TrackList: Tracing Back Query Linguistic Diversity for Head and Tail Medical Knowledge in Open Large Language Models

Ioana Buhnila[♪] **Aman Sinha**[♡♪♫] **Mathieu Constant**[♪]
[♪]ATILF, CNRS, Université de Lorraine, Nancy, France
[♡]IECL, University of Lorraine, Nancy, France
[♫]Institut Strauss, Strasbourg, France
Correspondence: firstname.lastname@univ-lorraine.fr

Abstract

While humans can easily produce various types of answers, such as definitions, examples or paraphrases, Large Language Models (LLMs) struggle to provide correct answers to medical questions that require diverse answer formats. In this paper, we introduce TrackList, a fine-grained linguistic and statistical analysis pipeline to investigate the impact of the pretraining data on LLMs answers to diverse linguistic queries. We also propose RefoMed-EN, a medical dataset consisting of 6,170 human-annotated medical terms alongside their corresponding definitions, denominations, exemplifications, explanations, or paraphrases. We investigated whether the high or low frequency of a concept (head or tail knowledge) impacts the language model's performance for answering medical questions. We evaluated the quality of the LLM's output using syntactic and semantic similarity metrics, statistical correlations and embeddings. Results showed that the LLM's answer quality for definition-type questions is the highest, while for the exemplification-type being the lowest. Additionally, we showed that for definition-type medical questions ("What is multiple sclerosis?"), LLMs are prone to paraphrase more for popular medical concepts, and less on more specialized medical knowledge.

1 Introduction

As Large Language Models are becoming widely used in many question-answering tasks and are applied to different fields, such as health and medicine, the need to explain and interpret the generated responses has become crucial. In this paper, we evaluated the most widespread real world use-case of LLMs, which is open question-answering, meaning the model gives a free text answer to a user question (Shailendra et al., 2024). LLMs are most efficient in question-answering tasks that cover popular concepts (head knowledge) (e.g. *cancer*, *asthma*) (Mallen et al., 2023; Li et al., 2024a)

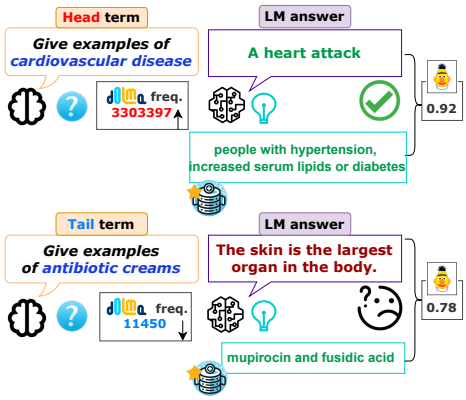

Figure 1: Language Models (LMs) tend to generate more hallucinated *definition-style* outputs even when directly asked to answer an *example-style* query for tail knowledge terms.

and when giving definitions to concepts. However, their performance drops when tested on torso or tail knowledge (*ideatory apraxia* or *erythematous angina*) (Sun et al., 2023; Kandpal et al., 2023) or when tackling complex questions (Daull et al., 2023). Our study analyzed the impact of query linguistic diversity on QA performance in the medical field, focusing on five different types of questions according to pragmatic functions of human communication (definitions, exemplifications, explanations, denominations and paraphrases). We demonstrated that LLMs perform worse when asked to give different types of answers, such as examples, to tail (or rare) medical knowledge (see Figure 1).

As black box and proprietary models such as ChatGPT showed limitations in terms of explainability and reproducibility (Zhao et al., 2024a; Liesenfeld et al., 2023; Ravichandran et al., 2024), more research should be conducted on open-source language models, to foster a deeper understanding of how LLMs process language and knowledge, especially in sensitive fields such as medicine. This study explored non-proprietary and open-source Large Language Models, OLMo (Groen-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 413–425

eveld et al., 2024) and Pythia (Biderman et al., 2023). We chose these specific language models because they are fully open, thus fostering interpretability and reproducibility. These LLMs were released together with model weights, inference code, training/evaluation code, and pretraining corpus, DOLMA (Soldaini et al., 2024) for OLMo, and The Pile (Gao et al., 2020) for Pythia.

The research questions we investigated are the following: **(RQ1)** *Does the frequency of a medical concept in the pretraining data influence a language model's answer quality?* **(RQ2)** *How does the linguistic diversity and complexity of questions impact the language model ability to give accurate answers?* **(RQ3)** *To what extent are language models paraphrasing the data from their parametric memory for head and tail knowledge?* Based on previously research results on the general domain (Wang et al., 2024), we investigated whether frequent medical terms (head) lead to distributional memorization, and less frequent medical terms (tail) result in lower downstream task performance due to reliance on parametric knowledge.

The contributions of our paper are the following:

1. We proposed a fine-grained analysis and evaluation pipeline, **`TrackList`**, to **`tra`**ce ba**`ck`** query **`li`**nguistic diver**`s`**ity of head and **`t`**ail knowledge in Open Large Language Models. The pipeline is useful for evaluating LLMs' answers on various query types according to word frequency in the pretraining dataset. We will share the `TrackList` code with the NLP community to ensure reproducibility and to foster new research on LLMs' linguistic abilities for open QA tasks[1].

2. We share `RefoMed-EN`, an English dataset of 6,170 human annotated medical terms together with their context, discourse markers, corresponding paraphrases, lexical and pragmatic functions. The dataset is a non-contaminated benchmark extracted from human written texts and it is shared with open license together with the `TrackList` pipeline.

3. We conducted a detailed analysis of the LLMs' performance according to the linguistic complexity of the user question. We showed that LLMs have limited understanding of linguistic differences in queries, as they are more prone to giving definition-style answers to most queries, even when explicitly asked to output an example or a paraphrase for a medical concept.

