BEAR: A Unified Framework for Evaluating Relational Knowledge in Causal and Masked Language Models

Jacek Wiland* **Max Ploner*** **Alan Akbik**

Humboldt Universität zu Berlin
Science Of Intelligence
{jacek.wiland, max.ploner, alan.akbik}@hu-berlin.de

Abstract

Knowledge probing assesses to which degree a language model (LM) has successfully learned relational knowledge during pre-training. Probing is an inexpensive way to compare LMs of different sizes and training configurations. However, previous approaches rely on the objective function used in pre-training LMs and are thus applicable only to masked or causal LMs. As a result, comparing different types of LMs becomes impossible. To address this, we propose an approach that uses an LM's inherent ability to estimate the log-likelihood of any given textual statement. We carefully design an evaluation dataset of 7,731 instances (40,916 in a larger variant) from which we produce alternative statements for each relational fact, one of which is correct. We then evaluate whether an LM correctly assigns the highest log-likelihood to the correct statement. Our experimental evaluation of 22 common LMs shows that our proposed framework, BEAR, can effectively probe for knowledge across different LM types. We release the BEAR datasets and an open-source framework that implements the probing approach to the research community to facilitate the evaluation and development of LMs.

1 Introduction

Pre-trained language models (LMs) are the backbone of current state-of-the-art NLP approaches. A key property is the syntactic and semantic knowledge stored in their internal parameters, allowing them to generalize beyond given training data when fine-tuning for a specific downstream NLP task. Due to their importance and the large number of proposed LMs, prior work has sought to improve the ability to measure the amount of factual knowledge encoded in LMs, thereby facilitating the comparison of different LMs (Petroni et al., 2019; Po-

---

*Equal contribution

The capital of France is [MASK].

(a) LAMA probe: Single-subtoken mask prediction

The capital of Uganda is Thimphu.
... Kampala.
... Buenos Aires.
... Bandar Seri Begawan.

(b) BEAR probe: Rank answer options of arbitrary length

Figure 1: Comparison of the LAMA and BEAR probes. Both probes query LMs given a template (here in black), the subject of the relation (blue), and the object (orange). LAMA masks the object and predicts a single token as the answer. In BEAR, we create separate textual statements for a set of potential answers and select the statement with the highest (pseudo) log-likelihood as assigned by the LM. This method allows us to include multi-token answers and evaluate causal and masked LMs.

erner et al., 2020; Cao et al., 2021; Kalo and Fichtel, 2022).

The LAMA probe (Petroni et al., 2019) is the seminal work in studying commonsense and relational knowledge in LMs and is widely used for inexpensive evaluation and model comparison (see Youssef et al. (2023) and Cao et al. (2023) for an overview). The idea is to use relational knowledge from an existing knowledge base (KB) and create cloze-style statements for an LM to fill in.

For instance, the entities "France" and "Paris" may be connected through the HAS-CAPITAL relation in a given KB, indicating that Paris is the capital of France. From this, LAMA constructs the sentence "The capital of France is [MASK]" and evaluates whether an LM predicts the correct token to complete this factual sentence. LAMA, therefore, effectively reuses the masked language modeling objective of the bidirectional family of LMs (Devlin et al., 2019) to probe for knowledge. This example is shown in Figure 1a.

Limitations of LAMA. While LAMA offers a straightforward approach to probing, it also has significant limitations.

First, LAMA requires the correct answer to be part of the evaluated LM's subtoken vocabulary, restricting the relational knowledge that can be tested to single-subtoken answers (such as "Paris" in Figure 1a). Thus, it cannot test for relational facts with long or rare answers (as shown in Figure 1b).

Second, and most importantly, LAMA relies on the masked language modeling objective. This makes LAMA inapplicable for LMs trained with other objectives. Therefore, it excludes causal LMs such as the GPT-family of models (Radford et al., 2019). To the best of our knowledge, no factual knowledge probe currently exists that applies to both masked and causal LMs.

Third, various prior works have noted limitations of the relational data used in the LAMA probe, such as (1) a heavily skewed answer space, favoring some answers over all others (Jiang et al., 2020b; Zhong et al., 2021; Cao et al., 2021), (2) overly revealing entity names (Poerner et al., 2020), (3) and issues involving knowledge with multiple correct answers, causing correct answers to be counted as errors (Kalo and Fichtel, 2022).

These limitations of both the probing approach and the data impair LAMA's ability to measure and compare the relational knowledge of different LMs accurately.

Contributions. To address these issues, we propose BEAR, a unified knowledge probe for both causal and masked LMs. Instead of casting the evaluation as a token prediction problem over the entire vocabulary of an LM, we present a set of answer options for each relation instance, create a textual statement for each option, and leverage the inherent ability of each LM to assign a log-likelihood score to statements, thereby ranking these options. See Figure 1b for an illustration.

We argue that this approach has numerous benefits in that it (1) allows us to evaluate both masked and causal LMs, (2) imposes no restrictions on the answer space, (3) allows us to design a new evaluation dataset that addresses a range of issues such as answer skews and multiple correct answers noted in prior work. In more detail, our contributions are:

1. We present an analysis of the weaknesses of the LAMA probe and follow-up works to derive desiderata for the BEAR probe (see Section 2).

2. We propose to query knowledge as a multiple-choice selection problem in which an LM estimates the (pseudo) log-likelihood of a given answer template with each choice filled in (see Section 3).

3. We construct a novel evaluation dataset that reflects the desiderata identified in our analysis (see Section 4).

4. We use BEAR to evaluate a range of common masked and causal LMs (see Section 5).

To enable the community to employ the proposed probing method and dataset, we publicly release[1] the evaluation framework *lm-pub-quiz* (based on the *minicons*, Misra, 2022) as well as the dataset *BEAR*[2].

2 LAMA and Follow-Up Work

We discuss the technical details of the LAMA probe first, followed by an analysis of its weaknesses.

LAMA evaluation data. The LAMA benchmark was originally composed of four separate datasets named after their respective sources: SQuAD (Rajpurkar et al., 2016), GoogleRE[3], ConceptNET (Speer and Havasi, 2012), and T-REx (Elsahar et al., 2018). However, subsequent research mainly focused exclusively on the T-REx subset. Its knowledge base comprises 41 *relations* derived from Wikidata. Each relation contains at most 1,000 *relation instances* in the form of subject-relation-object triples $\langle s, r, o \rangle$ where s is the subject (e.g., "France"), r the relation (e.g., HAS-CAPITAL), and o an object (e.g. "Paris").

There are three types of relations in LAMA: 1-1 (one-to-one, e.g., HAS-CAPITAL), N-1 (many-to-one, e.g., HAS-LANGUAGE), and N-M (many-to-many, e.g., SHARES-BORDER-WITH). Relations of the N-1 type allow multiple subjects to relate to one object, while the latter permits many subjects to be associated with numerous objects.

Relation identifiers. All relations are linked to a corresponding relation in Wikidata and thus have

---

unique IDs. For instance, the CAPITAL-OF relation in LAMA corresponds to Wikidata relation P1376 (see Table 1 for more examples). It facilitates comparison across different datasets since all follow-up works to LAMA, including BEAR, derive their relations from Wikidata.

Templates. Each relation in LAMA has a textual template with placeholders for subject and object. For CAPITAL-OF, the template is "[X] is the capital of [Y].", where [X] is a placeholder for the subject, while [Y] is the placeholder for the object. At test time, the subject of a given relation is filled in the template, while the object is replaced by a [MASK]-token. This procedure results in a masked sentence (e.g., "Paris is the capital of [MASK].") for which the LM is tasked to predict the masked token.

2.1 Issue 1: Single Subtoken Answers

As noted by Petroni et al. (2019), LAMA is restricted to single-subtoken answers for factual knowledge queries. This limitation causes issues as LMs split most words into multiple subtokens, and most LMs differ in how they perform the splits. To illustrate, consider how the country name "Togo" is tokenized by different versions of BERT: the bert-base-cased model splits the word into two subtokens ([To, ##go]), whereas the bert-base-uncased variant preserves it as a single subtoken ([togo]).

