KAMEL 🐪: Knowledge Analysis with Multitoken Entities in Language Models

Jan-Christoph Kalo J.C.KALO@VU.NL
Knowledge Representation and Reasoning Group, Vrije Universiteit Amsterdam

Leandra Fichtel L.FICHTEL@TU-BS.DE
Institute for Information Systems, Technische Universität Braunschweig

Abstract

Large language models (LMs) have been shown to capture large amounts of relational knowledge from the pre-training corpus. These models can be probed for this factual knowledge by using cloze-style prompts as demonstrated on the LAMA benchmark. However, recent studies have uncovered that results only perform well, because the models are good at performing *educated guesses* or recalling facts from the training data. We present a novel Wikidata-based benchmark dataset, KAMEL 🐪, for probing relational knowledge in LMs. In contrast to previous datasets, it covers a broader range of knowledge, probes for single-, and multi-token entities, and contains facts with literal values. Furthermore, the evaluation procedure is more accurate, since the dataset contains alternative entity labels and deals with higher-cardinality relations. Instead of performing the evaluation on masked language models, we present results for a variety of recent causal LMs in a few-shot setting. We show that indeed novel models perform very well on LAMA, achieving a promising F1-score of 52.90%, while only achieving 17.62% on KAMEL. Our analysis shows that even large language models are far from being able to memorize all varieties of relational knowledge that is usually stored knowledge graphs.

1. Introduction

In recent years, researchers have started exploring the capabilities of LMs to store relational knowledge. The seminal paper, *Language Models as Knowledge Bases?* has shown that pre-trained LMs can be probed for a factual triple, e.g. (Paris, capital, France) by simply transforming the triple into a cloze-style sentence: *Paris is the capital of [MASK].* to probe a masked LM like BERT and RoBERTa for relational knowledge typically stored in large knowledge graphs [Petroni et al., 2019]. Most research focused on the T-REx subset of the LAMA dataset consisting of 41 Wikidata relations. The best model in the seminal paper (BERT-large) achieved a P@1 of 32.3% in completing the cloze-style sentences on LAMA. After the publication of the original benchmark dataset, a variety of domain-specific knowledge probing datasets like BioLAMA [Sung et al., 2021], MedLAMA [Meng et al., 2022], and KMIR [Gao et al., 2022] have been published. They show that there is a large interest in exploring how much relational knowledge is stored in pre-trained LMs. Recently, also the idea of knowledge base completion with pre-trained LMs has been investigated [Alivanistos et al., 2022, Li et al., 2022].

Since the quality of the model's predictions strongly depends on the cloze-style prompt that was manually defined for the LAMA dataset, different techniques for automatic prompt learning [Jiang et al., 2020, Bouraoui et al., 2020, Shin et al., 2020, Zhong et al., 2021] or

fine-tuning [Fichtel et al., 2021] were proposed. Given an additional training dataset with triples, these techniques can improve the performance on LAMA to 48.6% [Zhong et al., 2021]. However, two recent papers have shown that the actual performance of the LMs on LAMA is mostly due to *educated guessing* [Cao et al., 2021] and due to *recalling* knowledge from the training dataset [Zhong et al., 2021]. Even randomly initialized models can get a very high precision on the test dataset by memorizing the training dataset. Also, no official training dataset which can be used to optimize prompts is available, so people relied on the dataset provided by Shin et al. [Shin et al., 2020].

LAMA is mostly evaluated only with masked LMs, even though the original paper also evaluated a couple of other kinds of models. Evaluating causal LMs is problematic since no standard evaluation routine exists that has shown good performance. To overcome existing limitations of LAMA and to enable probing of causal LMs with typical Wikidata-like knowledge, we present a new dataset: **KAMEL** 🐪. KAMEL comprises knowledge about 234 relations from Wikidata with a large training, validation, and test dataset. We make sure that all facts are also present in Wikipedia so that they have been seen during the pre-training procedure of the LMs we are probing. Most importantly we overcome the limitations of existing probing datasets by (1) having a larger variety of knowledge graph relations, (2) it contains single- and multi-token entities, (3) we use relations with literals, and (4) have alternative labels for entities. (5) Furthermore, we created an evaluation procedure for higher cardinality relations, which was missing in previous works, and (6) make sure that the dataset can be used for causal LMs.

We evaluate how a large variety of causal LMs: GPT2-XL, OPT-1.3b, OPT-6.7b, OPT-13b, and GPT-J-6b. We show that, as suspected, the results on the original LAMA benchmark overestimate the performance of LMs for Wikipedia-like knowledge by comparing the results from LAMA to the results on our KAMEL dataset. While state-of-the-art models achieve an F1 score of more than 50% on LAMA, these models only achieve 17.7% on KAMEL. Furthermore, the results on KAMEL strongly vary between different types of relations. While for some relations, we achieve an F1-score of 93.00%, others end up with only 0%. Our dataset, the evaluation scripts, and the scripts for dataset creation are available on GitHub[1]

2. Related Work

Recent works have shown that LMs contain relational knowledge as contained in knowledge graphs [Petroni et al., 2019, Roberts et al., 2020]. The idea by Petroni et al. is to probe a masked LM for triples/facts from a knowledge graph as follows: The triple (Barack Obama, speaksLanguage, English) can be translated into the sentence *Barack Obama can speak [MASK].* The goal is to complete the triple, given the subject entity (`Barack Obama`) and the relation (`speaksLanguage`) by predicting a single object entity. When prompting the masked LM with the masked sentence, it is returning an ordered list of (single token) words. If the top prediction is *English*, we assume that the LM *knows* the respective fact. When the model, however, predicts *Indonesian*, the model's predictions would be counted as incorrect, even though Obama actually speaks both languages as stated in Wikidata. The research on probing LMs for factual knowledge has sparked further research investigating

1. https://github.com/JanKalo/KAMEL

how relational knowledge can be extracted from large LMs [Safavi and Koutra, 2021] and how this knowledge can be used to support knowledge graphs [Razniewski et al., 2021].