[1]https://github.com/ATILF-UMR7118/TrackLIST

2 Related Work

Methods such as In-Context Learning (ICL) (Dong et al., 2024), Chain-of-Thought (CoT) prompting (Wu et al., 2023), and Retrieval Augmented Generation systems (RAG) (Lewis et al., 2020) have contributed to further interpret the LLM's generated response, and help mitigate hallucinations (Akbar et al., 2024; Huang et al., 2024; Asai et al., 2023). In the medical domain, LLM's output needs to be correct, therefore methods that combine ICL and RAG showed promising results, such as improving the accuracy of GPT4-Turbo's answers to oncology questions from 62.5 to 83.3% (with ICL) and 79.2% of questions (with RAG) (Iivanainen et al., 2024). Moreover, QA frameworks like Self-BioRAG that answer biomedical questions by reflecting and retrieving relevant documents showed a 7.2% improvement on average over the state-of-the-art open Small Language Models (SLM) of 7b or less (Jeong et al., 2024).

Large medical QA datasets like MedQuAD (Ben Abacha and Demner-Fushman, 2019) and MEDIQA (Abacha et al., 2019) gather different types of medical data, such as treatment, symptoms, definition, susceptibility or prevention. These datasets are built for medical experts, more than for the general public interested in understanding medical concepts (Nguyen et al., 2023). Moreover, these datasets are shared in a XML format, not allowing easy implementation for LLMs prompts. Furthermore, the dataset is divided in more medical than linguistic criteria. Therefore, our dataset, `RefoMed-EN`, is filling a gap in the benchmark leaderbord for query types in the medical field.

While evaluating LLMs answers to open questions is difficult, recent studies showed that human annotation is still needed when the generated answer does not match a gold standard (Kamalloo et al., 2023). In our work, we combine both human evaluation and automatic evaluation through similarity metrics that we further develop in Section 3.2. Determining what is head and tail knowledge is a research question in itself, as recent studies investigated this in detail (Mallen et al., 2023; Li et al., 2024b). We considered that the word frequency in the pretraining data as the popularity metric for our

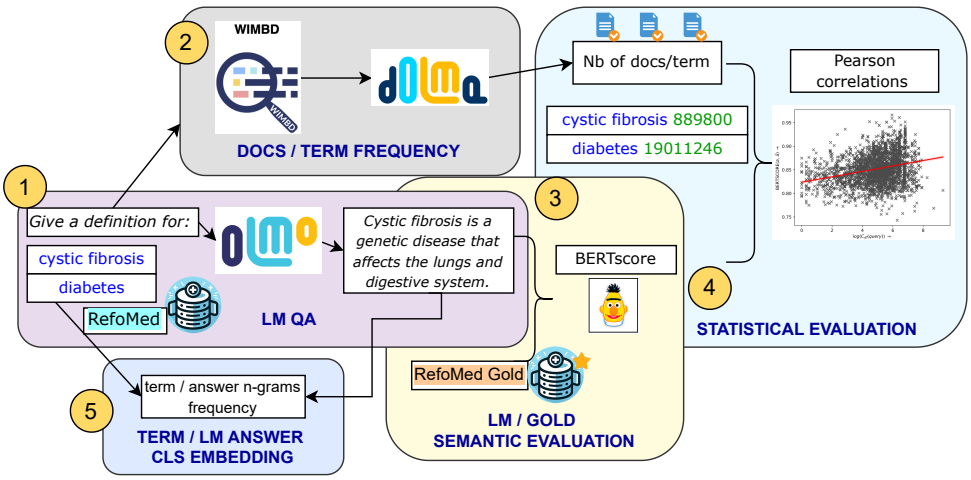

Figure 2: The pipeline of our method represented in five steps. 1) The zero-shot inference QA task using the medical concepts from the RefoMed-EN dataset. The dataset was divided in subdatasets according to the query type (detailed presentation in section 3.4). 2) We obtained the frequency in terms of number of documents for each RefoMed-EN concept. 3) We calculated the BERTScore between the LMs output and the RefoMed-EN gold standard. 4) These two values were used to compute Pearson correlations. 5) We computed a probability metric between the CLS embedding score between the term and the n-grams of the output, and their frequency in the pre-training corpora.

medical dataset.

Recent studies investigated whether LLMs memorize or generalize knowledge by exploring open language models. Wang et al. (2024) introduced the concept of distributional memorization, to measure the correlation between the LLM output probabilities and the frequency of the pretraining data. The authors evaluated the task performance with 3 to 5 n-grams search into The Pile (Gao et al., 2020), the pretraining corpus of the language model Pythia (Biderman et al., 2023). They found that LLMs generalize more in reasoning-intensive tasks, while they memorize more in simpler knowledge-intensive tasks (Wang et al., 2024). We test this hypothesis on the medical domain, to evaluate how knowledge specific concepts impact the LLMs answers' linguistic diversity.

Concurrent work investigated the concept of linguistic diversity in LLMs answers by evaluating their lexical, syntactic and semantic distribution compared to human linguistic richness (Guo et al., 2024). In our study, we investigated the linguistic complexity of LLM queries from a syntactic, semantic and pragmatic perspective, applied to a knowledge intensive task, medical QA. We present our method below.

3 Methodology

We illustrate our method in Figure 2. We conducted our experiments using four fully open language models, OLMo-1b, OLMo-7b, OLMo-7b-instruct, and Pythia-1b. We used the WIMBD tool's API (Elazar et al., 2023) to access the pre-training corpora DOLMA (for OLMo) and the `infini-gram` library (Liu et al., 2024) for The Pile dataset (for Pythia). We counted the number of documents that contain a certain medical concept, thus determining its frequency. We present our detailed rationale in the section below.