An analysis of 194 UN member country names is an excellent example of how such a restrictive condition affects the size of a hypothetical dataset. When restricting answer space to single tokens, 32% and 27% of available country names would have to be discarded for cased and uncased versions of BERT, respectively. Worse, the RoBERTa models (Liu et al., 2019) that uses a BPE-based tokenizer would split 88% of all country names. Refer to Table 1 for a list of how many LAMA instances need to be discarded when evaluating xlm-roberta-base (Conneau et al., 2020) and roberta-base models.

Comparison of different LMs. Because the tokenizer bundled with each model differs, comparing various LMs becomes only possible if the models tokenize the answers in the same way. To address this, practitioners are currently reverting to using the intersection of single-token vocabularies derived from all LMs being compared. However, in practice, this further limits the scope of relational knowledge that can be included in the evaluation.

ID	Relation	xlm-roberta-base	roberta-base
P30	ON-CONTINENT	74.46 %	80.21%
P31	INSTANCE-OF	28.85%	67.35%
P36	HAS-CAPITAL	45.80%	89.76%
P37	HAS-LANGUAGE	30.85%	45.13%
...	...	...	...
P1303	INSTRUMENT	58.69%	100.00%
P1376	CAPITAL-OF	32.05%	81.62%
	Mean	31.73%	62.86%

Table 1: Ratio of discarded instances due to multi-token answers in xlm-roberta-base and roberta-base.

Prior work. Various prior works address the issue of predicting multi-subtoken words for single [MASK] tokens (Ghazvininejad et al., 2019; Jiang et al., 2020a; Kalinsky et al., 2023; Shen et al., 2020; Robinson et al., 2023). Jiang et al. (2020a) provided a selection of algorithms to tackle predicting multi-token entities. However, they require a specification of further parameters, such as the number of subtokens to generate. Kalinsky et al. (2023) proposed generation approaches that either require additional training or the use of an external network, making them inapplicable to the purpose of evaluating knowledge contained in pre-trained weights through a zero-shot approach.

2.2 Issue 2: Multiple Correct Answers

LAMA expects exactly one correct answer to each knowledge query and rates other factually correct answers as errors. To illustrate this, consider the query "Germany shares a border with [MASK]", to which LAMA expects the answer "Poland". All other correct answers, such as "Denmark" are marked as incorrect. This issue affects all N-M
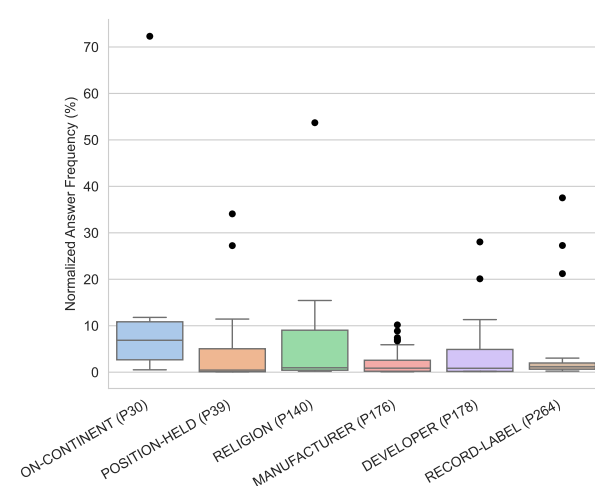

Figure 2: The normalized answer frequency of selected relations in the LAMA probe. The outliers are marked with dots. In some relations, a majority class accounts for more than 50% of all instances.
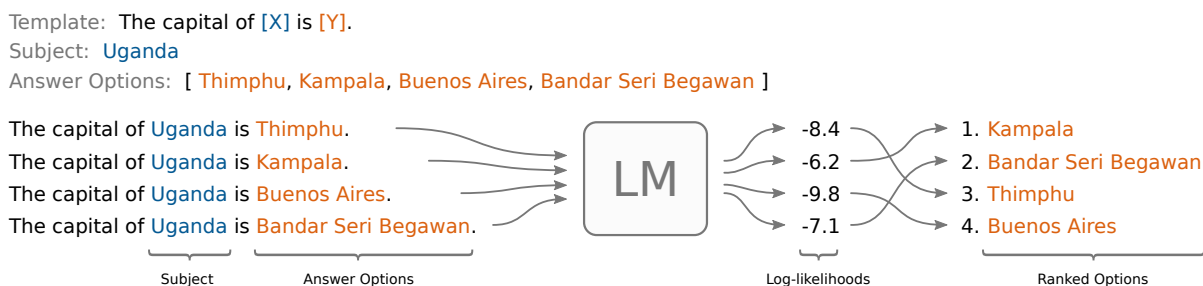
Template: The capital of [X] is [Y].
Subject: Uganda
Answer Options: [Thimphu, Kampala, Buenos Aires, Bandar Seri Begawan]

The capital of Uganda is Thimphu.		-8.4	1. Kampala
The capital of Uganda is Kampala.	LM	-6.2	2. Bandar Seri Begawan
The capital of Uganda is Buenos Aires.		-9.8	3. Thimphu
The capital of Uganda is Bandar Seri Begawan.		-7.1	4. Buenos Aires
Subject Answer Options		Log-likelihoods	Ranked Options

Figure 3: For each answer option, a sentence is passed to the LM (here using the template: "The capital of [X] is [Y]." and the subject "Uganda"). The log-likelihood scores assigned by the LM are then used to rank the answer options.

relations in LAMA.

Prior work. KAMEL (Kalo and Fichtel, 2022) address this by allowing the LM to generate an arbitrary number of answers using a template instructed via few-shot prompting, experimenting with ranges of 1-10 answers per instance. Subsequently, they evaluate the predictions using standard measures of precision and recall. However, their approach relies on the text generation ability of causal LMs and thus cannot be applied to masked LMs.

2.3 Issue 3: Imbalanced Answer Distribution

The relations in T-REx have a highly unbalanced answer distribution (except the 1-1 relations), and in certain relationships, over half of the instances belong to the predominant class (see Figure 2). Zhong et al. (2021) noted that a model that always chooses the majority class outperforms some state-of-the-art LMs on selected relations.

To illustrate, consider the T-REx's ON-CONTINENT relation, which connects a location to the continent in which it is situated. Counter-intuitively, the majority class in this relation is "Antarctica", accounting for 72% of all instances.

Prior work. To account for this imbalance, Cao et al. (2021) created a balanced version of the LAMA probe called WikiUNI. It contains the same relations as T-REx but has a uniform answer distribution and was constructed to have the same number of subjects for every object. However, their dataset samples a highly skewed number of instances per relation, with 7 relations (out of 41) accounting for over 50% of all instances.

2.4 Issue 4: Rare Wikidata Entries

The above-mentioned example of "Antarctica" being an answer to over 72% of all instances in the ON-CONTINENT relation also points to another problem: An artifact of randomly sampling Wiki-

data for relation instances is that rare Wikidata entries are overrepresented. For example, ON-CONTINENT has a large number of small islands as subjects (e.g., "Umber Island" and "Brooklyn Island", both close to the Antarctic continent), many of which are unlikely to occur in a corpus outside of an encyclopedia like Wikipedia. We believe that the LAMA dataset unfairly favors LMs trained using Wikipedia. However, to the best of our knowledge, no prior work has addressed this issue.

2.5 Issue 5: Evaluation of Causal LMs

LAMA relies on the capability of masked language models to fill in masked tokens, making it unsuitable for causal LMs.

Prior work. To address this, Kalo and Fichtel (2022) proposed the KAMEL probe. Factual knowledge is probed by virtue of question statements for which the response is auto-regressively generated using the causal LM. To guide the generation approach, they prepend k few-shot examples into the prompt that present how the correctly formatted answer should look. However, since this approach relies on the language modeling objective of causal LMs, KAMEL does not apply to masked LMs.

