Towards Automated Evaluation of Knowledge Encoded in Large Language Models

Bruno Carlos Luís Ferreira, Catarina Silva, Hugo Gonçalo Oliveira
University of Coimbra, DEI, CISUC/LASI
Coimbra, Portugal
{brunof,catarina,hroliv}@dei.uc.pt

Abstract

Large Language Models (LLMs) have a significant user base and are gaining increasing interest and impact across various domains. Given their expanding influence, it is crucial to implement appropriate guardrails or controls to ensure ethical and responsible use. In this paper, we propose to automate the evaluation of the knowledge stored in LLMs. This is achieved by generating datasets tailored for this specific purpose, in any selected domain. Our approach consists of four major steps: (i) extraction of relevant entities; (ii) gathering of domain properties; (iii) dataset generation; and (iv) model evaluation. In order to materialize this vision, tools and resources were experimented for entity linking, knowledge acquisition, classification and prompt generation, yielding valuable insights and lessons. The generation of datasets for domain specific model evaluation has successfully proved that the approach can be a future tool for evaluating and moving LLMs "black-boxes" to human-interpretable knowledge bases.

Keywords: Natural Language Processing, Large Language Models, knowledge Base, Explainable Artificial Intelligence

1. Introduction

Nowadays, even those with minimal computer proficiency and a basic understanding of current technologies are likely aware and taking advantage of Large Language Models (LLMs). On the one hand, there are many upsides to these technologies, such as, efficiency, automation, and versatility (Strasser, 2023). On the other hand, there are many sectors of society that have reported downsides to their usage, including education (students, professors, researchers), companies (administrative work), and others (Fecher et al., 2023). Understanding the fact that humans will not go backwards, we need to address the current and future problems of such technologies.

Due to the rapid advancements and widespread acceptance of LLMs, numerous drawbacks of these technologies emerged. The well-known examples of some shortcomings are: factual errors (Wang et al., 2024), hallucinations (Ye et al., 2023), inconsistency (Elazar et al., 2021), and not being human-interpretable, i.e., "black-boxes" (Sun et al., 2022). These issues do not align with the principles of Responsible Artificial Intelligence. Also, LLMs are trained on large quantities of data that is not always easy to track, represented through opaque methods and not directly accessible. Therefore, we may add that, to some extent, LLMs do not adhere to the Findable, Accessible, Interoperable, Reusable (FAIR) principles (Wilkinson et al., 2016). Nevertheless, researchers are working to understand how to adapt FAIR guiding principles to FAIR AI models (Ravi et al., 2022).

Our objective is to contribute to more transparent and human-interpretable LLMs. Towards that vision, we propose an approach for automating the evaluation of knowledge in LLMs, across diverse domains. This process is key in our world because, as mentioned earlier, we are witnessing the widespread application of LLMs across various software, professions, and as auxiliary aid in a broad range of tasks.

The main contributions of this work are summarized as follows:

- The proposal of an end-to-end solution for automating domain-specific generation of evaluation datasets, applicable to any LLM and domain;
- An instantiation of the proposed approach with its application to two critical domains, finance and medicine.
- The evaluation of a broad range of masked language models in the previous domains, where we confirm the feasibility of the proposed approach and reveal limitations of such models when it comes to zero-shot domain knowledge acquisition.

In the remainder of this paper, we present the starting points and inspirations of our work, by describing the related work in Section 2. We proceed by detailing our general approach in Section 3. In

DLnLD: Deep Learning and Linked Data @LREC-COLING-2024, pages 76–85
21 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

Section 4, we instantiate the approach for two domains: financial and medical. Obtained results are further analysed, as well as challenges and limitations of the implementation. We conclude the paper in Section 6, with important takeaways and plans for future work.

2. Related Work

Since their early instantiations, researchers have explored Transformer Language Modes (LMs) as sources of knowledge (Petroni et al., 2019) and assessed them in tasks like relation completion, in a broad range of domains. Such an evaluation is typically supported by specifically-tailored datasets, such as LAnguage Model Analysis (LAMA), created semi-automatically from knowledge sources like Wikidata (Vrandečić and Krötzsch, 2014) and ConceptNet (Speer et al., 2017).