3.1 Formulation

We utilized a language model $\mathbb{M}_{\text{LLM}}$ and assumed access to a corpora C with $\mathcal{D}$ documents which was used to train $\mathbb{M}_{\text{LLM}}$. We represented the corpora C by a set of unique words $\{w_1, w_2, w_3, ...w_\infty\}$ present in at least one document $d \in \mathcal{D}$. Each of the document $d \in \mathcal{D}$ can be similarly represented by a subset of C as $\{dw_1, dw_2, dw_3, ...dw_\infty\}$ where dw_i represents word w_i present in document d. We defined a document-frequency count operator over corpora C and denote it by $\mathbf{C}_{df}$. This enabled us to count the number of documents present in corpora C which contained the provided term at least once[2].

We further utilized query-answer (Q, A) pairs where a question denoted by $q \in Q$ can be of different types as mentioned in Table 1 and an reference answer $a \in A$ which is borrowed from the RefoMed-EN dataset. In our experiment setup, we provides as input $q \in Q$ to the language model as

[2]For example, $\mathbf{C}_{df}$("*infection*" | C) outputs 543435456, which implies that there are 543435456 documents in C containing the term *infection* at least once.

follows:

$$\hat{a} = \mathbb{M}_{\text{LLM}}(q) \quad (1)$$

In the above equation, we marked the generated response from the $\mathbb{M}_{\text{LLM}}$ as $\hat{a}$. During evaluation, we considered different evaluation metrics (denoted by Ψ) to calculate the correctness (denoted by E) of the generated response $\hat{a}$ with respect to the gold reference $a \in A$, as follows:

$$E = \Psi(\hat{a}, a) \quad (2)$$

Lower value of E, for example, in case of BERTscore, implies less relevance of generated answer and vice versa.

3.2 Evaluation Metrics

As the goal of our study is to advance the interpretability of open large language models by describing their behavior, we chose simple and easy to interpret metrics that allowed us to analyze the relationship between query term frequency, gold answers, and generated LLM answers.

Tracing back semantic similarity. We calculated the BERTscore (Zhang et al., 2019) between the generated answers and the gold standard from the RefoMed-EN dataset.

Statistical correlation. We computed the Pearson correlation metrics to compare the BERTscore calculated before with the term's frequency in the pretraining corpus.

CLS cosine similarity. We computed the cosine similarity of sentence embeddings to compare the semantic similarity between the medical term and the n-grams of the generated answer.

3.3 Linguistically Annotated Dataset

We constructed RefoMed-EN by automatically translating RefoMed (Buhnila et al., 2024) from its original version in French to English using a licensed DeepL Translator API. RefoMed is a French annotated dataset of 6,170 annotated medical terms with their corresponding reformulations. The dataset is comprised of paraphrases and reformulations semi-automatically extracted from scientific and popularization medical texts and abstracts from the ClassYN (Todirascu et al., 2012) and CLEAR (Grabar and Cardon, 2018) French medical corpora. The dataset required extensive pre-processing as there were formatting and translation inaccuracies from the original annotation. We chose to translate this dataset because there is no freely available benchmark annotated on question types following pragmatic and linguistic theories for the medical field (to the best of our knowledge). More information about the different types of linguistic annotations can be found in Appendix A. The translation of a French dataset assures that there is no benchmark contamination between the pretraining corpora and the test dataset (Sainz et al., 2023; Li et al., 2024c). We share this new non-contaminated benchmark with the NLP community on GitHub. We present below the question types explored in this study.

3.4 Query Linguistic Diversity

Linguistically, the concept of *reformulation* is defined as textual and discursive act performed with a precise objective (Grabar and Eshkol, 2016). *Reformulations* have a well defined pragmatic role by expressing a content in a different semantic or lexical representation, adapted to a specific audience and communicational need, like science popularization or education. The linguistic diversity is correlated to the pragmatic usage of language in humans, such as asking for a definition, an explanation, reformulation or paraphrase of a concept, or to receive examples of a certain concept (Grabar and Eshkol, 2016).

In this work, we analyzed the role of reformulations in the case of medical knowledge popularization for laypeople and patients (Grabar and Eshkol, 2016). We focused on a knowledge intensive question-answering task taking into account the five most common types of questions that require reformulation processes, as shown with examples in Table 1. We present the number of questions by type from the RefoMed-EN dataset in Table 2.

4 Experimental Setup

This section presents the experiments we conducted to evaluate the generated text according to query types and term frequency. Experiments were done on a P100 NVIDIA GPU, for an individual runtime of $\leq$ 15 hours including different steps involved.

4.1 Linguistic Diversity and Frequency for Task Performance

We explored the impact the query linguistic diversity has on the quality of the generated answer. We evaluated the task performance of OLMo and Pythia models in a zero-shot QA setting by conducting several experiments:

Pragmatic function	Definition	Query type	Example
`Definition` (**DEF**)	a difficult or technical concept is defined to ease comprehension	Give a definition for / What is *multiple sclerosis*	Multiple sclerosis (MS) is a *disease of the nervous system of immune origin*
`Exemplification` (**EX**)	the meaning of a concept is illustrated through examples of types and subtypes of entities	Give examples of *heart and vascular diseases*	Heart and vascular diseases include *heart attacks, angina, stroke, sudden cardiovascular death and the need for heart surgery*
`Denomination` (**DEN**)	a concept is reformulated through a semantically similar concept, without simplification	Give another denomination for *motor neuron disease*	Pharmacotherapy for pain management in *amyotrophic lateral sclerosis* (motor neuron disease)
`Paraphrase` (**PARA**)	the concept is reformulated through a easy to understand semantically similar synonym	Give a paraphrase for *hypotension*	Hypotension, i.e. *low blood pressure*, frequently occurs in newborns
`Explanation` (**EXP**)	the concept is explained through its process or a part of it	Give an explanation for *autoimmune diseases*	These are autoimmune diseases that can be explained by the fact that *the organism produces an antibody against the person's skin*

Table 1: Query types, definitions and examples according to pragmatic functions annotated on the RefoMed-EN corpus. These queries were used for prompting in our QA task.