Further Related Work

Further related work has attempted to mitigate the sensitivity of the evaluation on the timing of pre-training (and hence the corpus used) (Dhingra et al., 2022; Mallen et al., 2023; Onoe et al., 2022). Nonetheless, as these studies do not directly address the concerns outlined in the previous subsections, we will not discuss them in detail. Interested readers are referred to the cited papers for more information.

3 BEAR Probe

We base our evaluation on using an LM's inherent ability to estimate the log-likelihood of a given sentence. Our main idea is to restrict the space of possible answers for each relation instance and create a set of options that are then ranked by their log-likelihood values.

3.1 Ranking Options using Log-Likelihood

Our approach requires a dataset of $\langle s, r, o \rangle$ relation instances, where for each relation r there exists (at least) one template t and a set of answer options a_i with $i \in \{1, ..., k\}$ that includes the correct answer.

Creating options to rank. For each relation instance, we create k natural language statements using the template, instance's subject s, and each of the possible relation's objects a_i as parts of a textual statement.

Figure 3 illustrates this process for the example relation instance $\langle$"Uganda", HAS-CAPITAL, "Kampala"$\rangle$ and the template "The capital of [X] is [Y]". This example's set of potential answers is ["Kampala", "Thimphu", "Buenos Aires", "Bandar Seri Begawan"]. For each potential answer, we create a separate statement.

Predicting log-likelihood. For each generated statement, we predict the log-likelihood score $\log \hat{p}(a|t)$. As the template is the same for each of the answer options (i.e., $\hat{p}(t)$ is constant) and we are only interested in ranking them, it is sufficient to compute the log-likelihood for the entire sentence:

$$\log \hat{p}(a_i|t) = \log \hat{p}(a_i, t) - \log \hat{p}(t)$$

Since causal LMs are trained to predict a log-likelihood of each token given the previous context, the log-likelihood of a sentence is simply the sum of the log-likelihoods of each token.

Log-likelihood in masked LMs. A sentence-level log-likelihood is not clearly defined for an LM trained using the masked language modeling objective. However, Salazar et al. (2020) and Kauf and Ivanova (2023) offer two variants of how to retrieve a pseudo log-likelihood score for a given text. Both approaches use multiple forward passes. Salazar et al. (2020) simply mask each token once while keeping the remaining context unmasked. Kauf and Ivanova (2023) improve on this by additionally masking all tokens right to the current

token belonging to the same word. This approach fixes the issue of assigning disproportionate likelihoods to multi-token words. We use the latter in our approach.

Ranking the results. Finally, the statements are ranked by their log-likelihood scores. This is illustrated in Figure 3 (right-hand side).

3.2 Evaluation Metric

To evaluate the amount of knowledge encoded in each model, we score whether the top-ranked statement is the correct answer for each relation instance. Previous work (such as Petroni et al., 2019) additionally considered answers with higher ranks. The mean precision $P@k$ (for a given rank k) is commonly reported.

However, given that we evaluate on a constrained answers space, we believe the first rank to be sufficient. In the case of $k = 1$, this is identical to the accuracy metric. We report the average over all relation instances in our evaluation data as the BEAR score.

4 BEAR Dataset

Our proposed probing approach requires a dataset with a restricted answer space. Following the analysis in Section 2 we additionally desire (1) the answer space to be balanced, (2) a single correct answer per relation instance, (3) a balanced number of instances per relation, and (4) a focus on knowledge that could reasonably be found in corpora other than Wikipedia.

4.1 Selecting Relations

We use the 234 relations of KAMEL as a starting point and manually remove two-thirds of these. This curation process was conducted independently by two researchers (authors of this paper), and disagreeing judgments were discussed in detail to reach a final decision. The most common reasons for excluding a relation were:

- A relation (after filtering) had too few objects with a desired number of instances (i.e., given the relation's statistics, it was impossible to build a balanced answer space within our constraints).

- We expect the relational information to undergo significant changes (hence, it is not time-invariant). For example, we assume the RESI-DENCE relation, linking an individual to their

2397

place of residence, to be highly susceptible to change over time. Including such relations without accounting for their temporal context would unfairly benefit LMs trained on datasets from the same time period as the evaluation data.

- A relation with many instances that have incomplete object listings in Wikidata since this may cause correct answers to be counted as errors. For example, the MADE-FROM-MATERIAL relation, which connects an object and the material it is made of, often contains only a few primary components as objects.

- Overly diverse subject or object space, making the semantics of the relation overly broad and impairing our ability to design meaningful templates. For instance, the relation COUN-TRY connects various entity types, such as events, ships, roles, websites, TLDs, codes of standards, and many more categories, to a country.

As a result of this process, we selected 78 relations for inclusion in the larger variant of BEAR.

4.2 Selecting Relation Instances

For the selected relations, we retrieve relation instances from Wikidata[4] and employ a number of filtering steps to ensure that only instances meeting our desiderata remain.

Filtering subjects and objects. We first filter down the space of eligible subjects and objects. We remove all Wikidata entities (i.e., subjects and objects) that do not have an English label. Following prior work (Poerner et al., 2020), we additionally remove all subjects with overly revealing entity names. For example, predicting the name of the company that produced the "Apple Watch" is straightforward since the correct answer (i.e., "Apple") is part of the subject (i.e., "Apple Watch"). The similarity is computed via the token overlap and fuzzy string matching (Bachmann, 2023).

Ensuring a coherent answer space. Even in our curated set of relations, some relation instances are connected to outlier object types. For instance, the HEAD-OF-GOVERNMENT relation, which typically connects a country to a specific named person (e.g.,

"Joe Biden"), would in some cases connect to a job title instead (e.g., "president").

To increase coherence and ensure that our templates are meaningful, we utilized GPT4 (OpenAI, 2023) to flag answers which stand out (see Figure 11 in Appendix F for the template that was used) and decided on a case-by-case basis whether to accept these changes. This process also helped us check the relations for potential issues.

Sampling a balanced dataset. For this initial set of entities, we sampled the remaining relation instances such that (1) each relation has a uniform distribution of objects in the answer space, (2) each relation has approximately the same number of instances overall, and (3) there is no overlap between various entity names within a relation. During sampling, we gave preference to Wikidata entities with Wikipedia pages in multiple languages to focus on well-known entities that might reasonably be found in corpora outside of Wikipedia.

This process yields a total of 40,916 instances for our 78 relations.

4.3 Further Refinement

The large variant of the dataset consists of a broad range of subjects and captures a considerable amount of relational knowledge found in Wikidata. However, considering the extensive scope of the probe and the evaluation scheme employed, the time required for a complete assessment is considerable. Hence, we further developed a subset of the full probe designed to be more time-efficient.

We imposed a constraint on the answer space, setting an upper limit of 25 answers for 1:N relations. This upper limit was set to keep the task challenging while making the evaluation more efficient. To further refine this subset, we applied a filtering criterion based on entity popularity, excluding entities (objects and subjects) linking to Wikipedia pages with fewer than 10,000 page views (between 2016 and 2023). This step ensures the relevance and recognition of the included entities. Subsequently, through an iterative process, we identified optimal configurations that achieve a balanced answer space with a target of approximately 150 instances. Configurations deemed infeasible were systematically removed (in total, 18 relations, leaving 46 relations). The remaining possible answers were then sorted based on their subjects' median page view count and checked in a manual review. During this review, we examined the answers and

---

their respective subject samples, leading to the exclusion of problematic subject instances. We then included subjects with the highest page view counts.

In the case of 1-1 relations, the process was similar but using a limit of 60 answers/instances per relation. Additionally, we selected a random sample of instances rather than ranking them by popularity for these relations.

In the following sections of the paper, we will use "BEAR" to refer to this subset and "BEAR$_{big}$" to denote the full probe.

4.4 Templates

We create three templates for each relation to better safeguard against template-specific biases. We source the initial templates from the existing LAMA dataset, utilize GPT4 to create additional ones (the used prompt can be found in Figure 12 in Appendix F), and manually select those that best match our subjects and answer spaces. Finally, we query GPT4 with each of the templates applied to 5 random subject-object pairs from each relation to check for linguistic correctness (the used prompt can be found in Figure 13 in Appendix F).