In the following years, various contributions adopted a similar approach (Bouraoui et al., 2020; Mickus et al., 2023; Gromann et al., 2024), i.e., probe LLMs and evaluate them on a set of knowledge sources comprised of a set of facts. Besides LAMA, other datasets were used. For instance, despite originally created for assessing static word embeddings in analogy solving, BATS (Gladkova et al., 2016) has also been used for evaluating Transformer LMs. BATS, created for English but translated to other languages (Mickus et al., 2023; Gromann et al., 2024), covers four groups of relations: inflexion morphology, derivational morphology, lexicographic semantics, encyclopedic semantics.

The goal of relation completion is to, given a subject and a predicate (relation), obtain suitable values for the object. It may resort to prompting a Transformer LMs, including BERT-based masked language models (Petroni et al., 2019; Mickus et al., 2023; Gromann et al., 2024), where the object is masked; or generative based models, including GPT-3 (Gonçalo Oliveira and Rodrigues, 2023) or BLOOM (Gromann et al., 2024), where the object is generated.

Knowledge acquisition from LMs has also raised interest from the Semantic Web community, which is confirmed by the challenge on Knowledge Base Construction from Pre-trained Language Models (LM-KBC) (Singhania et al., 2022; Kalo et al., 2023). Evaluation was based on a datasets comprising diverse world-knowledge relations (e.g., BandHasMember, FootballerPlaysPosition, PersonCauseOfDeath), each including a set of subjects and a list of ground-truth objects per subject-relation-pair.

Despite the existence of datasets like those used in the previous works, essential for evaluating LLMs, they are inherently limited. Some were created manually (Singhania et al., 2022), while others, despite involving some automatic procedure (Petroni et al., 2019), required specific planning (e.g., in the selection of relations and definition of inclusion criteria). They are created with a specific goal and, once created, remain static.

The automation of data collection from specific domains and the dynamic generation of datasets represents a promising avenue. Therefore, we propose a methodology for the automation of the creation of datasets for multiple domains that can be used for evaluating LLMs.

3. Proposed Approach

We see knowledge acquisition from LLM as a way towards more transparent models. While it is impossible to represent everything in a model as a single Knowledge Graph (KG), in theory, a smaller KG can be extracted on specific domains. The models will perform differently for different domains.

So, our vision involves the possibility of evaluating any model in any domain of interest. This requires specific methods for turning user-provided seeds (e.g., domain data) into relevant domain knowledge (e.g., entities, relations), and for assessing to what extent such knowledge can be obtained from the model.

We propose an approach for this vision, depicted in Figure 1. It is based on the automatic generation of datasets given a collection of textual documents on the target domain. Datasets will contain knowledge on the target domain, guided by the input collection (seeds), but effectively extracted from a human-created KG. Briefly, the proposed approach encompasses four main steps:

1. Extraction of relevant entities from domain data;
2. Gathering of domain-related entity properties from a human-created knowledge graph;
3. Generation of an evaluation dataset on domain knowledge;
4. Automation of LLMs evaluation.

More formally, from a set of entities E extracted from the input collection, a subset of domain-relevant $E' \subset E$ is gathered. For each entity $e \in E'$, a set of domain properties P is then obtained from a knowledge graph G. With entities e and their respective properties $p \in P$, a dataset of triples $t(e, p, o)$ is finally generated.

3.1. Extract Relevant Entities

The set of relevant entities E is first extracted from the collection of textual documents. We rely on en-