	DEF	EX	DEN	PARA	EXP
`ClassYN EX`	207	305	131	30	32
`CLEAR EX`	919	379	401	73	67
RefM-EN-EX	*1126*	*684*	*532*	*103*	*99*
`ClassYN GP`	862	343	235	285	149
`CLEAR GP`	883	470	175	124	100
RefM-EN-GP	*1745*	*813*	*410*	*409*	*249*
RefM-EN	**2871**	**1497**	**942**	**512**	**348**

Table 2: Distribution of question types across subcorpora in RefoMed-EN (RefM-EN). EX denotes expert-oriented texts; GP targets the general public. Question types are ordered by frequency (left to right).

- We prompted the LLM to give a short answer (**a**) to the given question (**q**) as an expert in the field: `"You are a medical expert. Answer the following question in a short sentence"`. The queries were divided by question type, as shown in Table 2.
- We evaluated the quality of the generated answers for each type of query by computing the BERTscore between the generated answer and the gold standard, the human annotated paraphrases and definitions from RefoMed-EN.
- We calculated the Pearson correlation between the BERTscore and the frequency of the medical concept in the pretraining corpora.

4.2 CLS Embbedings and Co-occurrence Probability

We investigated the link between words embeddings and frequency in the pre-training corpus by conducting the following experiments:

- Firstly, we computed the **cosine similarity score** between the embeddings of the simple or multi-word medical term to be explained or defined, and the embeddings of the answer generated by the language models. We used the sentence-transformer model `paraphrase-MiniLM-L6-v2`. In order to do an exhaustive comparison, we dissolve the reference answer (a) and generated answer ($\hat{a}$) into all possible n-grams. Then, we compute the pairwise cosine similarity between 2, 3, 4 and 5 possible n-gram pairs.
- Secondly, we calculated the **probability score** between two document-frequency counts: 1) the frequency of the query term (q) in the corpora (C), and 2) the frequency of the term together with the generated answer of the language model in the corpus. We used the `WIMBD` tool to found the frequencies in terms of number of documents from DOLMA, and `infini-gram` library for The Pile.

$$P_{cooccurence} = \frac{\mathbf{C}_{df}(q|\hat{a}, C)}{\mathbf{C}_{df}(q|C)} \quad (3)$$

We calculated different n-gram combinations, of 2 to 5 n-gram length. We kept the top-3 semantically meaningful values. We show the statistical correlation between these variables in the Results section.

4.3 Tracing Back Head and Tail Knowledge

We listed the most popular medical concepts (`head`) as those appearing most frequently in the parametric memory, and the least popular medical concepts (`tail`) as having the smallest number of corresponding documents in the parametric memory.

Term	Docfreq	Examples
disease	75724477	A disease is a medical condition that affects the body's structure or function
cancer	50918098	A disease that affects the cells
anxiety	35107524	Anxiety is a feeling of fear and uneasiness
testosterone-inhibiting	115	The answer is testosterone-inhibiting
biological tissue damage	503	Give examples of biological tissue damage
lifelong neurological consequences	29	Explain the neurological consequences of the disease

Table 3: Examples of LLM paraphrasing. Frequent terms are paraphrased, while rare terms yield outputs similar to or derived from pre-training data.

We traced back 100 concepts (1.62%), where 50 were head concepts (0.8%) and 50 percent tail concepts, to the pretraining corpus. We discarded the very long tail concepts from RefoMed-EN that had zero frequency. However, it is important to note that WIMBD's search is exact match based (while ignoring special characters and punctuation), and that some zero frequency terms are very long and technical, such as "disorder of bronchial ventilation" or "SMN1 gene-related proximal spinal muscular atrophy". To compare expert (EX) to general public (GP) datasets, we split this number evenly between RefoMed-EN-EX and RefoMed-EN-GP. We analyzed word level frequencies in 100 documents for each term downloaded with WIMBD.

As the linguistic diversity and the quality of the pretraining data is extremely important in the task performance evaluation, we conducted a close up analysis of corresponding documents in the DOLMA corpus. We analyzed the texts for a selected number of `head` and `tail` concepts.

	DEF	EX	DEN	PARA	EXP
O1b	0.27	0.05	0.15	0.11	0.12
O7b	0.23	0.16	0.15	0.14	0.14

Table 4: Pearson Correlation coefficient (ρ<0.05) between BERTscore(a, $\hat{a}$) and $\texttt{log}(\mathbf{C}_{df}(q))$ for OLMo-1b (O1b) and OLMo-7b (O7b).

5 Results and Analysis

Statistical correlations between medical concepts frequency in the parametric memory. To answer **(RQ1)**, we computed the correlation coefficient between BERTscore and query term frequency C_{df}. We show the Pearson correlations scores coefficient between the BERTscore (a, $\hat{a}$) and log(C_{df} (q)) on the full dataset of 6,170 terms and gold paraphrases, in Table 4. The best scores were obtained with OLMo-1b on the DEF query type (0.27). While the value indicates a weak correlation, it is the highest among all query types, showing some link between the semantic similarity of the generated answer compared to the gold standard answer, and the frequency of the term in DOLMA. Second best scores are on DEN (0.15), while PARA and EXP have lower values (0.11 and 0.12). The lowest correlations score were obtained on the EX query type (0.05). We obtained similar results for the bigger model, OLMo-7b, for DEF (0.23), DEN (0.15), and EXP (0.14). OLMo-7b has better scores for EX (0.16) and PARA (0.14), which shows that the model handles the complex linguistic task better. We conducted a manual analysis of the quality of the generated answers for each query type, described in the next section that tackles **(RQ2)**.