4.5 Resulting Dataset

The final dataset consists of 60 relations and 7,731 items. Most of these relations are 1:N, each featuring a restricted answer space ranging from 5 to 25 possible answers, with an average of roughly 23 answers. The answer space is also balanced so that each answer appears the same number of times across all instances. Each answer has between 6 and 30 instances, with an average of approximately 6.5 instances per answer. The dataset also contains 14 1:1 relations that contain only one instance per answer.

For a detailed comparison of these statistics with LAMA and KAMEL, see Table 2.

5 Experiments

We present an experimental evaluation using the BEAR probe on a selection of LMs, compare our results to earlier probes, and discuss the results.

Compared LMs. We compare a total of 22 LMs, as listed in Table 3: This includes 6 masked LMs from the BERT, RoBERTa, XLM-RoBERTa families, each in their base and large variants. Additionally, we include 16 causal LMs from the GPT

Dataset	LAMA	KAMEL	BEAR$_{big}$	BEAR
# Instances	31,479	46,800	40,916	7,731
# Relations	41	234	78	60
Literals	no	yes	no	no
1:1 Relations	0		14	14
N:1 Relations	7		64	46
N:M Relations	34		0	0
N:M Instances	1,035	4,296	0	0
Avg. Instances per Relation	830.2	1,400[5]	596.9	149.8

Table 2: Descriptive dataset statistics: BEAR compared to LAMA (T-REx subset) and KAMEL (figures for KAMEL and LAMA from Kalo and Fichtel, 2022). Avg. Instances per Relation only includes relations with more than one instance per answer.

and OPT families, along with newer models such as Llama2, Gemini, and Mistral. We assess 5 different sizes for the OPT models to examine the relationship between the BEAR score and increasing model sizes.

BEAR score. We compute the BEAR score for each of the three template options per relation individually and report the average across templates as well as the standard deviation.

5.1 Main Results

Table 3 lists the results for all LMs in consideration. We present the overall BEAR score and the scores for the subsets of 1:1 and N:1 relations only. We find that scores are generally low for all models, highlighting the challenging nature of our benchmark, as it queries factual information with strong detractors. In addition, we make a number of observations:

BEAR scores are higher for larger LMs. In line with our expectations, we find that larger models consistently outperform their smaller counterparts. For a better illustration, we present a plot of accuracy against model size in Figure 4. This trend of steady accuracy improvement with increasing model size is evident across all tested model families. Interestingly, the smallest change is observed among BERT models, where the performance of `bert-base-cased` and `bert-large-cased` across all of the relations is roughly on par. We further observe that recent models generally achieve higher performance, with `Mistral-7B-v0.1` (2023) and `gemma-7b` (2024) surpassing the `Llama2-7b` (2023) model (when compared at identical parameter counts), which itself significantly outperforms the `opt-6.7b` model from 2022.

Model	Type	# params	BEAR	BEAR$_{1:1}$	BEAR$_{1:N}$
Llama-2-13b-hf	CLM	13b	66.9% ± 1.0%	66.5% ± 1.6%	67.0% ± 1.1%
Mistral-7B-v0.1	CLM	7.0b	65.4% ± 1.1%	64.5% ± 1.2%	65.5% ± 1.1%
gemma-7b	CLM	7.0b	63.7% ± 1.3%	63.5% ± 0.7%	63.8% ± 1.4%
Llama-2-7b-hf	CLM	7.0b	62.4% ± 1.3%	62.2% ± 1.1%	62.4% ± 1.3%
gemma-2b	CLM	2.0b	51.5% ± 1.0%	53.1% ± 1.3%	51.3% ± 1.0%
opt-30b	CLM	30b	47.9% ± 0.5%	45.8% ± 1.0%	48.2% ± 0.6%
opt-13b	CLM	13b	45.4% ± 0.8%	43.5% ± 2.1%	45.7% ± 0.6%
opt-6.7b	CLM	6.7b	43.8% ± 1.1%	42.5% ± 1.0%	43.9% ± 1.2%
opt-2.7b	CLM	2.7b	37.3% ± 0.9%	35.6% ± 0.7%	37.5% ± 1.0%
opt-1.3b	CLM	1.3b	31.5% ± 0.8%	31.3% ± 0.6%	31.5% ± 0.9%
gpt2-xl	CLM	1.6b	26.2% ± 0.7%	24.1% ± 1.6%	26.5% ± 0.6%
gpt2-large	CLM	812M	22.2% ± 0.6%	20.1% ± 1.8%	22.5% ± 0.5%
roberta-large	MLM	355M	21.5% ± 0.8%	22.0% ± 1.1%	21.5% ± 0.8%
bert-large-cased	MLM	335M	19.9% ± 0.5%	16.6% ± 1.0%	20.3% ± 0.5%
opt-350m	CLM	350M	19.6% ± 0.6%	18.6% ± 1.2%	19.7% ± 0.6%
gpt2-medium	CLM	355M	19.0% ± 0.8%	16.0% ± 2.6%	19.4% ± 0.6%
bert-base-cased	MLM	109M	18.4% ± 0.4%	15.0% ± 1.1%	18.8% ± 0.4%
roberta-base	MLM	125M	16.4% ± 0.7%	15.8% ± 1.8%	16.5% ± 0.8%
opt-125m	CLM	125M	16.4% ± 0.5%	14.0% ± 1.3%	16.7% ± 0.4%
xlm-roberta-large	MLM	561M	14.3% ± 0.3%	14.9% ± 1.7%	14.3% ± 0.5%
gpt2	CLM	137M	13.5% ± 0.8%	9.4% ± 2.1%	14.0% ± 0.7%
xlm-roberta-base	MLM	279M	11.4% ± 0.2%	11.4% ± 1.1%	11.4% ± 0.2%
Random Baseline	-	-	4.7%	1.7%	5.1%

Table 3: Models investigated in this work (Devlin et al., 2019; Jiang et al., 2023; Gemini-Team, 2023; Liu et al., 2019; Radford et al., 2019; Touvron et al., 2023; Zhang et al., 2022) sorted by their BEAR score (weighted average over all relations) and as the mean over all templates (with the standard error).
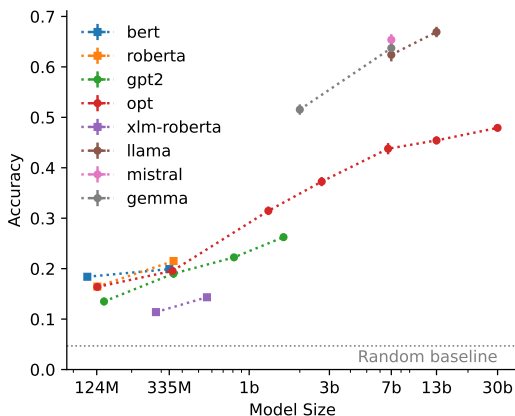

Figure 4: Probing scores of different models on BEAR. Model size is represented on a log scale.

Better BEAR scores for masked LMs. When comparing models by their parameter count, we note a slight advantage of masked over causal LMs. This may indicate that the masked language modeling objective, encouraging deep bidirectionality, is more effective in capturing factual knowledge.

Impact of multilingual training data. We note that the two XLM-RoBERTa models are among the lowest-scoring models in the benchmark. We hypothesize that this diminished performance of the XLM models may stem from its pre-training on multilingual corpora and a focus of BEAR on English-language entities.

Impact of templates. We further evaluate the im-

pact of template choice on the BEAR score. A full analysis of all relations is provided in Figure A in the Appendix.

In line with the observation of Elazar et al. (2021), we find that LMs are sensitive to how they are queried. For instance, in the CAPITAL-OF relation, the accuracy of bert-base-cased drops approximately by 80% when using "[Y] has its governmental seat in [X]" instead of "The capital of [X] is [Y].". While some states have their government seat in a different location than their official capital city, this fact alone cannot account for the significant disparity in observed performance. Interestingly, we don't observe any drop in performance between these templates in the case of other models (e.g., Llama2 models). We hypothesize that this might be due to BERT's primary training on Wikipedia, which has limited exposure to diverse writing styles.