The original LAMA Dataset The first dataset for knowledge analysis in LMs was **LAMA** [Petroni et al., 2019]. LAMA consists of three subsets: (1) facts from GoogleRE, (2) T-REx, and (3) ConceptNet. However, most follow-up works focused on the T-REx probing subset with 41 Wikidata relations with at most 1000 triples per relation. The object of these triples (the answers) consists of a single token in the vocabulary of the used LMs. Hence, the comparison between different models is usually difficult and requires intersecting vocabularies from different models to achieve comparable results. LAMA comprises 1-1, 1-n, and n-m relations and usually uses averaged Precision@k as a metric. Follow-up papers usually only report Precision@1. While the original LAMA paper only created a test dataset, follow-up papers have suggested different training datasets that could be used to develop better methods for prompting LMs (c.f. the next paragraph). The most used training dataset has been created by Shin et al. [Shin et al., 2020]. This dataset contains at most 1000 triples per relation from Wikidata triples. Even though the prediction quality of LMs to complete knowledge graph triples seem to be very promising, recent findings have shown that many improvements might come from memorization from this training dataset [Cao et al., 2021, Zhong et al., 2021]. Zhong et al. have pointed out, that the distributions of LAMA and the training set are highly skewed [Zhong et al., 2021]. Even a simple baseline that always predicts the majority entity label from the training data can already achieve an accuracy of 17.3%. Also randomly initialized LMs can already achieve 21% accuracy if fine-tuned on the training data. These critical results show that it is still unclear how much factual knowledge is actually contained in pre-trained LMs. Therefore, we aim at overcoming the limitations of LAMA to better investigate the question of whether LMs could serve as knowledge graphs.

Other Existing Benchmark Datasets A more difficult version of LAMA is **LAMA-UHN** [Poerner et al., 2019]. It removes simple triples where, e.g., nationality, can be guessed from the person's name with high probability. More recently, also domain specific probing datasets for the biological(**BioLAMA** [Sung et al., 2021]) and medical(**MedLAMA** [Meng et al., 2022]) domain have been created. **TempLAMA** is adding a factual domain to facts [Dhingra et al., 2021]. Similar to our work, **FewShot-LAMA** evaluates the few-shot capability of LMs by prompting the model with a couple of example prompts [He et al., 2021]. A very recent benchmark dataset for probing language modes additionally checks their knowledge reasoning capabilities on OWL-based axioms: **KMIR** [Gao et al., 2022].

Prompt Learning While the original LAMA paper used one single manually created prompt template per relation, automatically learned prompts can improve the prediction quality significantly. Early approaches worked on mining sentences from large text corpora and weighting them based on training data [Bouraoui et al., 2020, Jiang et al., 2020]. The predictions can even be further improved by learning the prompts through backpropagation for creating a discrete prompt [Shin et al., 2020, Haviv et al., 2021]. More recent approaches have learned only the embeddings instead of discrete prompts [Zhong et al., 2021, Qin and Eisner, 2021] A different way to boost the quality of fact extraction from LMs is by providing an additional context paragraph [Petroni et al., 2020]. Instead of learning the prompt from training data, it is also possible to perform adaptive fine-tuning by continuing training with

the pre-training objective on triple training data [Fichtel et al., 2021]. Instead of learning prompts and predicting words as for the other probing approaches, in [Meng et al., 2022], LM's embeddings are used for probing.

Question Answering Probing a language model for factual knowledge is also related to question answering. On the one hand, our work is related to knowledge graph question answering (KGQA). Its goal is to answer natural language questions from a structured knowledge graph[Cao et al., 2022]. While it has some similarities with our task (answering natural language questions about knowledge graph facts), KGQA is mostly about translating natural language queries into a structured query language, e.g., SPARQL.

Closed-book question answering[Roberts et al., 2020] is about answering natural language questions directly from a fine-tuned language model without using any external data source. This is pretty similar to probing a language model for knowledge graph facts, however, the goal is different. We think that there is more value to probing a language model for relational knowledge than just evaluating a downstream task like question answering because it actually gives the possibility to compare language models and knowledge graphs tasks independently. The insights from this analysis help understand how far LM and KG can complement each other in a variety of downstream tasks. The insights from this work therefore can be applied to a variety of tasks, e.g., question answering, knowledge base completion, knowledge base error detection, and relation extraction.

3. Creating the Probing Dataset and Evaluating Causal Language Models

The basic idea of this work is to create and evaluate a new dataset to probe relational world knowledge in large LMs. We focus particularly on causal models which have gained more attention than masked models over the previous years. One basic requirement for probing relational knowledge was that we only probe facts that the LMs could have seen during the pre-training phase. Thus, our goal was to only use facts that we know have been mentioned in Wikipedia, a corpus that is part of most training corpora. With a similar idea in mind, the LAMA dataset was based on T-REx [Elsahar et al., 2018], a distantly supervised dataset between Wikidata and Wikipedia. Unfortunately, T-REx has a rather low quality, due to the low entity extraction quality of the tool DBpedia Spotlight.

3.1 Dataset Creation

To overcome this drawback, we decided to use a novel tool for annotating large text corpora with Wikidata triples as presented in [Huguet Cabot and Navigli, 2021]: The cRocoDiLe tool uses entity annotations based on Wikipedia hyperlinks, is also distantly supervised, but additionally uses a textual entailment step to filter out distant supervision noise and improve the overall quality. Furthermore, we extended the original corpus which was used only on Wikipedia abstracts to the complete English Wikipedia from the Wikipedia version of 2022-03-21. For Wikidata entity linking we use the dump from 2021-12-11. Overall, this procedure leads to 9,872,196 distinct triples and 1493 different Wikidata relations. In contrast, T-REx contains only 371 relations with overall 382,900 distinct triples.

Since we want to provide a difficult probe, we remove those triples where the full object entity label is contained in the subject label. To assure this, n-grams., n-1-grams, ..., 1-

grams of the subject labels are created where n is the number of words in the object label. E.g. the triple (irrational number, opposite of, rational number) is not filtered out because the object label has two words, $n = 2$, and the 2-grams and 1-grams of the subject are: [irrational number; irrational; number]. Hence, the object label *rational number* is a subset of the subject label *irrational number*, but not part of the list of n-grams, so the triple is not removed from our dataset. Whereas the triple (Kenya national rugby league team; sport; rugby league) is removed because the object *rugby league* is in the list of n-grams and it would be too easy for an LM to predict this.