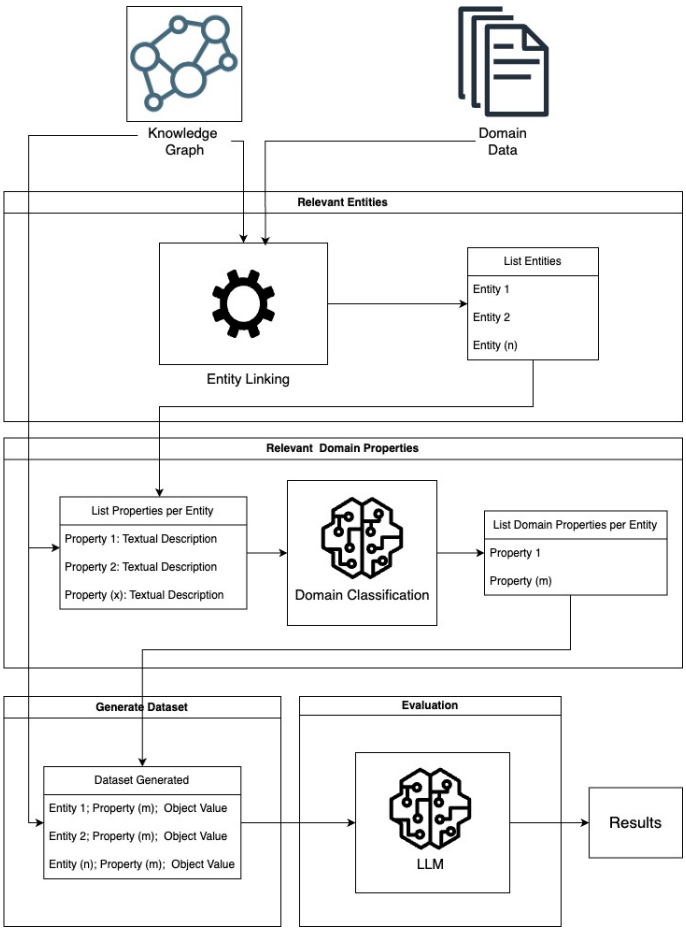

Figure 1: Approach for automating the evaluation of LLMs in given domains.

tity linking because we want to get ground knowledge on these entities, e.g., from a given KG.

Moreover, not all extracted entities will be relevant for the domain. So, having in mind the goal of assessing knowledge on the target domain, these should not be considered. Here, we assume that the most frequent entities in the input collection will also be the most relevant for the domain, and focus on these, i.e., extracted entities are ranked by their frequency, and only the top-$\mathtt{n}$ are used. These will constitute the set E'.

3.2. Gather Domain Properties

With each entity linked to a KG, many properties can be obtained. However, not all of them will be specific of the target domain. To consider only domain-related properties $\mathtt{p} \in \mathtt{P}$, we can use a text classifier trained for labelling the domain of given text. The input can be the name of the property but, if available, a longer description of the property can be used. If such classifier is not available and not enough annotated data is available for training, one can always opt for zero-shot text classification.

3.3. Dataset Generation

The last step is the construction of the dataset to be used for evaluation. From the each $\mathtt{e} \rightarrow \mathtt{p}$ pair, objects $\mathtt{o}$ are obtained from the KG, resulting in triples `(e, p, o)`. A possible triple is `(Portugal, hasCapital, Lisbon)`. The resulting dataset is a collection of such triples.

3.4. Automation of the Evaluation

Having a domain-oriented dataset of triples is an enabler of many evaluation possibilities where different approaches can be taken. One is follow the examples of probing, where we produce a sentence (prompt), provide it as input for a LLM, and evaluate the output of the model.

Depending on the target LLM, the creation of the prompt should be different, i.e., prompt engineering is strongly involved in this process.

4. Experimental setup

To materialise our vision, we experimented with different resources for entity linking, knowledge acquisition, classification and prompt generation,

which resulted in an initial implementation, described in this section.

The diagram in Figure 2, instantiates the one in Figure 1, in what constitutes our first implementation. For creating a dataset, we first need to acquire domain knowledge from the provided textual data, i.e., domain-relevant entities e and properties p. This is performed with the help of public sources of knowledge, namely DBpedia[1] (Lehmann et al., 2015) and Wikidata[2] (Vrandečić and Krötzsch, 2014), which can be queried from their respective SPARQL endpoints.

The production of this initial instance could help us identify challenges and problems, so we opted to streamline this implementation. For that, we opted to use existing applications and models in order to build and test our approach. A description of the implementation is detailed below.

4.1. Extract Relevant Entities

Entities were extracted from all the documents in the input collection. Since all of them should be on the target domain, we assume that the most frequently occurring entities are also the most relevant for the domain.

Entity extraction is made with the help of DBpedia Spotlight (Daiber et al., 2013), a tool for entity linking. Therefore, more than just identifying entities in text, Spotlight connects them to DBpedia resources, and thus to the Linked Open Data cloud, where knowledge about the target entities can be obtained from.

Instead of the obvious choice of using DBpedia directly, we opted for Wikidata as a Knowledge Base (KB) for being based on statements and having a community-curated ontology. This makes it, expectedly, more reliable than DBpedia, which is automatically generated from Wikipedia documents. Therefore, we must map the DBpedia entities to their Wikidata entries. For that, we relied on `owl:sameAs`, an Web Ontology Language (OWL) property that indicates that two URI references refer to the same thing in the world. Since there are `owl:sameAs` cross-links between DBpedia and Wikidata, we can use them to get the Wikidata URI corresponding to a DBPedia entity, i.e., `<Wikidata URI> owl:sameAs ?sameAsResource`.