5.2 Comparison to LAMA

To compare BEAR and LAMA, we consider their common subset of relations and utilize our proposed log-likelihood based evaluation technique. See Figure 5 for an illustration. We find BEAR to be a more challenging probe compared to T-REx, the dataset used in LAMA. We attribute this to several design choices, namely the balanced answer space and the absence of overly informative entity names, forcing LMs to rely solely on the knowl-

Figure 5: Comparative analysis of model performance on identical subsets of relations and templates in the T-REx (LAMA) and BEAR datasets using the log-likelihood based evaluation.
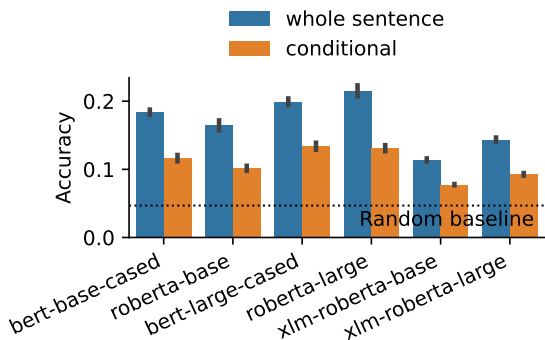

Figure 6: Aggregated accuracy (measured on BEAR) when using the sum over all tokens in the complete statement vs. answer-tokens only.

edge encoded within its parameters.

We also find that the performance disparity evaluated on BEAR's relations is less pronounced than across the same subset of T-REx's relations. Such difference may indicate that models pre-trained on Wikipedia (like BERT) have an advantage on LAMA over those not trained on Wikipedia (such as GPT2) due to a potential train/test data overlap. For a detailed performance comparison on a per-relation basis, refer to Figure 8 in the Appendix. For example, bert-base-cased achieves a very high performance on T-REx's MANUFACTURER relation (ID: P176) but a significantly lower score on the corresponding BEAR subset.

5.2.1 Ablations

Conditional scores. To compute the pseudo log-likelihood of a statement in masked language models, one forward pass per token is required. However, masking only the tokens that are part of the an-

swer would significantly reduce the required computation. One may expect that the representation of the answer tokens may be sufficient to predict the likelihood of each answer. Still, our experiments (see Figure 6) indicate there is a significant[6] drop in performance when using the conditional score (i.e., the score of the answer tokens conditioned on the context instead of the pseudo log-likelihood of the complete sentence). Coincidentally, for entities represented by a single token, the conditional score matches the score predicted for the [MASK] token, similar to the approach first used in the LAMA probe.

Sum vs. mean of the log-likelihood. We investigate how performance varies by scoring the sentences using both sum and mean reduction methods. We observe that scoring the sentence by computing the mean over the token log-likelihoods tends to yield inferior performance for the probe. Figure 7 (in Appendix A) illustrates the results of this ablation study.

To understand why this might be the case, consider how the word 'souvenir' can be broken down into subword tokens: 'so', '##uven', and '##ir'.[7] The first token 'so' may have a relatively small log-likelihood, '##uven' a bit higher (since it's conditioned on the first token), and finally '##ir' a log-likelihood of almost 0 since there are few other ways to continue the statement. In contrast, the single-token word 'gift' may have a medium log-likelihood, which may still be lower than the average of the three-token word 'souvenir'. This example illustrates that the mean can inflate the predicted probabilities of longer answers or sentences.

We suggest summating the tokens' log-likelihoods for both masked and casual language models in future experiments.

6 Conclusion

We presented BEAR, a relational knowledge probe applicable to both causal and masked LMs. Since our proposed approach imposes no restrictions on the evaluation data, we created a large evaluation dataset that addresses issues of answer skews, domain and template bias, and the correctness of facts identified by ourselves and prior work. We publicly release BEAR for use by the research community.

---

[6]P-value of 1.1×10^{-11}; using a Student's t-test for paired samples

[7]The example was introduced by Kauf and Ivanova (2023) to make a different point but is also relevant here.

Limitations and Risks

The knowledge probe we present in this paper follows the approach of earlier probes and, as such, tests only for factual, relational knowledge. This kind of knowledge includes classic relationship types such as the place of birth of persons, their time of birth, the genre of works of art, etc. However, one might be interested in testing a model for other types of more general commonsense knowledge, such as physical reasoning or general properties of concepts. Our probe does not test for such kinds of knowledge.

Furthermore, even though we devised heuristics to ensure that entities in BEAR are common enough to appear on Wikipedia pages of many different languages and record a certain number of page views, there remains a likely bias towards entities overrepresented on Wikipedia, giving an advantage to LMs trained on Wikipedia rather than more general corpora.

We see few risks in the BEAR probe itself but caution that knowledge probing is often used to assist in the research and development of LMs. As such, BEAR may contribute to developing LMs that malevolent actors might misuse.

References

Max Bachmann. 2023. RapidFuzz. Original-date: 2020-02-29T14:41:44Z.

Boxi Cao, Hongyu Lin, Xianpei Han, and Le Sun. 2023. The Life Cycle of Knowledge in Big Language Models: A Survey. ArXiv:2303.07616 [cs].

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. 2022. Time-Aware Language Models as Temporal Knowledge Bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

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.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. 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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Gemini-Team. 2023. Gemini: A Family of Highly Capable Multimodal Models.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-Predict: Parallel Decoding of Conditional Masked Language Models. 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 6112–6121, Hong Kong, China. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud,

Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models. *arXiv:2010.06189 [cs]*. ArXiv: 2010.06189.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Oren Kalinsky, Guy Kushilevitz, Alexander Libov, and Yoav Goldberg. 2023. Simple and Effective Multi-Token Completion from Masked Language Models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2356–2369, Dubrovnik, Croatia. Association for Computational Linguistics.

Jan-Christoph Kalo and Leandra Fichtel. 2022. KAMEL : Knowledge Analysis with Multitoken Entities in Language Models. In *Automated Knowledge Base Construction*.

Carina Kauf and Anna Ivanova. 2023. A Better Way to Do Masked Language Model Scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. Association for Computational Linguistics.

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:1907.11692 [cs].

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Yasumasa Onoe, Michael Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity Cloze By Date: What LMs Know About Unseen Entities. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 693–702, Seattle, United States. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

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.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging Large Language Models for Multiple Choice Question Answering.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5186–5198, Online. Association for Computational Linguistics.

Robyn Speer and Catherine Havasi. 2012. Representing General Relational Knowledge in ConceptNet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3679–3686, Istanbul, Turkey. European Language Resources Association (ELRA).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura,

Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Wikidata contributors. 2022. Dump of Wikidata of January 3rd 2022.

Paul Youssef, Osman Alperen Koraş, Meijie Li, Jörg Schlötterer, and Christin Seifert. 2023. Give Me the Facts! A Survey on Factual Knowledge Probing in Pre-trained Language Models. ArXiv:2310.16570 [cs].

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. 2022. OPT: Open Pre-trained Transformer Language Models. ArXiv:2205.01068 [cs].

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. 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. Association for Computational Linguistics.
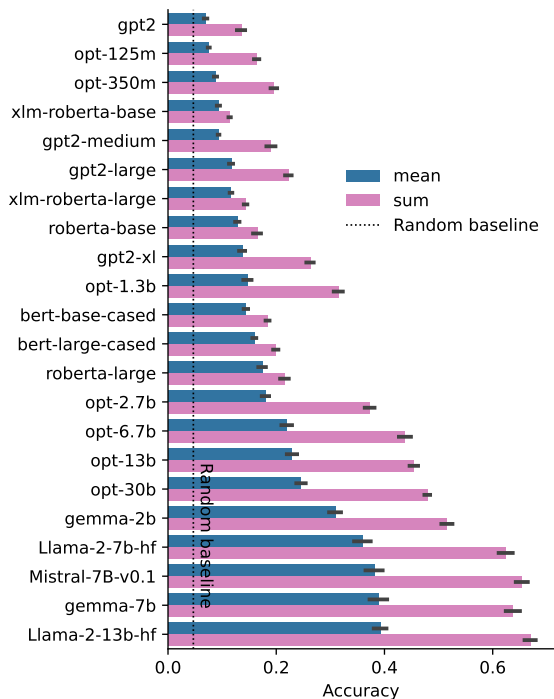
A Further Results

Figure 7: Aggregated accuracy of different retrieval variants on BEAR. The error bars indicate the standard error over three evaluations using the different templates.