Furthermore, we manually removed the relations that were not suitable for the task. (1) Relations with literals that are no single integers. All date relations were transformed into years only. If a relation can be expressed in multiple formats and is ambiguous without a metric, we removed it from the dataset as well. (2) We also removed relations about Wikidata meta information, (3) relations that are hard to interpret without using its Wikidata qualifiers, and (4) relations that are unsuitable for LM probing since they are too general concepts. A complete list of relations that were removed manually by us and the reasons can be found in the appendix in Table 5.

From the remaining triples of the filtered relations, we created *queries* consisting of one subject and one relation. The *answer* consists of a set of object entities. For example, considering the relation `P1412`(languages spoken, written, or signed) and the triples (Barack Obama, `P1412`, English) and (Barack Obama, P1412, Indonesian), we create the query (Barack Obama, `P1412`, ?). The answer is the set of objects: *English, Indonesian*.

We allow subject and object labels of arbitrary lengths. Furthermore, we do not only use rdf:labels for each of the objects but add all alternative labels as well. When we query the language model for *What countries has Taylor Chorney played for?*, the answer entity is `Q30`. It has the rdf:label *United States of America*. But in Wikidata for `Q30` also alternative labels such as *USA* or *the States* are stored. Using also these alternative labels for evaluation is necessary, because many language models actually predict *USA*, instead of the full name. As an additional restriction, we only allow queries that have at most 10 different answer entities. Longer answer sets are very rare but require significantly higher computational costs when running the benchmark.

As a final step, we randomly sampled 1400 queries for each of the 234 remaining relations, so that we could create a training set of size 1000, a validation set of size 200, and a test set of size 200. Due to this random sampling, many facts that we probe are actually about unknown entities, that, however, has an article in the English Wikipedia and are in Wikidata.

Dataset	KAMEL	LAMA
Number of Queries	46800	31479
Number of Relations	234	41
Avg. Number of Tokens	4.86	1 (2.87)[2]
Avg. Number of Labels	3.19	1
Queries with Multiple Results	4296	1035
Literals	yes	no

Table 1: Overview of the differences between LAMA and the test set of KAMEL.

Differences between LAMA and KAMEL KAMEL was designed to overcome several limitations of LAMA. We shortly summarize the key statistics in Table 1. LAMA officially only consists of a test set. Follow-up works have created their own development and training datasets. KAMEL, however, provides all splits in an adequate size.

Furthermore, we cover a wider range of knowledge: 41 vs. 234 relations. KAMEL is extracted from a larger dataset. Therefore it covers more triples per relation. It has a substantial number of queries with multiple answers. We think that this is an important requirement for comparing an LM and a KG. Furthermore, KAMEL uses multiple labels for the entities to improve the evaluation quality. Furthermore, LAMA has only evaluated single token entities. As previous research has shown that predicting entities with multiple tokens is more difficult. These single token answers are from the intersection of the vocabularies of a variety of LMs. Consequently, they all are very frequent words and therefore often common entities that are easier to predict. For example, for most geographic relations in LAMA this implies that solutions are often countries or large, well-known cities. Without restricting to single-token entities, as in KAMEL, solutions can also be uncommon entities, e.g., smaller villages or municipalities. Another important difference is that KAMEL contains number literals and not only triples with entities.

3.2 Probing Causal Language Models

Previous work has mainly focused on probing masked LMs, whereas causal LMs have become more popular most recently, these could hardly be probed for relational knowledge using LAMA. Evaluation of masked LMs was rather simple since the LMs were only supposed to predict a single token that was compared to the gold answer. If the top one prediction matched, the triple was counted as correct.

Causal models for text generation can generate texts of arbitrary length. Hence, the evaluation of long predictions is way more complex than for masked LMs. A straightforward idea would be to generate text of a fixed length to answer a factual question. If all correct answers are contained in the generated text, the recall would be 100%. An assessment of the precision is rather difficult since the assessment of wrong answers in the generated text is hard. To overcome this issue, we decided to only perform few-shot evaluations, where the few-shot examples demonstrate how the output should be formatted.

Few-Shot Evaluation For each test triple, we randomly sampled k training triples of the same relation and created prompts presenting how the answers should look like. As an example, for the Wikidata relation `P1412` (*languages spoken, written or signed*) and the test query (Natalie Portman, `P1412`, ?), a 5-shot example could look as follows. We used % as an end-of-sentence token and a semicolon to separate answers.

```
What languages does Barack Obama speak? English;Indonesian%
What languages does Albert Einstein speak? English;German%
What languages does Confucius speak? Old Chinese%
What languages does António Guterres speak? Portuguese;English;Spanish;French%
What languages does Chimamanda Ngozi Adichie speak? English;Igbo;Nigerian Pidgin%

What languages does Natalie Portman speak?
```

2. 1 when using the original LAMA evaluation. With the GPT2 tokenizer, as in our experiments, it is 2.87.

From the prompt examples, the model learns how the answer should be formatted and which end-of-sentence token to use. Hence, in the optimal case, it would generate the text `English;Spanish;Hebrew;Japanese;French%`, leading to precision and recall of 100%. If only a single example is provided, the model does not necessarily stick to the desired format and might answer in a longer sentence, which we will need to consider when computing the evaluation metrics.

Prompt Templates The prompts used throughout this paper have been manually curated by the authors and are available as supplemental material for further evaluations. We have written exactly one single prompt per relation in our dataset, leading to 234 prompts, and an additional single prompt for a relation in the LAMA dataset that is not part of our dataset. All prompts are created as simple and short questions based informed by Wikidata's property descriptions and examples. For prompts asking for a timestamp, we explicitly mention that we only want the *year* as an answer to simplify evaluation.

Evaluation Metrics Since every test instance can have up to 10 answers, we evaluate precision, recall, and F1-measure for every single instance. Given a prediction of a model, e.g., `English; Spanish; French; Italian%` for the Natalie Portman example from above, we would perform a split on the prediction on semicolons. We remove all special characters and then perform an exact match comparison of every single answer part to the gold dataset. Since the prediction, in this case, contains 3 out of 5 results, and one incorrect prediction, we achieve a precision of 75% and a recall of 60%. We make sure to not only compare the original label of the gold entity but also all its alternative labels found on Wikidata. Thus, we can prevent the LM to make correct predictions that are counted as incorrect. In the case that generated text is not having the format taught in the few-shot examples, we employ the same metric. For example, *Natalie Portman speaks English and French.* would get precision and recall of 0%. However, in practice, these cases are extremely rare.