When computed separately by type of corpus, expert medical texts (RefoMed-EN-EX) and general public medical texts (RefoMed-EN-GP), the results are consistent: the scores are the highest for DEF and DEN. However, the scores on RefoMed-EN-EX are better than those on RefoMed-EN-GP for DEF, and even higher than the values on the full dataset (0.31 compared to 0.24 and 0.27). This result might indicate that in the expert texts, the task performance of the model on DEF is slightly more correlated to the term frequency than in the general public (as RefoMed-EN-EX has a higher number of technical terms). On the contrary, for DEN, scores are higher on RefoMed-EN-GP (0.19) than RefoMed-EN-EX (0.11), also surpassing the full dataset score (0.15). We analyzed these results manually in the next section.

QA task performance evaluation on linguistically diverse query types. To investigate **(RQ2)**, we conducted a qualitative analysis of the LLM's linguistic *understanding*. Our hypothesis was that language models show very high performance for the `definition` type query, as this type of query

is very frequent in knowledge intensive QA benchmarks (Rebboud et al., 2024; Zhao et al., 2024b; Fei et al., 2023). The results are consistent with our hypothesis (Table 4). On the flip side, we observed that the LLM does not completely *understand* the concept of `paraphrase` or `explanation`, as it generated definitions as answers instead.

We noticed that OLMo-1b obtained lower task performance with the *denomination* and *exemplification* types of queries, as it tends to either repeat the term (for denomination in particular) or give a definition style answer for both. The difficulties for these two queries come from linguistic and domain specific linguistic traits, such as:

- `Denomination query` - the LLM repeats the term instead of giving another synonym for the medical concept. Furthermore, the presence of highly technical long-tail terms (such as *lymphocytic bacterial meningitis*; freq=2) or opaque abbreviations (e.g. *FixM/F*; freq=11) renders the task even more difficult for the language model.

- `Exemplification query` - the LLM does not output examples (i.e instances, types, subtypes) when the query term is very technical, thus long-tail knowledge: (*Give examples of severe motor disorders* → *The answer is severe motor disorders*). However, the model task performance increases for head knowledge (*1. Give examples of cardiovascular disease* → *a heart attack*; *2. Give examples of psychological factors* → *The answer is: Psychological factors include: fear, anxiety, and depression*).

Tracing back head and tail knowledge in the pretraining corpus. We show the 100 analyzed terms and their document and term frequencies in Table 6 and Table 7 in Appendix B. We looked into a total of 5,074 documents from the DOLMA corpus, where 5,000 contained the `head` words and 74 the `tail` for the tail concepts. For each `head` term out of the 50, we downloaded 100 documents where this term appeared at least once. For `tail` terms, we looked for concepts that appeared in 1 to 3 documents in the DOLMA pretraining corpus.

The `long tail` terms with very low frequency (>4) are very technical long multi-word terms, such as "pseudo-rhizomelic polyarthritis", "lymphocytic bacterial meningitis", "CoA HMG reductase inhibitors". However, there are also very long multi-word terms in RefoMed-EN-GP, like "work-related musculoskeletal disorders of the upper limbs and neck" or "shiftworker sleep disorder" with a term frequency of 2 and 1 in the whole DOLMA corpus.

Criteria	Model	RefM-EN-EX	RefM-EN-GP	RefM-EN
Diversity	OLMo-1b	-0.3904	-0.5710	-0.4403
	Pythia-1b	-0.0206	-0.1202	-0.0317
Scalability	OLMo-7b	-0.3656	-0.3308	-0.3501

Table 5: Pearson correlation between CLS cosine similarity (query term, response-top3-ngrams) and the cooccurence probability of frequency of the term together with its top3 n-grams. (Red color denotes ρ>0.05)

Regarding the `head` terms, the most frequent term from RefoMed-EN is "life", which appears in 497M documents in the pretraining corpus, while the second most frequent is "control" with 230M documents. Other head terms are "conditions" (130M), "function" (114M), "skills" (126M) and typical medical head concepts such as "treatment" (113M), "pain" (80M), "disease" (75M) and "cancer" (50M). While analyzing the term frequency, we noticed that terms such as "control", "function" and "client" appear in code texts, thus irrelevant for your medical knowledge QA task.

CLS n-gram embeddings on head and tail terms. We explored **(RQ3)** in line with Wang et al.'s (2024) hypothesis: LLMs generalize more in reasoning-intensive tasks, and they memorize more in simpler knowledge-intensive task (like in the case of tail knowledge). We conducted the CLS n-gram analysis on the list of 100 head and tail concepts to verify this hypothesis in the medical field. We calculated the CLS embedding BERTscore between the medical term and n-grams of the generated answer of different sizes (2, 3, 4 and 5 n-grams). We kept the top 3 best BERTscore for each term and we compared them with the frequencies of the term together with theses n-grams in the pretraining corpus. We show the distribution of the results in Table 5.

The negative distribution scores (-0.44 on RefoMed-EN for OLMo-1b) indicate that the language model tends to create its own sentences and does not take full information package directly from DOLMA. In terms of scalability, OLMo-7b is inline with OLMo-1b (-0.35 on RefoMed-EN), while another family of models, Pythia-1b, show very low distribution scores. This suggests that the LLM prioritizes semantic paraphrasing (backed by the semantic evidence in relation to the CLS similarity) (see examples in Table 3), and it is less likely to reproduce the same content from the pretraining corpus, as proven by the syntactic evidence related to the n-gram occurrence. OLMo-1b's scores are

better for head terms and for the medical texts destined to general public texts (-0.71). This finding demonstrated that the model used more diverse vocabulary for popular medical knowledge, and less technical medical concepts, particularly in the expert texts (+0.07). For the bigger model, OLMo-7b, we obtained similar values for both types of texts, popular and scientific. This observation supports previous studies on general language (Wang et al., 2024), indicating that the LLM memorizes more on knowledge intensive questions in the medical domain, than for popular medical terms.