A.1 Results on BEAR_big

In addition to the accuracy on the further refined BEAR dataset, we report the scores on BEAR_big (see Table 4). The ranking is very similar and differs only in a few positions.

A.2 Ablation: Pseudo log-likelihood metric

While in preliminary experiments on LAMA, we observed a higher benefit from using the `within_word_l2r` variant, it has only a slightly higher mean score than `original` (see Figure 10). This difference is not significant when using the `sum` variant (p-value of 0.52 on a Student's t-test for paired samples). However, the difference is large when using the `mean` variant (and significant with p-value of 0.025)

A.3 Impact of Pre-Training Data on the BEAR Score

In order to verify the actual ability of the BEAR probe to measure knowledge contained in models' pre-trained weights, we set up an ablation study. We hypothesize that models trained solely on domain-specific datasets, without exposure to the general knowledge tested by BEAR, will show

significantly reduced performance compared to those trained on more general datasets, given that the models have similar architectures and sizes. A family of BioGPT models (Luo et al., 2022) was based on the GPT-2 architecture with the sole differences arising from the pre-training data: specifically, they were trained on PubMed abstracts and titles rather than on data from web crawls as it was the case for the GPT models. The evaluation results confirm our hypothesis, with the BioGPT model achieving an average score of 10.84% and its large variant reaching 13.6% on BEAR, both trailing behind the `gpt2-medium` (19.0%) and `gpt2-xl` (26.2%) models.

B Noise Levels

Crowdsourced knowledge bases like Wikidata often contain noisy and inaccurate data. To reduce this issue, we developed heuristics to select better-known entities that are more likely to be verified and corrected. Nonetheless, the potential for noise leakage exists. To evaluate the extent of such noise, we performed a noise levels analysis, in which we manually reviewed 100 randomly chosen examples from various relations for both BEAR_big and LAMA (T-REx) probes. We cross-validated the accuracy of the information from Wikidata with alternative sources, confirming 96% and 97% of examples for BEAR_big and T-REx, respectively[8].

Additionally, we evaluated whether the object of the relation truly represents the only correct answer as our benchmark (expecting a single correct answer) would mark those answers as incorrect. About 11% of BEAR's answers included multiple correct responses within its answer range, significantly lower than T-REx's 35%[9], which uses BERT's vocabulary as its answer space.

Thus, while BEAR_big and LAMA exhibit similar noise levels, BEAR_big demonstrates higher reliability by effectively reducing the incidence of multiple correct answers. Moreover, the refined subset of BEAR, due to additional filtering steps, is expected to decrease these values even further.

C GPU Compute Time

Due to the diversity of GPU hardware utilized in our experiments, we have chosen not to present cumulative GPU hours. Instead, to provide insights into the computational requirements of our

---

[8]A two-sided z-test yields a p-value of 0.6698
[9]A two-sided z-test yields a p-value of 0.0001

Model	Type	# params	BEAR	BEAR$_{1:1}$	BEAR$_{1:N}$
Llama-2-13b-hf	CLM	13b	42.0% ± 0.6%	54.3% ± 1.3%	41.2% ± 0.6%
Mistral-7B-v0.1	CLM	7.0b	41.0% ± 0.6%	52.6% ± 1.2%	40.2% ± 0.6%
gemma-7b	CLM	7.0b	39.5% ± 0.8%	52.0% ± 1.1%	38.6% ± 0.8%
Llama-2-7b-hf	CLM	7.0b	37.5% ± 0.8%	50.3% ± 1.1%	36.6% ± 0.7%
gemma-2b	CLM	2.0b	29.1% ± 0.6%	41.4% ± 1.1%	28.2% ± 0.6%
opt-30b	CLM	30b	25.6% ± 0.3%	35.7% ± 0.9%	24.9% ± 0.3%
opt-13b	CLM	13b	24.2% ± 0.5%	33.6% ± 1.6%	23.5% ± 0.5%
opt-6.7b	CLM	6.7b	23.2% ± 0.7%	33.1% ± 0.5%	22.5% ± 0.7%
opt-2.7b	CLM	2.7b	19.3% ± 0.4%	27.6% ± 0.8%	18.7% ± 0.3%
opt-1.3b	CLM	1.3b	16.0% ± 0.4%	23.3% ± 1.0%	15.5% ± 0.4%
gpt2-xl	CLM	1.6b	12.9% ± 0.2%	17.8% ± 1.2%	12.6% ± 0.2%
roberta-large	MLM	355M	11.1% ± 0.4%	17.1% ± 0.8%	10.7% ± 0.4%
gpt2-large	CLM	812M	10.7% ± 0.2%	14.0% ± 1.5%	10.5% ± 0.2%
bert-large-cased	MLM	335M	10.1% ± 0.3%	11.8% ± 0.7%	10.0% ± 0.3%
bert-base-cased	MLM	109M	9.6% ± 0.3%	11.5% ± 1.2%	9.4% ± 0.3%
opt-350m	CLM	350M	9.5% ± 0.2%	13.4% ± 0.8%	9.2% ± 0.2%
gpt2-medium	CLM	355M	9.1% ± 0.3%	11.3% ± 1.9%	8.9% ± 0.2%
roberta-base	MLM	125M	8.4% ± 0.3%	11.8% ± 1.8%	8.1% ± 0.4%
opt-125m	CLM	125M	8.0% ± 0.2%	9.5% ± 0.8%	7.9% ± 0.2%
xlm-roberta-large	MLM	561M	7.4% ± 0.2%	11.2% ± 1.6%	7.1% ± 0.1%
gpt2	CLM	137M	6.4% ± 0.3%	5.8% ± 1.6%	6.5% ± 0.2%
xlm-roberta-base	MLM	279M	5.8% ± 0.1%	8.7% ± 0.9%	5.6% ± 0.0%
Random Baseline	-	-	2.5%	0.5%	2.7%

Table 4: Models investigated in this work evaluated on BEAR$_{\text{big}}$ (weighted average over all relations) and as the mean over all templates (with the standard error).

research, we offer a selection of representative examples highlighting individual experiments. The duration for the BEAR probe evaluations notably differed across models. For instance, evaluating the `opt-6.7b` model on the larger variant of the probe using the Nvidia A100 (80GB) machine required approximately 14 hours, whereas `gpt2` averaged about 1 hour. Additionally, causal language models were evaluated faster due to their method for calculating sentence log-likelihoods. For instance, evaluating the `bert-base-cased` model, roughly the same size as `gpt2`, on the entire BEAR probe took four times as long.

Moreover, as stated in the main text, assessing the large variant of BEAR requires a significant amount of time. For example, `bert-base-cased` requires almost 4.5 hours when evaluating the BEAR$_{\text{big}}$ probe with Nvidia RTX 3090. Yet, this duration drops to around 16 minutes when using the smaller BEAR probe on the same hardware.