As the last step, we perform macro-averaging per relation and then average the relation's scores to get an overall performance score.

Even though, we strict the analysis in this work to few-shot prompting, we created a large training and validation set that offers the possibility to also perform other kinds of evaluations. In our opinion, it would be particularly interesting to extend recent prompt learning techniques that have been built for LAMA to work on multi-token entities and higher cardinality relations so that they can be evaluated on our dataset as well.

4. Experiments

4.1 Experimental Setup

All our experiments have been performed on our newly created KAMEL dataset using 234 Wikidata relations with 200 instances per relation in the test set, 1000 in the training set, and 200 in the validation set. However, we are not using the validation set in our work, but only performing few-shot learning by taking the few-shot examples randomly from the training set. Additionally, we use LAMA with the training dataset created by Shin et al.([Shin et al., 2020]) as a comparison to our dataset. For all experiments, we have used

	1-shot			5-shot			10-shot		
Model	P	R	F1	P	R	F1	P	R	F1
GPT-J-6b	10.27%	10.20%	10.24%	16.22%	15.81%	16.01%	17.46%	17.13%	17.30%
GPT2-xl	7.11%	7.05%	7.08%	10.11%	10.00%	10.06%	11.10%	11.00%	11.05%
OPT-1.3b	7.02%	6.91%	6.97%	10.87%	10.61%	10.74%	11.50%	11.18%	11.34%
OPT-6.7b	10.19%	10.09%	10.14%	15.65%	15.20%	15.42%	16.67%	16.24%	16.45%
OPT-13b	10.96%	10.88%	**10.92%**	16.42%	16.22%	**16.32%**	17.76%	17.48%	**17.62%**
OPT-13b*	9.63%	9.55%	9.59%	14.72%	14.54%	14.63%	16.21%	15.96%	16.08%

Table 2: Overview of the result of all models and the different few-shot scenarios. OPT-13b* only uses rdf:labels without alternative labels for comparison.

manually created prompts that were written by the authors of this paper. We created only a single prompt per relation that was re-used for all experiments.

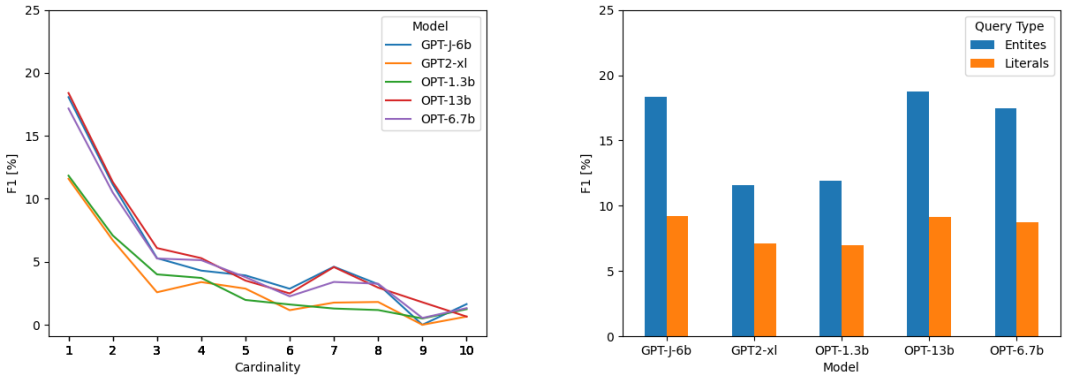
Our experiments comprise five different models from the GPT [Radford et al., 2019, Wang and Komatsuzaki, 2021] and OPT [Zhang et al., 2022] group that were downloaded from the Huggingface model hub. OPT-1.3b and GPT2-xl are the models with the smallest parameter number having 1.5 and 1.3 billion parameters. GPT-J-6b and OPT-6.7b have a similar number of parameters, namely 6.7 billion. OPT-13b is the biggest model in our evaluation with 13 billion parameters. We evaluate them on our test set and calculate precision, recall, and F1 score as explained in Section 3.

The reported numbers are from a single run only, however we choose new few shot examples for every single query instance, randomly. Thus, the standard deviation for averaged scores per model are very low. Our investigation has shown that the standard deviations are below 0.1% in F1-score.

4.2 Results

We first give an overview of the results of the different models and their performance in different few-shot settings presented in Table 2. We first look at the performance of different types of test triples. We analyze the performance for relations with different cardinality and compare queries with literals in comparison to normal entities. Afterward, we have a look at the top and worst-performing relations. Finally, we also present a short analysis of the differences between the results on the LAMA dataset and on our dataset.

Overview Comparing the models performing at the different few-shot scenarios, GPT-J-6b, and OPT-6.7b achieve similar results (c.f. Table 2). This is as expected since they have a roughly equal parameter number. OPT-13b as the biggest model performs best, but does not show a big difference to the results of e.g. GPT-J-6b with 10-shot: OPT-13b achieves only an F1-score of 17.62% compared to 17.30%. Considering the poor results of the smallest model of OPT, which reaches only 12.08% for the F1-score in the 10-shot scenario, a minimum number of parameters is apparently important, so that a higher degree of knowledge can be stored in the LM. However, in our experiments, a large number of parameters does not necessarily correlate with better performance. Looking at the precision and recall values, it is notable that they do not differ much. This is because most of the



(a) (b)

Figure 1: (a) F1-scores for all models against cardinality of queries. (b) F1-scores for all models split into entity and literal queries each

queries have cardinality 1 and therefore only a single answer. Hence, the precision and recall per query are either 0 or 1.

The last row in the table depicts an OPT-13b run with 10 shots, only using rdf:labels without alternative labels. OPT-13b* performs almost 10% (relative to OPT-13b) worse than with alternative labels at 10-shot. This shows that indeed, using alternative labels is giving us a more precise estimation of the knowledge in language models.