6 Discussion

Scaling up small language models does not necessarily improve performance. We replicated our experiments on a bigger model, OLMo-7b. We notice a similar trend for definition-type questions: both OLMo-1b and OLMo-7b exhibit positive Pearson correlation coefficient (p) between BERTscore and query term frequency. Interestingly, we noticed that OLMo-7b shows a weaker trend (p=0.23) compared to OLMo-1b model (p=0.27), this implies that with increase in the size of language models the ability to deal with tail knowledge does not necessarily increase. Regarding overall comparison of CLS n-grams scores on head and tails (see Table 5), we further notice a similar trend between OLMo-1b and OLMo-7b models. However, these findings might change when testing even bigger models.

Human evaluation shows hallucinations are not scale related. We conducted a manual analysis of 400 answers given by the four language models for the 100 head and tail terms dataset (Figure 3). We evaluated task performance and diversity of models by looking directly into the data for hallucinations. Our analysis showed that Pythia-1b is more prone to generating hallucinatory texts (+22%) compared to OLMo-1b, as it was previously shown (Groeneveld et al., 2024). As for bigger models, OLMo-7b is hallucinating more (+19%) than its Instruct version. The best model remains OLMo-1b (33% hallucinated answers), followed by OLMo-7b-Instruct (39%). Pythia-1b demonstrates an opposite trend for definition-type questions (DEF) as compared to OLMo-1b, with a comparatively low Pearson correlation coefficient of -0.103. However, OLMo-7b shows a similar positive trend for paraphrase-type questions (PARA) with a coefficient of 0.14.

We want to stress that results might differ with bigger models. Nevertheless, the main goal of our study was to analyze smaller sized LLMs to support reproducibility and frugality for responsible use of generative AI tools.

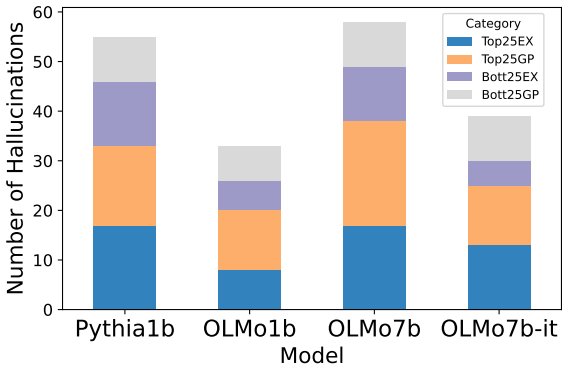

Figure 3: Manual analysis of hallucinations on 100 head and tail terms. Best model is OLMo-1b, with the lowest rate of hallucinations (33%).

More linguistic diversity metrics are needed. BERTscore has limitations in capturing the semantic aspects of query types. For example, for denomination-types queries where the LLM only repeated the medical term in the query, the BERTscore will be very high, but not relevant. In a concurrent study, Guo et al. (2024) compared lexical, syntactic and semantic distribution of LLMs texts to human gold answers. Lee and Lee (2023) proposed a pipeline to identify 220 popular handcrafted linguistic features. However, these features are not thoroughly tested on different writing tasks, and more research needs to be done on their utility for diverse query types.

7 Conclusion

Our study introduced TrackList, a pipeline to analyze, trace back and evaluate a language model's answer to diverse linguistic queries. We showed that frequency of terms in the pretraining corpus impacts performance: LLMs tend to give inaccurate answers for head terms and more accurate answers for tail terms in the medical domain. Furthermore, the LLMs studied were prone to paraphrase more on popular medical terms, and less on very scientific terms. Our linguistic analysis demonstrated that language models tend to give definition-type answers to different queries, even when asked to give examples or paraphrases. This demonstrates that Small Language Models have a limited linguistic *understanding* (or representation) of the best fitted answer for diverse question types.

Our study focused on small size language models to investigate the inner working of LLMs without domain knowledge finetuning, DPO finetuning or RAG systems. Our purpose was to analyze how the pretraining corpus and the word frequency (head and tail) impacts the accuracy of the small LLMs answer to different linguistic types of questions. This research design was motivated by the fact that humans use language models as plug-and-play tools, without any finetuning methods. Future research using these methods might improve vanilla LLM's performance for the QA task. Moreover, LLM-as-a-Judge evaluation could be explored and compared with existing metrics.

Ethical Considerations

The dataset used for this experiments has open license (CC BY-NC 4.0) and can be used for research by the NLP community. The RefoMed-EN dataset contains no personal data or patient data.

Limitations

This work was conducted only on English using only fully open language models, OLMo (Groeneveld et al., 2024) and Pythia (Biderman et al., 2023). Due to computational power limitations, we conducted our analysis on Small Language Models (1b and 7b). Future studies could include more recent models, like OLMo2 (OLMo et al., 2024) and OLMo3 (Olmo et al., 2025), as well as concurrent data tracing tools such as OLMoTrace (Liu et al., 2025). Other open source LLMs give access to their pretraining data, like BLOOM (Workshop et al., 2022) available, but the big size of the dataset makes it difficult to explore.

Moreover, we are aware of this limitation for your semantic similarity evaluation, and we further investigate medical metrics such as MEDCON (Yim et al., 2023), or fact checking metrics and tools like FACTSCORE (Min et al., 2023), FIRE (Xie et al., 2024), or LOKI (Li et al., 2025).

References

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379.

Shayan Ali Akbar, Md Mosharaf Hossain, Tess Wood, Si-Chi Chin, Erica Salinas, Victor Alvarez, and Erwin Cornejo. 2024. Hallumeasure: Fine-grained hallucination measurement using chain-of-thought reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15020–15037.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20:1–23.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Ioana Buhnila. 2023. *Une méthode automatique de construction de corpus de reformulation*. Ph.D. thesis, Université de Strasbourg.

Ioana Buhnila, Aman Sinha, and Matthieu Constant. 2024. Retrieve, generate, evaluate: A case study for medical paraphrases generation with small language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 189–203.

