for the remaining classes the agreement spans from .61 for posology and .80 for side effects (Table 3).

Class	A1/A2	A1/A3	A2/A3	AVG
Molecule	1	1	1	1
Usage	.69	.67	.68	.68
Posology	.61	.62	.66	.63
Drug interaction	.66	.66	.65	.66
Side effects	.80	.76	.75	.78

Table 3
IAA for the GS

To assess the inter-annotator agreement (IAA) for the creation of the gold standard, we employed two different metrics: pairwise F1 score [15, 16] and token-level agreement percentage [17]. The pairwise F1 score was used to calculate the IAA for the "Molecule" and "Usage" labels, as the information contained in the text for these entities refers to unique and well-defined concepts. This metric provides a balanced measure of the precision and recall of the annotations, allowing us to quantify the level of

agreement between annotators on the identification of these specific entities.

On the other hand, for the "Dosage", "Drug Interaction", and "Side Effect" classes, we opted to use the token-level agreement percentage as the IAA metric. This choice was motivated by the fact that these classes involve variable text spans, which can be more challenging to align between annotators. Before calculating the token-level agreement percentage, we performed preprocessing steps on the annotated portions, removing punctuation marks (such as - and • that indicate a list) and Italian stopwords from the Spacy Italian language model⁶. The token-level agreement percentage provides a more granular assessment of the consistency in the identification of the relevant text segments, which is crucial for the accurate extraction of these types of entities from the source documents.

GS Post-processing To ensure high consistency among annotations and to remove additional information that does not meet the specified annotation criteria, we perform a post-processing step. During this phase, we review the GS, using recurring patterns and regular expressions to clean the data and correct errors. We also carry out manual cleaning to produce the final GS.

For instance, when applicable, we remove the drug name mentioned in the posology specification (e.g., one tablet of drug_name once a day) so that only the general information related to the molecule is retained.

The resulting evaluation dataset contains XXX annotated molecules, XXX drug interactions, XXX usage information, and XXX side-effects (Table 4).

Class	Tot. Entities	Unique Entities
Molecule	657	657
Usage	2159	2113
Posology	831	827
Drug interaction	8617	8458
Side effect	36748	30313
Total	49012	42368

Table 4

Annotated entities for each class

4.2. Results

The expected results should be presented as a list of entities for each of the classes of information about each drug. To obtain the result lists, we consider the annotated terms and their simplifications as unique entities e.g., the span *livelli aumentati di calcio nel sangue (ipercalcemia)* (elevated levels of calcium in the blood (hypercalcemia)) is listed as two separate entities that are *livelli aumentati*

⁶https://spacy.io/models/it#it_core_news_lg

di calcio nel sangue and *ipercalcemia*.

This choice aims at accounting for both entities as possible correct answers.

For instance, for the drug **NATRILIX**, the expected results are as it follows:

- Usage: *pressione sanguigna elevata, ipertensione arteriosa essenziale*
- Molecole: *indapamide*
- Dosage: *1 compressa al giorno*
- Side_effect: *eruzioni cutanee, bassi livelli di potassio nel sangue, vomito, porpora ...*
- Drug_interaction: *litio, chinidina, idrochinidina, disopiramide (...)*

For the drug **Trevid**, the correct answers would be:

- Usage: *carezza di vitamina D*
- Molecole: *colecalfiferolo*
- Dosage: *3-4 gocce al giorno*
- Side_effect: *livelli aumentati di calcio nel sangue, ipercalcemia, livelli aumentati di calcio nelle urine, ipercalcemia, debolezza, astenia, reazioni allergiche, appetito ridotto (...)*
- Drug_interaction: *anticonvulsivanti, barbiturici, colestipolo, colestiramina, orlistat (...)*

Since this is an information extraction task in a zero-shot setting based on PILs, it is expected that LLMs will be able to extract the exact terminology used in the different sections of the PILs and provide a list of terms. The performance will be evaluated based on the metrics described in 4. Potential limitations in accurately assessing the performance of LLMs may arise from: 1) the variability in the models' choice of terms to extract, and 2) the provision of terms and their simplifications as two entities. In these cases, forcing the LLMs to provide a more structured and less ambiguous output might help, as currently the gold standard does not account for a set of synonyms to handle variability in the output, or employing additional metrics to address the second case.

5. Limitations

One important limitation of the DIMMI dataset is the disclaimer provided by the Italian Medicines Agency (AIFA) regarding the content available on their website in section A. Disclaimer⁷. AIFA states that all the information and services offered on their website are provided "as is" and "with all faults". The Italian Medicines Agency, therefore, does not provide any kind of warranty, either explicit or implied, regarding the content, including, without limitation, the legality, ownership, suitability, or fitness for particular purposes or uses.

⁷<https://www.aifa.gov.it/en/copyright>

This disclaimer from the data source raises concerns about the reliability and quality of the patient information leaflets (PILs) that were used to construct the DIMMI corpus. While the dataset has been carefully curated and annotated, the underlying data may contain errors, inaccuracies, or other issues that are not explicitly acknowledged by the original provider. Researchers and developers using the DIMMI dataset should be aware of this limitation and exercise caution when relying on the information contained within the corpus, particularly for critical applications or decision-making processes.