Spotlight will extract numerous entities, but not all of them will be especially-relevant for the target domain. To get the most relevant for the domain, before mapping to Wikidata, extracted entities are ranked by frequency of occurrence in the input documents, and we focus only on the top-ranked.

[1] https://www.dbpedia.org/

[2] https://www.wikidata.org/

4.2. Gather Domain Properties

The next step is to gather domain-relevant properties involving the domain entities. While it is trivial to get from Wikidata every property involving the selected entities, i.e., `<Wikidata URI> ?property ?value`, as it happens to the input text, not all properties will be relevant for the domain. However, in this case, selecting the most frequent properties will lead to many false positives, because of generic properties held by most entities. These include generic properties connecting to the entity class (e.g., *subclass of*) or to its source (e.g., *described by source*). We thus rely on a supervised classifier for discriminating between domain-relevant properties and other.

Depending on the domain, we might need to train our own classifier or resort to zero-shot learning. Yet, for many domains, state-of-the-art text classifiers are available. An example is RoBERTa-base, fine-tuned[3] in a dataset[4] based on the Human ChatGPT Comparison Corpus (HC3) (Guo et al., 2023), which classifies text in a broad range of domains.

Since the labels of some properties can be limited, whenever possible, we classify a text resulting from concatenating the property descriptions to the name of the property, i.e., `?property: ?schema:description`. Descriptions are longer and more in line with the data used for training available text classifiers. For example, for the property *retirement age* (P3001), the following text would be classified: "*retirement age: the age at which most people normally retire from work*".

The result is a set of domain-relevant triples `t(e, p, o)`. These are used for creating the dataset, where the goal would be to, given a subject (domain-relevant entity) and a property, obtain a valid object, e.g., `(Australia, retirement age, 67)`.

4.3. Automating the Evaluation

A possible approach for assessing an LLMs with the created dataset requires the definition of prompts. Specifically, the triple should be transformed to natural language sequences where the object is missing, to be completed by the model. Evaluation will rely on the proportion of triples for which the model completion is a valid object.

There are two types of LLMs: Generative Language Models (GLMs), where the model predicts

[3] https://huggingface.co/rajendrabaskota/hc3-wiki-domain-classification-roberta