Xavier Daull, Patrice Bellot, Emmanuel Bruno, Vincent Martin, and Elisabeth Murisasco. 2023. Complex qa and language models hybrid architectures, survey. *arXiv preprint arXiv:2302.09051*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, and 1 others. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, and 1 others. 2023. What's in my big data? *arXiv preprint arXiv:2310.20707*.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Songyang Zhang, Kai Chen, Zongwen Shen, and Jidong Ge. 2023. Lawbench: Benchmarking legal knowledge of large language models. *arXiv preprint arXiv:2309.16289*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, and 1 others. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Natalia Grabar and Rémi Cardon. 2018. Clear-simple corpus for medical french. In *ATA*.

Natalia Grabar and Iris Eshkol. 2016. Why do we reformulate? automatic prediction of pragmatic functions. In *HrTAL 2016*.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, and 1 others. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2024. Benchmarking linguistic diversity of large language models. *arXiv preprint arXiv:2412.10271*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2024. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Sanna Iivanainen, Jarkko Lagus, Henri Viertolahti, Lauri Sippola, and Jussi Koivunen. 2024. Investigating large language model (llm) performance using in-context learning (icl) for interpretation of esmo and nccn guidelines for lung cancer.

Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. 2024. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *arXiv preprint arXiv:2401.15269*.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Bruce W Lee and Jason Lee. 2023. Lftk: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Dongyang Li, Junbing Yan, Taolin Zhang, Chengyu Wang, Xiaofeng He, Longtao Huang, Jun Huang, and 1 others. 2024a. On the role of long-tail knowledge in retrieval augmented large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–126.

Haonan Li, Xudong Han, Hao Wang, Yuxia Wang, Minghan Wang, Rui Xing, Yilin Geng, Zenan Zhai, Preslav Nakov, and Timothy Baldwin. 2025. Loki: An open-source tool for fact verification. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 28–36.

Huihan Li, Yuting Ning, Zeyi Liao, Siyuan Wang, Xiang Li, Ximing Lu, Wenting Zhao, Faeze Brahman, Yejin Choi, and Xiang Ren. 2024b. In search of the long-tail: Systematic generation of long-tail inferential knowledge via logical rule guided search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2348–2370.

Yucheng Li, Yunhao Guo, Frank Guerin, and Chenghua Lin. 2024c. An open-source data contamination report for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 528–541.

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. 2023. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th international conference on conversational user interfaces*, pages 1–6.

Jiacheng Liu, Taylor Blanton, Yanai Elazar, Sewon Min, YenSung Chen, Arnavi Chheda-Kothary, Huy Tran, Byron Bischoff, Eric Marsh, Michael Schmitz, and 1 others. 2025. Olmotrace: Tracing language model outputs back to trillions of training tokens. *arXiv preprint arXiv:2504.07096*.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

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.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, and 1 others. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Ajay Madhavan Ravichandran, Julianna Grune, Nils Feldhus, Aljoscha Burchardt, Roland Roller, and Sebastian Möller. 2024. Xai for better exploitation of text in medical decision support. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 506–513.

Youssra Rebboud, Pasquale Lisena, Lionel Tailhardat, and Raphael Troncy. 2024. Benchmarking llm-based ontology conceptualization: A proposal. In *ISWC 2024, 23rd International Semantic Web Conference*.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787.

Pasi Shailendra, Rudra Chandra Ghosh, Rajdeep Kumar, and Nitin Sharma. 2024. Survey of large language models for answering questions across various fields. In *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 520–527. IEEE.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, and 1 others. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.

Amalia Todirascu, Sebastian Padó, Jennifer Krisch, Max Kisselew, and Ulrich Heid. 2012. French and german corpora for audience-based text type classification. In *LREC*, volume 2012, pages 1591–1597.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2024. Generalization vs memorization: Tracing language models' capabilities back to pretraining data. *arXiv preprint arXiv:2407.14985*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Skyler Wu, Eric Meng Shen, Charumathi Badrinath, Jiaqi Ma, and Himabindu Lakkaraju. 2023. Analyzing chain-of-thought prompting in large language models via gradient-based feature attributions. *arXiv preprint arXiv:2307.13339*.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2024. Fire: Fact-checking with iterative retrieval and verification. *arXiv preprint arXiv:2411.00784*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10(1):586.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024a. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024b. Knowledgefmath: Knowledge-intensive math reasoning in finance domains. In *62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 12841–12858. Association for Computational Linguistics (ACL).

A RefoMed Linguistic Annotation

The RefoMed (Buhnila et al., 2024) has extensive human annotation validated by human inter-annotator agreement. The annotators were French bilingual speakers of different levels of expertise in linguistics and NLP: senior researchers, PhD and master students specialized. The annotation was conducted on the original version in French and then automatically translated in English with DeepL API. The translation was cleaned and verified, however, some translation mistakes might be scattered in the 6,170 lines dataset.

The annotation comprises a linguistic marking of the lexical relations that exists between the reformulation and the medical term. We show a full example of annotation in Table 9. Different relations were tagged, such as *hyperonymy, hyponymy, meronymy and synonymy*. In the case of pragmatic functions that we explored in this study, the annotators marked the reason to the use of the reformulation in the text: to give a *definition, examples, explanations, denominations* (other name for the same concept), and *paraphrases* of the medical term. RefoMed has annotations of discourse markers that help automatically identify reformulations (all pragmatic types presented above) inside the same sentence, such as "is a", "meaning", "also called", "refers to" (Buhnila, 2023).

B Head and Tail Query Terms from RefoMed-EN

We extract the 50 query terms from head and tail of the Dolma corpus. See Table 6 for 25 terms from expert and general public domain and for each term, the table also includes the number of documents (C_{df}) present in the Dolma corpus. Similarly, we extract 50 terms from tail distribution of Dolma corpus (Table 7 and Table 8).