Query Cardinality In Figure 1a, we present the average F1 score for queries of different answer sizes (cardinality) for the 5 models in the 10-shot setting. The general trend is that all models perform best for queries with cardinality 1. This category is also dominating our dataset, therefore the F1 scores for cardinality 1 are very close to the average scores presented in Table 2. Queries with higher cardinality naturally become more difficult, which is also reflected in the graph. While queries with cardinality 2 still have an F1 score above 5%, for higher cardinality the average F1 scores are between 1% and 5%. These queries are most often answered incorrectly. One reason for that is that we take random few-shot examples from the training set. These, however, often only comprise a single answer so the model does not know that it is supposed to predict multiple answers. This problem could be possibly prevented by a more elaborate prompt design, which however is not the focus of this work.

Literals and Entities In Figure 1b, we compare the performance between queries that have entities as an answer and queries that have numbers as an answer. The first observation is that number queries seem to be significantly more difficult. While entity queries can be answered with an F1 score above 10%, often even above 15%, number queries often only have half the F1 score. Also within the number relations, the spectrum is large. Queries for some relations can be answered rather well, achieving F1 scores between 10% and 20%. This is either due to a very small number of possible answers, or the subject entity already

ID	Label	F1
P4743	animal breed	93.00%
P30	continent	91.58%
P1412	languages spoken	56.41%
P17	country	55.12%

ID	Label	F1
P47	shares border with	0.00%
P570	date of death	0.00%
P1066	student of	0.00%
P569	date of birth	0.00%

Table 3: Performance of some best and worst relations with OPT-13b and 10-shots.

containing the correct number prediction [3]. On the other hand, there are many relations with numbers, where the F1 score is close to 0%. The year of birth or the year of death of a person is almost always predicted incorrectly.

Detailed Analysis To get a better idea of how the performance is on different Wikidata relations, we present some of the top relations measured by F1 score and some of the worst performing relations in Table 3 [4]. The model achieves the best performance of over 90% for the *animal breed* relations. This is due to data skewness in Wikidata and therefore also in our dataset. Most animals with relation `P4743` are racehorses, all having the same breed *Thoroughbred*. Also, the other 3 relations in our list have very few possible object labels each. Similar to the LAMA dataset, also KAMEL has many triples with the object *Antarctica* for the relation continent, hence the high F1 score. Besides that, also many geographic and language-specific relations have performed very well. This is a pattern that was also already observed on LAMA.

Due to the relations that can easily be answered by just picking the most frequent label from the training dataset. Such a simple baseline already achieves an F1 score of 9.7% in our dataset.

A large number of relations are extremely difficult for the LMs and therefore only achieve an F1 score of 0.00%. As an example, the relation *shares border with*, stating which municipalities and countries share a border with each other. While this is rather simple for countries, most test instances are about bordering municipalities, often very small ones. Such niche knowledge is hardly present in any of the LMs even though these facts are part of their training corpora.

As already mentioned before, relations to numbers, here *years of birth* and *death*, are very difficult. For *birth year*, all models correctly predict a valid yea, consisting of 4 digits. The predictions get better the bigger the model gets. However, the average difference over all test instances for the birth year is 59 years for OPT-13b. For the *death year* this difference is even 84 years. Also, the distribution of predictions for both varies significantly. For the birth year, some predictions are only off a single year, others up to 433 years.

Comparison to LAMA As an additional analysis, we evaluate OPT-13b on LAMA and compare the quality of the same subset of relations in KAMEL. This analysis does not consider `P530`, since this property is not contained in KAMEL, because there were not enough instances. For this experiment, we used the same evaluation procedure for

3. Such easy-to-answer queries are usually removed from our test dataset. Some exceptions, however, were not covered by our removal heuristic
4. A detailed list of all relations can be found in the Appendix Table 6.

Dataset	1-shot			5-shot			10-shot		
	P	R	F1	P	R	F1	P	R	F1
LAMA	39.83%	39.83%	39.83%	50.18%	50.18%	50.18%	52.90%	52.90%	52.90%
KAMEL	16.21%	16.10%	16.15%	22.38%	22.01%	22.19%	24.28%	23.71%	23.99%

Table 4: Performance between 40 LAMA relations and the subset in KAMEL for OPT-13b.

LAMA as for KAMEL. Since LAMA always has a single correct answer per query, precision and recall are always identical. The performance of OPT-13b on LAMA is 52.90% F1 in the best setting with 10-shots (c.f.Table 4). This outperforms all approaches presented for prompt learning on LAMA, even though they usually use around 1000 training examples per relation, instead of only 10. The results on the subset of 40 relations on KAMEL are higher than the average on the whole KAMEL, with an F1 score of 23.99%, but significantly lower than for LAMA. Hence, KAMEL is significantly more difficult; also when only evaluating the same relations as in LAMA.

5. Conclusions

We present KAMEL, a new dataset for probing LMs for relational knowledge, and a simple method on how to probe causal LMs in a few-shot setting. Our new dataset overcomes a variety of limitations of previous benchmark datasets and covers a significantly larger variety of relations. Thus, we can better investigate the memorization of factual world knowledge in pre-trained LMs. In extensive experiments, we show that indeed novel causal models can achieve a very high precision on the LAMA dataset, but only a rather low precision on our new dataset. Concretely, the best model has only achieved an F1 score of 17.62% on KAMEL. While particularly geographic relations are usually memorized incredibly well, knowledge graphs often contain lots of niche knowledge that cannot be recalled by the LM. Also, numeric literals seem to be more difficult than predicting only entities. Overall, our results show that even large recent language models are far from being able to serve *as knowledge graphs*.

For future work, it would be interesting to look into improving the prompts to increase the performance of models on KAMEL. Optimizing the few-shot examples could already give higher F1 scores, and therefore more realistic numbers. Additionally, zero-shot evaluations might be interesting, but so far have to lead to bad results for the small models that we have evaluated. Also using the full training dataset to learn better prompts, similar to what was done for the LAMA dataset, might be an interesting option that might increase the model's performance on our probing dataset. Additionally, we think that our dataset is a prime candidate to further explore knowledge base completion tasks using LMs.

Acknowledgments

This research is partially funded by Huawei Amsterdam Research Center.