[4] https://huggingface.co/datasets/rajendrabaskota/hc3-wiki-intro-dataset

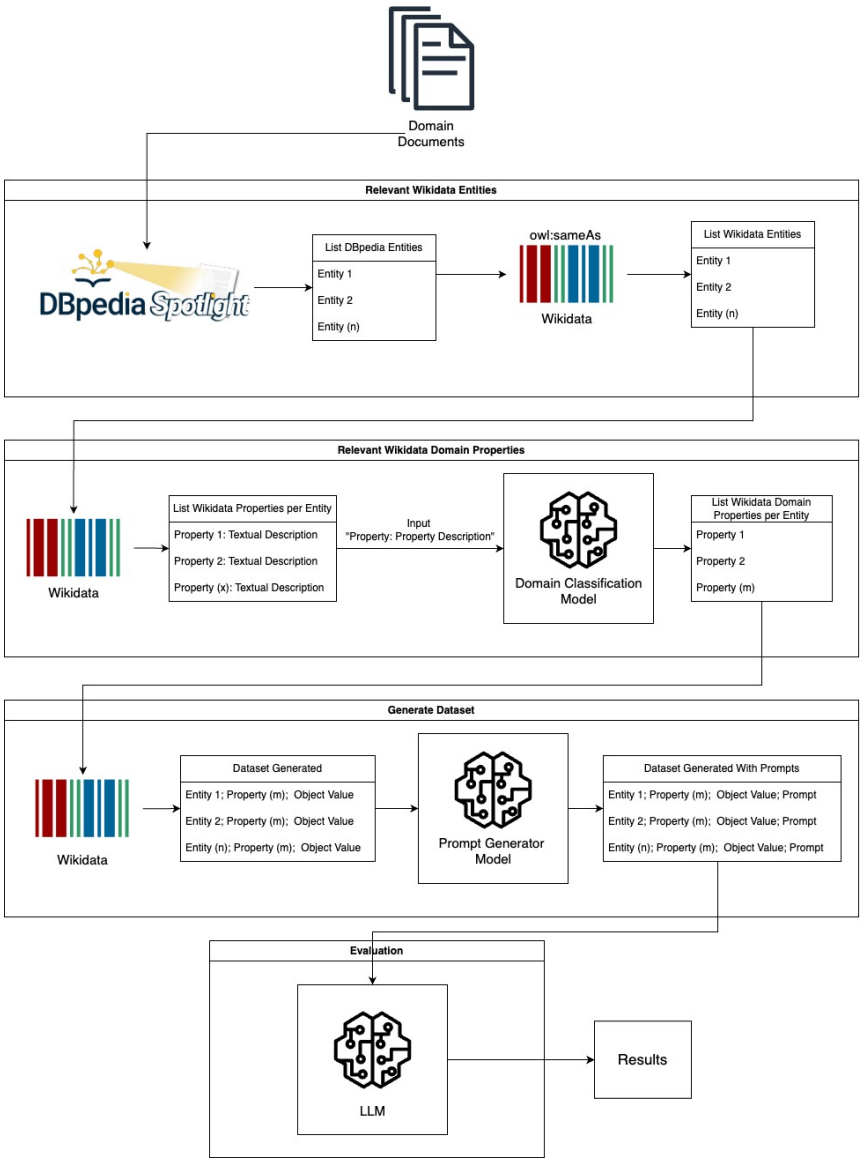

Figure 2: Representation of the instantiated implementation.

the next tokens in for a given sequence, while attending only to tokens on the left, models such as Generative Pretrained Transformer (GPT) (Radford et al., 2018) like; Masked Language Models (MLMs), where the model predicts the value of a masked token in a sequence, while attending to both the tokens in the left and right contexts, models as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) like. Depending on the type of language modelling, the prompt that interacts with the model should be different.

The experimentation reported in this work is limited to MLMs, where the boundaries of the predictions are easier to define. Having this in mind, we needed to define prompts with a masked token. For example, after replacing the object of the triple `(Australia, retirement age, 67)` by a mask, the result would be `(Australia, retirement age, [MASK])`. For simplicity, we decided to use the masked token always in place of the `object`. To provide a natural input similar to the data used for model training, i.e., natural language text, the triple is finally transformed to a prompt like "The retirement age in Australia is [MASK]".

Creating such prompts for every single property would be tedious and a limitation of the proposed approach. Therefore, prompts are generated automatically, with the help of a 7B open generative LLM, Large Language Model Meta AI (LLaMA) 2 (Touvron et al., 2023), used in its quantized version through the ollama[5] tool. LLaMA 2 was instructed to produce a sentence based on the triple provided with the following prompt: "You are

[5] https://ollama.ai/

a model that only converts triples into sentences and nothing else. You get a triple as input, for example ['Portugal', 'currency', '[MASK]'], and you need to transform the triple into a simple human-readable sentence, for example: 'The currency of Portugal is [MASK]' or '[MASK] is the currency of Portugal' or 'Portugal has [MASK] has its currency'. Choose only the best sentence possible and return it." The option for a completely automatic generation of prompts brought pros and cons, to be discussed further ahead.

5. Evaluation of MLMs

We tested our implementation using two different datasets, one focused on the financial domain and the other on the medical domain. Both domains hold significant significance in our society due to their impact. This is advantageous for the analysis of our initial implementation as it enables us to comprehend (at a high level) the general knowledge of both domains, i.e., related entities and properties.

Having that into account, we can evaluate two components that resulted from our implementation:

1. The generated dataset for each domain;

2. The output predictions of the models;

In terms of the generated datasets, we can verify 1) if the `n` top entities found and 2) if the relevant entities properties extracted are relevant for each domain.

For the evaluation of the models, we decided to follow the norm and use the mean precision at `k` (`P@k`). The value is 1 if the object is ranked among the top `k` results, and 0 otherwise. We used `k = 1`, `k = 5`, and `k = 10`.

We considered a broad range pre-trained MLMs, namely, BERT (Devlin et al., 2018)[6], Robustly Optimized BERT Approach (RoBERTa) (Liu et al., 2019)[7], A distilled version of BERT (DistilBERT) (Sanh et al., 2019)[8], Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA) (Clark et al., 2020)[9], and A Lite BERT (ALBERT) (Lan et al., 2019)[10].

[6] https://huggingface.co/google-bert/bert-base-uncased

[7] https://huggingface.co/FacebookAI/roberta-base

[8] https://huggingface.co/distilbert/distilbert-base-uncased

[9] https://huggingface.co/google/electra-base-generator

[10] https://huggingface.co/albert/albert-base-v2

5.1. Generated Datasets

In this initial implementation we aimed to use datasets as a source of "seeds", i.e., given a dataset from a specific domain it is more likely to obtain entities related to that domain.