RefoMed-EN-EX			RefoMed-EN-GP		
Term	DocFreq	TermOcc	Term	DocFreq	TermOcc
life	497066067	13012	professional	161724849	6894
control	230950452	12218	version	159636360	15503
intervention	23396121	5022	surgery	37481694	8558
conditions	136013609	4499	conditions	136013609	4499
function	114612170	14331	skills	126936290	8171
Treatment	113612643	6963	treatment	113612643	6863
Pain	80378923	24840	tool	92230908	13708
patients	78742642	5985	condition	88989985	8446
disease	75724477	9078	scale	88228556	12443
'client'	65025830	9365	long term	83709724	3505
behavior	64089027	6194	Pain	80378923	24840
tests	57368874	21505	disease	75724477	9078
Cancer	50918098	9858	foot	67911272	8771
drugs	44267374	11980	volume	67588213	5281
Cells	43057710	14193	exercise	64859465	7467
lifestyle	42983527	4540	Iron	35692516	11691
MS	38326252	17530	brain	60468547	7769
Surgery	37481694	8558	tests	57368874	21505
Anxiety	35107524	7560	drug	57092723	25334
diseases	34543764	3966	cancer	50918098	9858
IM	33368643	29766	technique	48805868	4938
protein	32887735	17620	procedure	47808948	21328
depression	30980124	9459	symptoms	44391219	5593
infection	27670132	6022	drugs	44267374	11980
facilitate	27181783	2858	MS	38326252	17530

Table 6: Head-50 query terms from RefoMed-EN from the Dolma corpus. We show 25 terms from RefoMed-EN-EX and 25 terms from RefoMed-EN-GP.

C Model Size and Budget

Experiments were done on a P100 NVIDIA GPU, for an individual runtime of $\leq$ 15 hours.

RefoMed-EN-EX		
Term	**DocFreq**	**TermOcc**
Horton's disease and Takayasu's arteritis	1	2
vesico-sphincter dyssynergia (VSD)	1	1
Motor Neurone Disease (MMN)	1	2
strain CF7968	1	1
region of around 100bp	1	1
inappropriate secretion of HAD	1	1
surgical haemorrhagic shock	1	1
psoriatic onychosis	1	3
extracellular proliferating bacteria	1	1
medical problems contributing to insomnia	1	1
pain resulting from a medical cause	1	1
Early administration of an amino acid solution	1	1
recent global weakening	2	2
Crohn's disease and hemorrhagic rectocolitis	2	2
proteins EWI-2 and CD81	2	2
Decompression sickness (DD)	2	6
low risk of bias threshold	2	2
Steinert's myopathy	3	6
cognitive disorders in Parkinson's disease without dementia	2	4
lymphocytic bacterial meningitis	2	3
forms of neurolupus	2	2
family of antithrombotics	2	2
dopamine stock solutions	2	2
CoA HMG reductase inhibitors	2	2
Jet lag and shift work syndromes	2	4

Table 7: Tail-25 query terms from RefoMed-EN-EX based on the Dolma corpus.

RefoMed-EN-GP		
Term	**DocFreq**	**TermOcc**
treatment of ovarian cancer by surgery	1	1
Pseudo-rhizomelic polyarthritis	1	1
Sydenhan's chorea	1	2
Uterine (uterine/endometrial) cancer	1	10
Work-related musculoskeletal disorders of the upper limbs and neck	1	2
reformulation of a food product	1	1
bronchiolitis obliterans syndrome and lymphocytic bronchitis	2	1
ImmunoGuard(R)	2	2
bronchiolitis obliterans syndrome and lymphocytic bronchitis	1	2
prophylaxis with immunoglobulin g	1	1
notion of therapeutic education	1	4
disabled child's education allowance	2	2
Blood Form Count	2	2
rare bullous skin diseases	1	1
psychiatric disorders and social anxieties	1	1
pain resulting from a medical cause	1	1
gushing liquid diarrhea	1	1
disorders of immediate memory	1	1
shiftworker sleep disorder	1	1
disorder of ocular refraction	1	2
Eulenburg's paramyotonia congenita	2	2
notion of therapeutic education	2	4
facio-truncular obesity	2	2
STBBI, STD	2	11
American mucocutaneous and cutaneous leishmaniasis	2	4

Table 8: Tail-25 query terms from RefoMed-EN-GP based on the Dolma corpus.

Context (C)	Myasthenia is an autoimmune disease triggered by autoantibodies that most often target the nicotinic acetylcholine receptor.
Term (T)	Myasthenia
Marker (M)	is a
Indicator (I)	disease
Reformulation (R)	an autoimmune disease triggered by autoantibodies that most often target the nicotinic acetylcholine receptor.
Lexical relation (LR)	hyperonymy
Pragmatic function (PF)	definition
C	Diabetic kidney disease (DKD), also known as diabetic nephropathy (DN) [...]
T	Diabetic kidney disease (DKD)
M	also known as
I	disease
R	diabetic nephropathy (DN)
LR	synonymy
PF	denomination
C	Antibiotics (chloramphenicol, tetracycline and doxycycline) are used in the treatment of this disease.
T	Antibiotics
M	()
I	disease
R	chloramphenicol, tetracycline and doxycycline
LR	hyponymy
PF	exemplification
C	We planned to include all RCTs and quasi-RCTs (in which assignment was not random, but was supposed to be unbiased, e.g., alternate assignment) involving treatment for paraneoplastic neuropathies.
T	quasi-RCTs
M	()
I	for example
R	in which assignment was not random, but was supposed to be unbiased
LR	meronymy
PF	explanation
C	We accepted as comparison arms a placebo control (sham intervention) or without injection, or another active treatment [...]
T	placebo control
M	()
I	intervention
R	sham intervention
LR	synonymy
PF	paraphrase

Table 9: Examples of the RefoMed fine-grained linguistic annotation for each lexical relation and pragmatic function used to determine the question types.