D Predictions' Confidence

Log-likelihoods assigned to textual statements by language models over a closed set of answers can be converted into values that can be interpreted as *probabilities*. Here, the idea involves transforming these negative values with the softmax function to obtain a normalized set of values that together resemble a probability distribution. These scores reflect the model's confidence in each possible answer within the set. Furthermore, by comparing the calculated probability distribution with a uniform distribution, we can derive a *uncertainty score*, which reflects the model's confusion for a given instance. This score is based on the Kullback-Leibler Divergence between the predicted probabilities and the uniform distribution. Specifically, we derive divergence between a predicted probability distribution P and a uniform distribution U over N possible answers (outcomes). We then normalize this score to make it scale-free and bounded between 0 and 1 (entropy is maximal for the uniform distribution). Finally, we take a complement of the obtained value (by subtracting it from 1) effectively reversing its interpretation, shifting the focus from divergence to similarity to the uniform distribution. The mathematical formulation of this metric can be found in Equations 1, 2, 3.

$$
\begin{aligned}
KL(P\|U) =&= \sum_{i=0}^{N} p_i \log \frac{p_i}{1/N} \\
&= \sum_{i=0}^{N} p_i \log p_i + \sum_{i=0}^{N} p_i \log N \\
&= \log N - H(P) \\
&= H(U) - H(P) \\
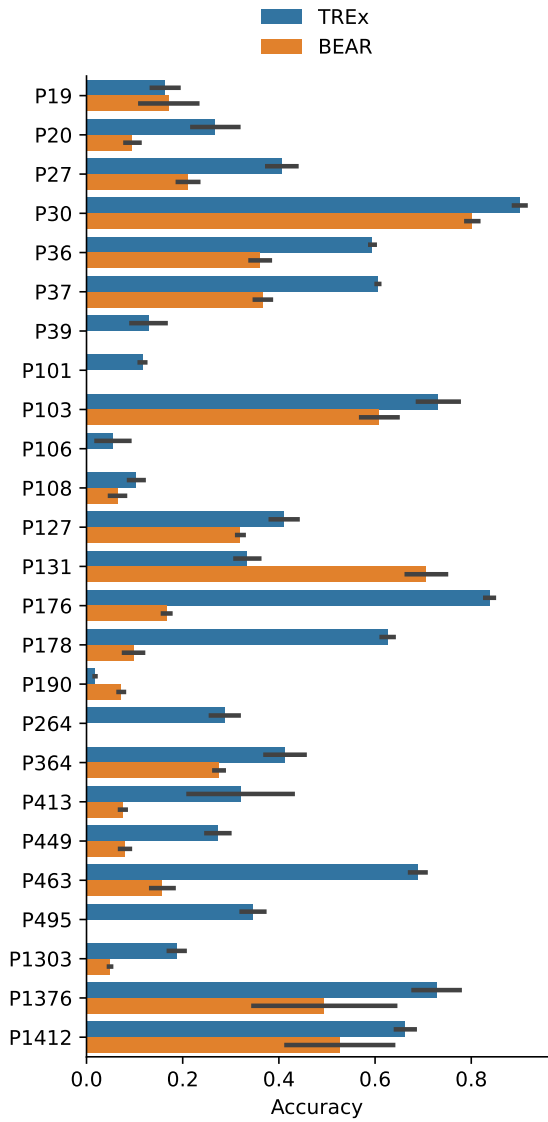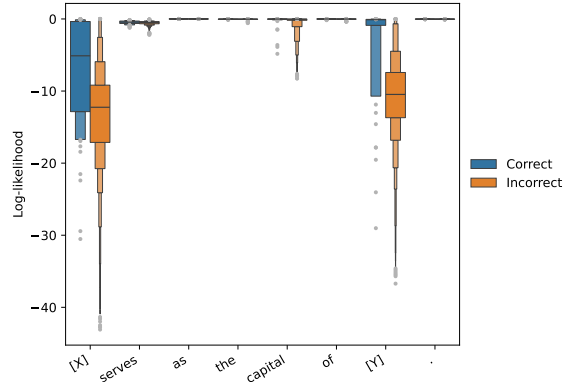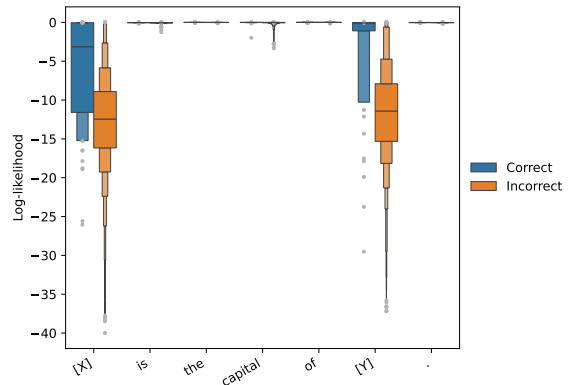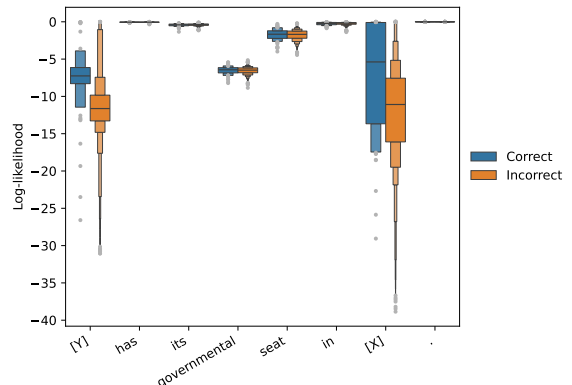&\in [0, H(U)]
\end{aligned}
\tag{1}
$$

Figure 8: Performance of `bert-base-cased` on a per-relation basis for both BEAR and T-REx probes. The results were obtained by summing pseudo log-likelihood scores (`within_word_l2r`)

(a) First Template: "[X] is the capital of [Y]."; Accuracy of 63%

(b) Second Template: "[Y] has its governmental seat in [X]."; Accuracy of 20%

(c) Third Template: "[X] serves as the capital of [Y]."; Accuracy of 65%

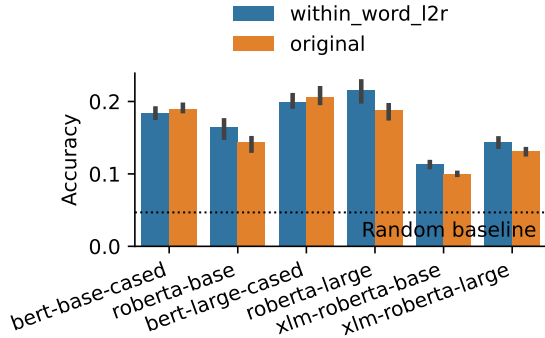Figure 9: `bert-base-cased` on P1376 (BEAR)

Figure 10: Aggregated accuracy of different retrieval variants on BEAR. The error bars indicate the standard error over three evaluations using the different templates.

$$\text{Normalized KL} = 1 - \frac{H(P)}{H(U)} \in [0, 1] \qquad (2)$$

$$\begin{aligned}\text{Uncertainty} &= 1 - \text{Normalized KL} \\ &= \frac{H(P)}{H(U)} \in [0, 1].\end{aligned} \qquad (3)$$

The interpretation of such a score is as follows: we observe the largest uncertainty when the predicted distribution approximates a uniform distribution, and the lowest when the entropy of the predicted distribution is null. The latter situation occurs with a point mass distribution, which allocates all the probability mass to one outcome.

E Error Analysis

We further conducted an error analysis on the mispredictions generated by the `Llama-2-13b-hf` model. The investigation did not reveal the presence of any systematic categories or clusters of errors. Some errors appeared to be similar to educated guesses. For instance, `Llama-2-13b-hf` outputs the highest log-likelihood for the factually incorrect statement such as "Kazım Ayvaz passed away in Turkey". However, this is not an unreasonable assumption, given that *Kazım Ayvaz* was a Turkish Olympic medalist born in Turkey. Even though the correct answer is *Sweden*, this illustrates that some of the model's errors may stem from logical assumptions instead of arbitrary mistakes.

On the other hand, certain errors appear to be random, with the predicted answer having no apparent connection to the subject. For example,

`Llama-2-13b-hf` assigned the highest score to the incorrect answer of *Massif Central* (highlands in France) when queried about the location of the Welsh mountain *Elidir Fawr*.