References

Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan-Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. Prompting as probing: Using language models for knowledge base construction. *arXiv preprint arXiv:2208.11057*, 2022.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. Inducing Relational Knowledge from BERT. In *Proc. of the Thirty-Fourth Conference on Artificial Intelligence*, AAAI'20, 2020.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases. pages 1860–1874, 2021. doi: 10.18653/v1/2021.acl-long.146.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6101–6119, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.422. URL https://aclanthology.org/2022.acl-long.422.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-Aware Language Models as Temporal Knowledge Bases. 2021. URL http://arxiv.org/abs/2106.15110.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Leandra Fichtel, Jan-Christoph Kalo, and Wolf-Tilo Balke. Prompt Tuning or Fine-Tuning - Investigating Relational Knowledge in Pre-Trained Language Models. In *Automatic Knowledge Base Construction (AKBC)*, pages 1–15, 2021. URL https://openreview.net/pdf?id=o7sMlpr9yBW.

Daniel Gao, Yantao Jia, Lei Li, Chengzhen Fu, Zhicheng Dou, Hao Jiang, Xinyu Zhang, Lei Chen, and Zhao Cao. Kmir: A benchmark for evaluating knowledge memorization, identification and reasoning abilities of language models, 2022. URL https://arxiv.org/abs/2202.13529.

Adi Haviv, Jonathan Berant, and Amir Globerson. BERTese: Learning to speak to BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3618–3623, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.316. URL https://aclanthology.org/2021.eacl-main.316.

Tianxing He, Kyunghyun Cho, and James Glass. An Empirical Study on Few-shot Knowledge Probing for Pretrained Language Models. (2), sep 2021. URL http://arxiv.org/abs/2109.02772.

Pere-Lluís Huguet Cabot and Roberto Navigli. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.204. URL https://aclanthology.org/2021.findings-emnlp.204.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How Can We Know What Language Models Know? In *Transactions of the Association for Computational Linguistics 2020 (TACL)*, volume 8, pages 423–438, 2020. doi: 10.1162/tacl_a_00324. URL https://doi.org/10.1162/tacl{_}a{_}00324.

Tianyi Li, Wenyu Huang, Nikos Papasarantopoulos, Pavlos Vougiouklis, and Jeff Z. Pan. Task-specific pre-training and prompt decomposition for knowledge graph population with language models. *ArXiv*, abs/2208.12539, 2022.

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4798–4810, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.329. URL https://aclanthology.org/2022.acl-long.329.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language Models as Knowledge Bases? In *Proc. 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, nov 2019.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. How Context Affects Language Models' Factual Predictions. In *Automatic Knowledge Base Construction (AKBC)*, 2020.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. BERT is Not a Knowledge Base (Yet): Factual Knowledge vs. Name-Based Reasoning in Unsupervised QA. 0, 2019. URL http://arxiv.org/abs/1911.03681.

Guanghui Qin and Jason Eisner. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online, June 2021. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Simon Razniewski, Andrew Yates, Nora Kassner, and Gerhard Weikum. Language Models As or For Knowledge Bases. 2021. ISSN 1613-0073. URL http://arxiv.org/abs/2110.04888.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.437. URL https://aclanthology.org/2020.emnlp-main.437.

Tara Safavi and Danai Koutra. Relational World Knowledge Representation in Contextual Language Models: A Review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1053–1067, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.81. URL https://aclanthology.org/2021.emnlp-main.81.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. pages 4222–4235, 2020. doi: 10.18653/v1/2020.emnlp-main.346.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. Can Language Models be Biomedical Knowledge Bases? pages 4723–4734, 2021. doi: 10.18653/v1/2021.emnlp-main.388.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5017–5033, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.398. URL https://aclanthology.org/2021.naacl-main.398.

Appendix A. Additional Results

A.1 Manually removed relations

Relation ID	Relation Label	Reason
P1269	facet of	Wikidata meta-fact
P793	significant event	requires qualifiers
P1352	ranking	requires qualifiers
P609	terminus location	requires qualifiers
P1343	described by source	requires qualifiers
P2283	uses	too general concept
P1889	different from	most subjects and objects have the same label
P460	said to be the same	most subjects and objects have the same label
P2257	event interval	requires a metric
P2067	mass	requires a metric
P2043	length	requires a metric
P2046	area	requires a metric

Table 5: Relations that were removed from the dataset manually and reasons for their removal.

A.2 Detailed Results per Relation for OPT-13b

ID	Label	Precision	Recall	F1 %
P4743	animal breed	93.00%	93.00%	93.00 %
P541	office contested	92.50%	92.25%	92.37 %
P30	continent	92.00%	91.17%	91.58 %
P8875	indexed in bibliographic review	94.75%	84.77%	89.48 %
P105	taxon rank	88.50%	88.50%	88.50 %
P467	legislated by	81.00%	81.00%	81.00 %
P4884	court	80.00%	80.00%	80.00 %
P196	minor planet group	78.50%	77.00%	77.74 %
P37	official language	67.53%	62.88%	65.12 %
P103	native language	63.00%	62.50%	62.75 %
P1412	languages spoken, written or signed	57.42%	55.43%	56.41 %
P17	country	55.50%	54.75%	55.12 %
P765	surface played on	54.00%	53.25%	53.62 %
P414	stock exchange	54.00%	52.50%	53.24 %
P1103	number of platform tracks	51.00%	50.75%	50.87 %
P407	language of work or name	49.50%	48.42%	48.95 %
P412	voice type	49.50%	48.25%	48.87 %
P27	country of citizenship	46.50%	46.25%	46.37 %
P1435	heritage designation	45.50%	45.50%	45.50 %
P664	organizer	44.50%	44.00%	44.25 %
P2597	Gram staining	44.00%	44.00%	44.00 %
P172	ethnic group	42.25%	42.50%	42.37 %
P461	opposite of	42.50%	42.00%	42.25 %