The datasets used for both domains are publicly available in HuggingFace. For the financial domain we used a dataset[11] that contains news sentences from *Yahoo-Finance* and for the medical domain we used a dataset[12] that contains abstracts of *Pubmed* articles.

Financial dataset was relatively small, containing 25k small sentences where a total of 7567 distinct entities were found. We used the top 200 entities to build our dataset. In Table 1 are present the 10 more relevant entities present in the financial dataset, which in our opinion seems right. There are entities that are directly related to the financial domain, e.g., "Inflation", "European Central Bank", "Yen", which is logical, and in the other hand there are entities that, although not directly related to the domain, is very understandable there presence, e.g., "Reuters", "Chief Executive Officer", "Apple".

Entity	Detail	QID
Inflation	Rise in price level over time	Q35865
China (Mexico)	Municipality Location	Q942154
Reuters	International News Agency	Q130879
Board of Governors of the Federal Reserve System	Governing Body of the US Federal Reserve System	Q5440396
Chief Executive Officer	highest-ranking corporate officer	Q484876
Artificial Intelligence	field of computer science	Q11660
European Central Bank	central bank of the European Union	Q8901
Yen	official currency of Japan	Q8146
Japan	island country in East Asia	Q17
Apple	Technology Company	Q312

Table 1: Top 10 entities extracted from the finance dataset.

In terms of financial related properties of the entities, our implementation extracted some of the following properties present in Table 2). The table is a small subset of five of the 120 obtained properties classified as relevant for the financial domain by the domain classification model. Under analysis, not all the 120 properties obtained are relevant for the domain, but a considerable part is.

Our resulting dataset, containing the entities and their respective properties, is com-

[11] https://huggingface.co/datasets/ugursa/Yahoo-Finance-News-Sentences

[12] https://huggingface.co/datasets/ccdv/pubmed-summarization

Relation	PID
has subsidiary	P749
owner of	P1830
retirement age	P3001
Indeed company ID	P10285
owned by	P127

Table 2: Relevant properties gathered for the finance domain.

posed of 1115 different triplets `(entity, property, object)`. An actual example of an extracted triple: `(Microsoft, owned by, [Bill Gates, BlackRock, The Vanguard Group])`, and from that triple the generated prompt to interact with the MLMs was "The owner of Microsoft is [MASK].".

Medical dataset is considerable bigger, with 130k abstracts from *Pubmed* Papers. As the abstracts are longer than the financial news sentences, we decided not to use the entire dataset. From a total of 130k abstracts we used 9k. We obtained a total of 16791 distinct entities which indicates how much more complex is this medical dataset (compared with the financial dataset used). The following Table 3 contains the 10 more relevant entities in the 9k abstracts used.

Entity	Detail	QID
Riboflavin	Chemical	Q130365
Cancer	Disease	Q12078
Protein	Biomolecule	Q8054
Mortality Rate	Measure Deaths	Q58702
Pain	Unpleasant Feeling	Q81938
Metastasis	Spread of a Disease	Q181876
Gene	Unit of Heredity	Q7187
Obesity	Excess Body Fat	Q12174
Brain	Organ	Q1073
Insulin	Pancreas Hormone	Q50265665

Table 3: Top 10 entities extracted from the medical dataset.

In terms of relevant medical properties, the domain classification model only obtained 11 distinct properties, which is considerably less compared with the financial domain, however, all the 11 extracted are related to the medical domain.

Relation	PID
health specialty	P1995
PatientsLikeMe condition ID	P4233
PatientsLikeMe treatment ID	P4235
possible treatment	P924
symptoms and signs	P780

Table 4: Relevant medical properties gathered.

The resulting dataset, containing the entities and their respective properties, is composed of 172 different triplets. An actual example of an extracted triple: `(stroke, health specialty, [neurology, neurosurgery])`, and from that triple the generated prompt was "The health specialty of stroke is [MASK].".

5.2. Models Evaluation

For the prompts generated we ran the mentioned masked LLMs, obtaining as a result a set of outputs that we could compare with the ground truth of the dataset created. Follows the results that each model obtained in each domain, financial (Table 5) and medical (Tabel 6).

Model	Acc@1	Acc@5	Acc@10
ALBERT	0.009	0.022	0.023
BERT	0.012	0.021	0.025
DistilBERT	0.009	0.022	0.025
ELECTRA	0.008	0.016	0.019
RoBERTa	0.008	0.030	0.036

Table 5: Accuracy results in the finance domain

Model	Acc@1	Acc@5	Acc@10
ALBERT	0.038	0.077	0.108
BERT	0.045	0.089	0.108
DistilBERT	0.051	0.102	0.115
ELECTRA	0.006	0.045	0.096
RoBERTa	0.045	0.096	0.108

Table 6: Accuracy results in the medical domain

The results obtained in both domains are subpar. That occurs for a multitude of reasons, i.e., challenges and limitations that exist in our implementation.