6. Ethical issues

Ethical considerations are crucial when working with a dataset that contains sensitive information from PILs. The DIMMI corpus, which is derived from the AIFA (Italian Medicines Agency) Database, must be handled with the utmost care and respect for individual privacy, data protection, and the diversity of the target population.

Additionally, the use of the DIMMI corpus for the development and evaluation of natural language processing models must be guided by ethical principles that consider the diversity of the target population. The models trained on this data should be designed and deployed in a way that respects individual privacy, avoids potential misuse or discrimination, and ultimately benefits the public good, regardless of ethnicity or age. Careful consideration should be given to the potential societal impact of the applications built upon the DIMMI dataset, ensuring that they are inclusive and equitable.

By upholding the ethical standards in the handling and utilization of the DIMMI corpus, the research community can ensure that the valuable pharmacological information contained in the PILs is leveraged responsibly and in a manner that prioritizes the well-being of patients and the general public, while respecting the diversity of the target population.

7. Data license and copyright issues

The DIMMI corpus has been created using the patient information leaflets (PILs) from the AIFA (Italian Medicines Agency) Database. As reported in the Web site⁸, the distribution license used by AIFA for these data is the Creative Commons Attribution (CC-BY) license, version 4.0. This license allows third parties to distribute, modify, adapt, and use the data, even for commercial purposes, with the sole requirement of providing attribution to the original source.

⁸<https://www.aifa.gov.it/en/copyright>

By making the DIMMI corpus available under the CC-BY 4.0 license, the dataset can be freely accessed, utilized, and built upon by the scientific community, contributing to the advancement of research and applications in the field of biomedical text mining and pharmacological information extraction.

Acknowledgments

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

References

- [1] W. H. Shrank, J. Avorn, Educating patients about their medications: the potential and limitations of written drug information, *Health affairs* 26 (2007) 731–740.
- [2] P. Rodríguez, R. Azarola, S. Lorda, B. Cantalejo, A. Danet, et al., Quality improvement of health information included in drug information leaflets. patient and health professional expectations, *Atencion primaria* 42 (2009) 22–27.
- [3] M. Á. Piñero-López, P. Modamio, C. F. Lastra, E. L. Mariño, 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 (2016) e100.
- [4] I. Segura-Bedmar, P. Martínez, Simplifying drug package leaflets written in spanish by using word embedding, *Journal of biomedical semantics* 8 (2017) 1–9.
- [5] M. Yuan, P. Bao, J. Yuan, Y. Shen, Z. Chen, Y. Xie, J. Zhao, Y. Chen, L. Zhang, L. Shen, et al., Large language models illuminate a progressive pathway to artificial healthcare assistant: A review, *arXiv preprint arXiv:2311.01918* (2023).
- [6] A. B. Abacha, E. Agichtein, Y. Pinter, D. Demner-Fushman, Overview of the medical question answering task at trec 2017 liveqa., in: *TREC, 2017*, pp. 1–12.
- [7] A. B. Abacha, Y. Mrabet, M. Sharp, T. R. Goodwin, S. E. Shooshan, D. Demner-Fushman, Bridging the gap between consumers' medication questions and trusted answers, in: *MEDINFO 2019: Health and Wellbeing e-Networks for All*, IOS Press, 2019, pp. 25–29.
- [8] V. Nguyen, S. Karimi, M. Rybinski, Z. Xing, 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), 2023, pp. 629–648.
- [9] M. Filannino, Ö. Uzuner, Advancing the state of the art in clinical natural language processing through shared tasks, *Yearbook of medical informatics* 27 (2018) 184–192.
- [10] R. Vaishya, A. Misra, A. Vaish, Chatgpt: Is this version good for healthcare and research?, *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 17 (2023) 102744.
- [11] P. Lee, S. Bubeck, J. Petro, Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine, *New England Journal of Medicine* 388 (2023) 1233–1239.
- [12] S. Gilbert, H. Harvey, T. Melvin, E. Vollebregt, P. Wicks, Large language model ai chatbots require approval as medical devices, *Nature Medicine* 29 (2023) 2396–2398.
- [13] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, CALAMITA: Challenge the Abilities of LAnguage Models in ITAlIAn, in: *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, Pisa, Italy, December 4 - December 6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.
- [14] L. Giordano, M. P. Di Buono, Large language models as drug information providers for patients, in: *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health)@ LREC-COLING 2024*, 2024, pp. 54–63.
- [15] G. Hripcsak, A. S. Rothschild, Agreement, the f-measure, and reliability in information retrieval, *Journal of the American medical informatics association* 12 (2005) 296–298.
- [16] L. Deleger, Q. Li, T. Lingren, M. Kaiser, K. Molnar, L. Stoutenborough, M. Kouril, K. Marsolo, I. Solti, et al., Building gold standard corpora for medical natural language processing tasks, in: *AMIA Annual Symposium Proceedings*, volume 2012, American Medical Informatics Association, 2012, p. 144.
- [17] C. Grouin, S. Rosset, P. Zweigenbaum, K. Fort, O. Galibert, L. Quintard, Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview, in: *Proceedings of the 5th linguistic annotation workshop*, 2011, pp. 92–100.