P6886	writing language	42.25%	40.75%	41.49 %
P364	original language of film or TV show	41.00%	41.00%	41.00 %
P1532	country for sport	40.50%	39.75%	40.12 %
P277	programming language	40.54%	37.17%	38.78 %
P1001	applies to jurisdiction	38.50%	38.25%	38.37 %
P641	sport	37.50%	37.50%	37.50 %
P991	successful candidate	37.50%	37.00%	37.25 %
P2094	competition class	38.00%	36.46%	37.21 %
P1971	number of children	37.00%	37.00%	37.00 %
P495	country of origin	37.00%	36.75%	36.87 %
P2348	time period	37.00%	36.00%	36.49 %
P7937	form of creative work	35.50%	35.50%	35.50 %
P2437	number of seasons	35.50%	35.25%	35.37 %
P7959	historic county	33.00%	33.00%	33.00 %
P306	operating system	34.43%	30.66%	32.44 %
P59	constellation	32.00%	32.00%	32.00 %
P5353	school district	31.50%	31.00%	31.25 %
P140	religion	31.00%	31.00%	31.00 %
P607	conflict	30.78%	30.14%	30.46 %
P126	maintained by	30.00%	30.00%	30.00 %
P1308	officeholder	30.50%	29.38%	29.93 %
P115	home venue	30.00%	29.67%	29.83 %
P183	endemic to	29.50%	29.50%	29.50 %
P1444	destination point	26.50%	26.50%	26.50 %
P2936	language used	22.78%	29.23%	25.61 %
P291	place of publication	25.50%	25.25%	25.37 %
P945	allegiance	24.00%	23.75%	23.87 %
P937	work location	24.10%	23.58%	23.84 %
P1877	after a work by	23.50%	23.50%	23.50 %
P241	military branch	23.50%	23.25%	23.37 %
P1350	number of matches played/races/starts	23.50%	21.42%	22.41 %
P186	made from material	22.75%	21.92%	22.33 %
P31	instance of	22.00%	22.00%	22.00 %
P1303	instrument	22.53%	19.31%	20.80 %
P118	league	21.00%	20.50%	20.75 %
P53	family	20.50%	20.25%	20.37 %
P1441	present in work	20.54%	18.17%	19.28 %
P2868	subject has role	19.42%	18.75%	19.08 %
P177	crosses	19.00%	19.00%	19.00 %
P159	headquarters location	18.50%	18.50%	18.50 %
P102	member of political party	18.25%	18.00%	18.12 %
P413	position played on team / speciality	18.00%	18.00%	18.00 %
P452	industry	18.00%	17.50%	17.75 %
P113	airline hub	18.00%	17.17%	17.57 %
P286	head coach	17.50%	17.50%	17.50 %
P611	religious order	17.50%	17.50%	17.50 %
P97	noble title	17.50%	17.50%	17.50 %
P1027	conferred by	18.00%	17.00%	17.49 %
P137	operator	17.00%	16.75%	16.87 %
P931	place served by transport hub	16.50%	16.50%	16.50 %
P3450	sports season of league or competition	16.50%	16.50%	16.50 %

P141	IUCN conservation status	16.00%	16.00%	16.00 %
P400	platform	15.55%	16.05%	15.80 %
P1376	capital of	15.35%	15.42%	15.38 %
P176	manufacturer	15.00%	15.00%	15.00 %
P2341	indigenous to	14.68%	14.42%	14.55 %
P6	head of government	14.50%	14.50%	14.50 %
P156	followed by	14.50%	14.25%	14.37 %
P178	developer	14.25%	13.75%	14.00 %
P749	parent organization	14.12%	13.75%	13.93 %
P1427	start point	13.50%	13.50%	13.50 %
P3764	pole position	13.50%	13.50%	13.50 %
P1142	political ideology	13.42%	13.17%	13.29 %
P36	capital	13.00%	13.00%	13.00 %
P201	lake outflow	13.00%	13.00%	13.00 %
P921	main subject	13.00%	12.75%	12.87 %
P1050	medical condition	13.00%	12.75%	12.87 %
P276	location	13.00%	12.50%	12.75 %
P415	radio format	12.50%	12.50%	12.50 %
P20	place of death	12.50%	12.50%	12.50 %
P65	site of astronomical discovery	12.50%	12.50%	12.50 %
P206	located in or next to body of water	13.00%	11.83%	12.39 %
P840	narrative location	12.00%	12.00%	12.00 %
P410	military rank	12.00%	12.00%	12.00 %
P127	owned by	11.50%	11.50%	11.50 %
P1433	published in	11.50%	11.50%	11.50 %
P371	presenter	12.88%	10.34%	11.47 %
P170	creator	11.50%	10.75%	11.11 %
P4552	mountain range	11.00%	11.00%	11.00 %
P740	location of formation	11.00%	11.00%	11.00 %
P674	characters	8.37%	15.93%	10.97 %
P149	architectural style	11.00%	10.75%	10.87 %
P136	genre	11.00%	10.25%	10.61 %
P3018	located in protected area	10.50%	10.50%	10.50 %
P200	inflows	9.92%	10.50%	10.20 %
P1416	affiliation	10.00%	10.00%	10.00 %
P179	part of the series	10.00%	10.00%	10.00 %
P106	occupation	10.00%	9.42%	9.70 %
P135	movement	10.00%	9.33%	9.66 %
P800	notable work	10.47%	8.83%	9.58 %
P2632	place of detention	10.25%	8.29%	9.17 %
P708	diocese	9.00%	9.00%	9.00 %
P1132	number of participants	9.00%	9.00%	9.00 %
P123	publisher	9.00%	9.00%	9.00 %
P915	filming location	9.50%	8.50%	8.97 %
P449	original broadcaster	9.00%	8.75%	8.87 %
P3602	candidacy in election	8.52%	8.67%	8.59 %
P2416	sports discipline competed in	8.50%	8.50%	8.50 %
P1056	product or material produced	8.50%	8.50%	8.50 %
P1923	participating team	8.23%	8.56%	8.39 %
P551	residence	8.50%	7.92%	8.20 %
P1411	nominated for	9.29%	7.25%	8.14 %