The analysis of the outcomes, focusing on their overall precision, is impractical, as there are numerous enhancements that need to be made, mainly in two areas: the selection of domain-relevant properties, and the generation of the prompt from the extracted triples.

However, if we analyse the results by property, there are some properties that the models obtained good performance. Table 7 shows two properties from each domain that all five models got reasonable results for. There are other properties that each model performed better, but these are the two properties in the "top 10" properties of each model.

5.3. Limitations Analysis

The quality of the overall results in both datasets generated and the evaluation of the LLMs was decreased by some decisions we took to accelerate the process of implementation. Several times,

	Financial		Medical	
Model	**name** Acc@10	**retirement age** Acc@10	**symptoms and signs** Acc@10	**PatientsLikeMe condition ID** Acc@10
ALBERT	0.143	0.125	0.077	0.231
BERT	0.286	0.500	0.385	0.192
DistilBERT	0.286	0.625	0.308	0.192
ELECTRA	0.286	0.750	0.308	0.154
RoBERTa	0.286	0.875	0.231	0.269

Table 7: Two of the best best performing properties in the financial and medical domains

those decisions were to apply already existing applications and available models, which was not optimal.

To acquire domain knowledge, we relied on DBpedia Spotlight for entity linking, but used Wikidata as the KB. At this stage, we encountered some challenges that we need to overcome. The initial obstacle is that employing `owl:sameAs` is not a flawless solution, as multiple entities can be identified within the same `owl:sameAs` query. We have decided to employ a second query to quantify the *inlinks* of each entity identified and select the entity with the highest number of *inlinks*. The solution is not perfect because the entity with most *inlinks* could not be the entity originally mentioned in the domain data.

To obtain the properties of the extracted entities, we employed a query for retrieving 400 properties. Since there is no built-in method for obtaining a ranking of the most relevant properties, we decided to introduce a limit to make our experiments possible. Without a limit, the query along with the domain classification task would take too long to run. Additionally, even though DBpedia Spotlight is a great resource, there are newer and better solutions for entity linking, solutions that should be considered in future work.

As mentioned previously, we relied on an existing model for the domain classification task. On the one hand, the proposed solution is robust in terms of implementation, however, not all properties were classified correctly. A future possibility would be to train a classifier for our needs, or, going in line with recent advances, use a powerful LLM with zero-shot or `n`-shot for the domain classification task.

The LLaMA 2 model was utilized for the generation of prompts from the triples obtained. This decision was judicious. However, some prompts generated did not adhere to the predetermined restrictions, thus we decided not to utilize them in the evaluation of the models. This is a significant challenge to address as we aim to automate the knowledge evaluation of LLMs.

6. Conclusion

LLMs have a significant impact on many sectors, jobs and tasks. They are now part of our lives and their usage will continue to grow. Given this trend, we advocate for the adoption of LLMs in a controlled manner.

We took inspiration from previous works and propose an approach for automating the evaluation of the knowledge in LLMs, based on the dynamic generation of datasets for given domains. In the paper, we describe our vision that encompasses four major steps: extraction of relevant entities, gathering domain properties, dataset generation, and automation of the evaluation.

To materialise our vision, we experimented with different resources for entity linking, knowledge acquisition, classification, and prompt generation. The result can be seen as its first implementation and constitutes important steps towards the automation of the evaluation of LLMs.

Datasets were generated for assessing the presence of knowledge in two domains, finance and medicine, and a range of MLMs were evaluated. Poor performance suggests that domain knowledge is limited in the models tested, which were trained on generic data. But we also note that the adopted zero-shot prompting approach was very straightforward, and did not go through specific prompt engineering. Moreover, the performed experimentation was useful for highlighting some limitations of the current implementation, which will be the focus of future work.

In the future, we intend to address several issues, such as the mapping of entities in DBPedia to Wikidata, which should allow for the gathering of additional domain-related properties, as well as for the generation of better prompts. For instance, we will consider entity linking tools that link directly to Wikidata, as well as the extraction of additional domain terms. We will also analyse datasets generated after different rankings on entities and properties, hopefully focusing more on the target domains, and devise balancing strategies.