P7153	significant place	8.00%	8.00%	8.00 %
P750	distributed by	8.00%	8.00%	8.00 %
P131	located in the administrative territorial entity	7.75%	8.00%	7.87 %
P50	author	8.00%	7.75%	7.87 %
P703	found in taxon	7.46%	7.75%	7.60 %
P710	participant	6.80%	8.46%	7.54 %
P2522	victory	7.50%	7.25%	7.37 %
P272	production company	7.50%	7.25%	7.37 %
P61	discoverer or inventor	7.50%	7.25%	7.37 %
P6379	has works in the collection	6.94%	7.58%	7.25 %
P138	named after	7.25%	7.25%	7.25 %
P488	chairperson	7.50%	6.75%	7.11 %
P19	place of birth	7.00%	7.00%	7.00 %
P463	member of	7.50%	6.50%	6.96 %
P669	located on street	7.00%	6.75%	6.87 %
P112	founded by	7.00%	6.04%	6.49 %
P279	subclass of	6.25%	6.25%	6.25 %
P366	use	6.50%	6.00%	6.24 %
P361	part of	6.17%	6.25%	6.21 %
P39	position held	6.50%	5.75%	6.10 %
P155	follows	6.00%	6.00%	6.00 %
P427	taxonomic type	6.00%	6.00%	6.00 %
P647	drafted by	6.00%	6.00%	6.00 %
P585	point in time	6.00%	6.00%	6.00 %
P101	field of work	5.83%	5.92%	5.87 %
P195	collection	6.00%	5.75%	5.87 %
P610	highest point	5.50%	5.50%	5.50 %
P1113	number of episodes	5.50%	5.50%	5.50 %
P144	based on	5.50%	5.50%	5.50 %
P180	depicts	4.93%	5.67%	5.27 %
P466	occupant	6.25%	4.42%	5.18 %
P1366	replaced by	5.00%	5.00%	5.00 %
P287	designed by	5.00%	4.75%	4.87 %
P69	educated at	4.92%	4.29%	4.58 %
P509	cause of death	4.50%	4.50%	4.50 %
P1346	winner	4.50%	4.12%	4.30 %
P1344	participant in	3.72%	4.75%	4.18 %
P108	employer	5.07%	3.50%	4.14 %
P706	located in/on physical feature	4.00%	4.00%	4.00 %
P175	performer	4.00%	4.00%	4.00 %
P2031	work period (start)	4.00%	4.00%	4.00 %
P580	start time	4.00%	4.00%	4.00 %
P676	lyrics by	3.75%	3.50%	3.62 %
P2044	elevation above sea level	3.50%	3.50%	3.50 %
P88	commissioned by	3.50%	3.50%	3.50 %
P576	dissolved, abolished or demolished date	3.50%	3.50%	3.50 %
P2032	work period (end)	3.50%	3.50%	3.50 %
P87	librettist	3.50%	3.25%	3.37 %
P86	composer	3.50%	3.25%	3.37 %
P40	child	2.79%	2.85%	2.82 %
P2975	host	2.67%	2.75%	2.71 %

P264	record label	3.00%	2.42%	2.68 %
P54	member of sports team	2.25%	2.95%	2.55 %
P355	subsidiary	2.47%	2.38%	2.42 %
P1192	connecting service	2.16%	2.50%	2.32 %
P451	unmarried partner	2.17%	2.00%	2.08 %
P119	place of burial	2.00%	2.00%	2.00 %
P571	inception	2.00%	2.00%	2.00 %
P582	end time	2.00%	2.00%	2.00 %
P22	father	2.00%	2.00%	2.00 %
P57	director	2.00%	2.00%	2.00 %
P1408	licensed to broadcast to	2.00%	2.00%	2.00 %
P559	terminus	1.75%	2.25%	1.97 %
P2789	connects with	1.76%	2.08%	1.91 %
P1365	replaces	2.00%	1.75%	1.87 %
P81	connecting line	2.00%	1.75%	1.87 %
P1830	owner of	1.49%	1.89%	1.67 %
P84	architect	1.50%	1.50%	1.50 %
P403	mouth of the watercourse	1.50%	1.50%	1.50 %
P171	parent taxon	1.50%	1.50%	1.50 %
P1101	floors above ground	1.50%	1.50%	1.50 %
P1082	population	1.50%	1.50%	1.50 %
P162	producer	1.50%	1.50%	1.50 %
P4908	season	1.50%	1.50%	1.50 %
P4647	location of first performance	1.50%	1.50%	1.50 %
P575	time of discovery or invention	1.50%	1.50%	1.50 %
P166	award received	1.18%	1.92%	1.46 %
P737	influenced by	1.14%	1.50%	1.30 %
P6087	coach of sports team	1.14%	1.42%	1.27 %
P58	screenwriter	1.25%	1.25%	1.25 %
P1598	consecrator	1.17%	1.17%	1.17 %
P150	contains administrative territorial entity	0.96%	1.09%	1.02 %
P25	mother	1.00%	1.00%	1.00 %
P606	first flight	1.00%	1.00%	1.00 %
P1619	date of official opening	1.00%	1.00%	1.00 %
P729	service entry	1.00%	1.00%	1.00 %
P577	publication date	1.00%	1.00%	1.00 %
P161	cast member	0.48%	1.39%	0.72 %
P1038	relative	0.54%	1.00%	0.70 %
P197	adjacent station	1.00%	0.50%	0.67 %
P3999	date of official closure	0.50%	0.50%	0.50 %
P619	UTC date of spacecraft launch	0.50%	0.50%	0.50 %
P344	director of photography	0.50%	0.50%	0.50 %
P802	student	0.50%	0.50%	0.50 %
P98	editor	0.50%	0.50%	0.50 %
P1327	partner in business or sport	0.50%	0.50%	0.50 %
P3373	sibling	0.50%	0.50%	0.50 %
P527	has part	0.40%	0.46%	0.43 %
P974	tributary	0.29%	0.75%	0.42 %
P190	twinned administrative body	0.25%	0.50%	0.33 %
P184	doctoral advisor	0.00%	0.00%	0.00 %
P26	spouse	0.00%	0.00%	0.00 %

P185	doctoral student	0.00%	0.00%	0.00 %
P1249	time of earliest written record	0.00%	0.00%	0.00 %
P1191	date of first performance	0.00%	0.00%	0.00 %
P47	shares border with	0.00%	0.00%	0.00 %
P570	date of death	0.00%	0.00%	0.00 %
P1066	student of	0.00%	0.00%	0.00 %
P569	date of birth	0.00%	0.00%	0.00 %

Table 6: Detailed results for OPT-13b for all the properties in KAMEL ordered by F1 score.