The proposed approach will open the door to an easier evaluation of the knowledge in LLMs, thus contributing to faster conclusions on the suitability of different models or prompting strategies.

Therefore, we plan to take advantage of it for further evaluating different state-of-the-art models, e.g., GPT 4 (Achiam et al., 2023), Gemma (Banks and Warkentin, 2024), or to compare the performance of generic models versus models pretrained or fine-tuned in domain data. We further plan to test the impact of different prompting strategies, including few-shot prompts, and different approaches for the automatic generation of prompts.

On top of this, KGs can be created from LLMs

and contribute to alternative human-interpretable representations of these "black-boxes" models. The pros and cons of each representation, or of their combination, should be further analysed, not only in terms of performance in different tasks, but also on aspects like transparency, consistency, and computational cost.

Acknowledgements

This work was partially supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI; and by FCT –Foundation for Science and Technology, I.P., within the scope of the project CISUC –UID/CEC/00326/2020 and by the European Social Fund, through the Regional Operational Program Centro 2020.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jeanine Banks and Tris Warkentin. 2024. Gemma: Introducing new state-of-the-art open models.

Zied Bouraoui, Jose Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7456–7463.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. 2013. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Benedikt Fecher, Marcel Hebing, Melissa Laufer, Jörg Pohle, and Fabian Sofsky. 2023. Friend or foe? exploring the implications of large language models on the science system. *Ai & Society*, pages 1–13.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Procs of NAACL 2016 Student Research Workshop*, pages 8–15. ACL.

Hugo Gonçalo Oliveira and Ricardo Rodrigues. 2023. GPT3 as a Portuguese Lexical Knowledge Base? In *Proceedings of the 4th Conference on Language, Data and Knowledge (LDK 2023), Vienna, Austria*, pages 358–363. NOVA CLUNL.

Dagmar Gromann, Hugo Gonçalo Oliveira, Lucia Pitarch, Elena-Simona Apostol, Jordi Bernad, Eliot Bytyçi, Chiara Cantone, Sara Carvalho, Francesca Frontini, Radovan Garabik, Jorge Gracia, Letizia Granata, Fahad Khan, Timotej Knez, Penny Labropoulou, Chaya Liebeskind, Maria Pia di Buono, Ana Ostroški Anić, Sigita Rackevičienė, Ricardo Rodrigues, Gilles Sérasset, Linas Selmistraitis, Mahammadou Sidibé, Purificação Silvano, Blerina Spahiu, Enriketa Sogutlu, Ranka Stanković, Ciprian-Octavian Truică, Giedrė Valūnaitė Oleškevičienė, Slavko Zitnik, and Katerina Zdravkova. 2024. MultiLexBATS: Multilingual Dataset of Lexical Semantic Relations. In *Proceedings of LREC-COLING (to appear)*. ELRA.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

Jan-Christoph Kalo, Sneha Singhania, Simon Razniewski, and Jeff Z Pan. 2023. Lm-kbc 2023: 2nd challenge on knowledge base construction from pre-trained language models. In *Joint proceedings of 1st workshop on Knowledge Base Construction from Pre-Trained Language Models (KBC-LM) and the 2nd challenge on Language Models for Knowledge Base Construction (LM-KBC)*, volume 3577 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu

Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Timothee Mickus, Eduardo Calò, Léo Jacqmin, Denis Paperno, and Mathieu Constant. 2023. „Mann "is to "Donna" as「国王」is to « Reine » Adapting the Analogy Task for Multilingual and Contextual Embeddings. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 270–283, Toronto, Canada. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Nikil Ravi, Pranshu Chaturvedi, EA Huerta, Zhengchun Liu, Ryan Chard, Aristana Scourtas, KJ Schmidt, Kyle Chard, Ben Blaiszik, and Ian Foster. 2022. Fair principles for ai models with a practical application for accelerated high energy diffraction microscopy. *Scientific Data*, 9(1):657.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sneha Singhania, Tuan-Phong Nguyen, and Simon Razniewski. 2022. Lm-kbc: Knowledge base construction from pre-trained language models. 3274.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Anna Strasser. 2023. On pitfalls (and advantages) of sophisticated large language models. *arXiv preprint arXiv:2303.17511*.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*, pages 20841–20855. PMLR.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Wenxuan Wang, Juluan Shi, Zhaopeng Tu, Youliang Yuan, Jen tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2024. The earth is flat? unveiling factual errors in large language models.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. 2023. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*.