Furthermore, on rare occasions, the accuracy of the model's predictions is compromised by issues such as noise leakage or the imperfect quality of data sourced from Wikidata (for the analysis of such noise, see Section B in the Appendix). For instance, a Wikidata entry mentioning *Ronaldo* refers to Ronaldo Luís Nazário de Lima, commonly known simply as Ronaldo, a well-known player for the Brazilian national football team. However, the model erroneously identifies *Ronaldo* as Cristiano Ronaldo, a renowned Portuguese footballer. Table 5 presents a selection of errors made by the `Llama-2-13b-hf` model and provides further information on prediction's confidence as described in Section D. Specifically, it gives the ranks of the correct and predicted statements as assessed by their log-likelihoods as well as their probability scores obtained by applying the softmax function over all log-likelihoods in the rankings.

F Prompts Used

All prompts were passed as 'system messages' to Chat-GPT4 API and are presented in Figures 11, 12, and 13.

```
You are a researcher assistant tasked to design an evaluation dataset to test relational
knowledge contained in language models. Specifically, you are given a label for a relation,
its description, and a list of possible answers. Your assignment is to identify words that
do not align with the majority category in a given list of answers given the relation label
and its description. Return your response as a Python tuple. The first element should be
a list containing the words that don't fit the majority category, and the second element
should be a string representing the category of the majority of answers. If all words fit
the category, return an empty list. Example format: (['Berlin', 'Warsaw'], 'countries').
```

Figure 11: Prompt used to flag words in the answer space of each relation. In addition to some relation metadata (label and description) the (intermediate) answer space was passed on the model.

```
As a research assistant, your task is to create an evaluation dataset to assess the
relational knowledge of language models. You are provided with a specific relation label,
its definition, and examples of subjects and objects related to it. Your objective is to
craft three semantically similar cloze sentence templates that embody this relation. Use
'[X]' as a placeholder for the subject and '[Y]' for the object (answer). Ensure that these
sentence templates are straightforward and devoid of superfluous elements. For instance,
given 'label': 'educated at', 'description': 'educational institution attended by subject',
'subjects': ['Einstein', 'Feynman'], 'objects': ['Princeton University', 'University of
Zurich'], your templates might be: ['[X] was educated at [Y].', '[X] studied at [Y].', '[X]
was a student at [Y].']. Present your response as a Python list.
```

Figure 12: Prompt used to generate new template variants. Alongside the relation metadata, including label and description, 6 subject-object pairs were provided as examples for each relation.

```
Evaluate the linguistic correctness of the following sentence. If it is correct, return
'Correct'. If incorrect, identify the error and suggest a revised sentence. For instance,
'I used to live in USA.' -> 'I used to live in the USA.', 'My name is John.' -> 'Correct'.
```

Figure 13: System message for GPT-4 API call used to identify potential problems with the templates for each relation. We assessed all three templates by populating them with 5 random subject-object pairs. Although the prompt was intended to detect linguistic issues in the templates, it also facilitated the discovery of further issues with the relation instances.

Relation	Index	Correct	Instance	Rank	Probability	Uncertainty
P20	65	False	Kazım Ayvaz passed away in Turkey.	1	0.6038	0.3963
		True	Kazım Ayvaz passed away in Sweden.	5	0.0132	
P69	65	False	Gary Trent Jr. was educated at the University of Washington.	1	0.7873	0.2573
		True	Gary Trent Jr. was educated at the Duke University.	2	0.0814	
P509	284	False	Camille Pissarro died from tuberculosis.	1	0.4592	0.5945
		True	Camille Pissarro died from sepsis.	2	0.1854	
P19	6	False	Didier Marouani was born in Belgium.	1	0.4284	0.563
		True	Didier Marouani was born in Monaco.	4	0.0839	
P1303	314	False	Rosamund Pike is a viola player.	1	0.3485	0.522
		True	Rosamund Pike is a cello player.	2	0.3046	
P206	198	False	Biscoe Islands is situated on the shores of the Indian Ocean.	1	0.7662	0.3375
		True	Biscoe Islands is situated on the shores of the Southern Ocean.	2	0.0588	
P178	171	False	iOS is developed by Google.	1	0.4219	0.3706
		True	iOS is developed by Apple Inc..	2	0.3389	
P19	9	False	Zahir Khan was born in Pakistan.	1	0.6817	0.2711
		True	Zahir Khan was born in Afghanistan.	2	0.2735	
P175	152	False	Love You Live was performed by David Bowie.	1	0.388	0.6016
		True	Love You Live was performed by The Rolling Stones.	2	0.2469	
P466	277	False	The occupant of El Sadar Stadium is Real Betis Balompié.	1	0.6635	0.3132
		True	The occupant of El Sadar Stadium is Club Atlético Osasuna.	5	0.0266	
P7937	365	False	The Swan of Tuonela is a form of waltz.	1	0.385	0.452
		True	The Swan of Tuonela is a form of symphonic poem.	2	0.3495	
P7959	371	False	Quin is located in the historic county of Cornwall.	1	0.2946	0.6588
		True	Quin is located in the historic county of County Clare.	4	0.0771	
P509	279	False	Cardinal Richelieu died from peritonitis.	1	0.3046	0.6528
		True	Cardinal Richelieu died from tuberculosis.	3	0.1668	
P4552	349	False	Elidir Fawr is part of the Massif Central.	1	0.22	0.699
		True	Elidir Fawr is part of the Snowdonia.	2	0.2023	
P466	275	False	Allianz Stadium is occupied by Stanford University.	1	0.2156	0.5922
		True	Allianz Stadium is occupied by Juventus F.C..	4	0.1084	
P364	231	False	Shake It All About was originally created in French.	1	0.2289	0.6675
		True	Shake It All About was originally created in Danish.	10	0.0152	
P1532	331	False	Ronaldo represents Portugal.	1	0.9522	0.0806
		True	Ronaldo represents Brazil.	2	0.0329	
P108	101	False	Mary Lou Williams works for Google.	1	0.2291	0.6821
		True	Mary Lou Williams works for Duke University.	6	0.07	
P177	162	False	The Alfred H. Smith Memorial Bridge crosses the Connecticut River.	1	0.2193	0.7419
		True	The Alfred H. Smith Memorial Bridge crosses the Hudson River.	3	0.1514	
P171	144	False	The parent taxon of Bifora is Asteraceae.	1	0.6527	0.3646
		True	The parent taxon of Bifora is Apiaceae.	2	0.1884	
P364	230	False	The original language of Padmaavat is Malayalam.	1	0.4069	0.5318
		True	The original language of Padmaavat is Hindi.	2	0.2162	
P427	257	False	The taxonomic type of Rhizocarpon is Claviceps purpurea.	1	0.6416	0.3663
		True	The taxonomic type of Rhizocarpon is Map lichen.	10	0.0071	
P291	214	False	Follow the Reaper got published in Japan.	1	0.1692	0.8415
		True	Follow the Reaper got published in Finland.	11	0.0206	
P69	69	False	William Thomas Blanford was educated at the University of Oxford.	1	0.6478	0.2856
		True	William Thomas Blanford was educated at the Imperial College London.	12	0.0003	
P449	262	False	The Son was originally broadcasted by CBS.	1	0.2422	0.736
		True	The Son was originally broadcasted by AMC.	3	0.1433	
P206	200	False	Lipari is situated on the shores of the Ionian Sea.	1	0.402	0.392
		True	Lipari is situated on the shores of the Tyrrhenian Sea.	2	0.3725	
P7937	363	False	Ombra mai fu is a form of opera.	1	0.4891	0.3061
		True	Ombra mai fu is a form of aria.	2	0.4493	

Table 5: A random selection of errors originating from `Llama 2 13B` model spanning a diverse sample of relations. Each instance is constructed with a subject (highlighted in blue), a template (in plain text) and an object (highlighted in orange). Probability values are obtained by applying the softmax function to the log-likelihood scores of all potential answers. Uncertainty is measured by the similarity to a uniform distribution (see Section D). The lower the uncertainty value, the higher the model's confidence in its prediction.
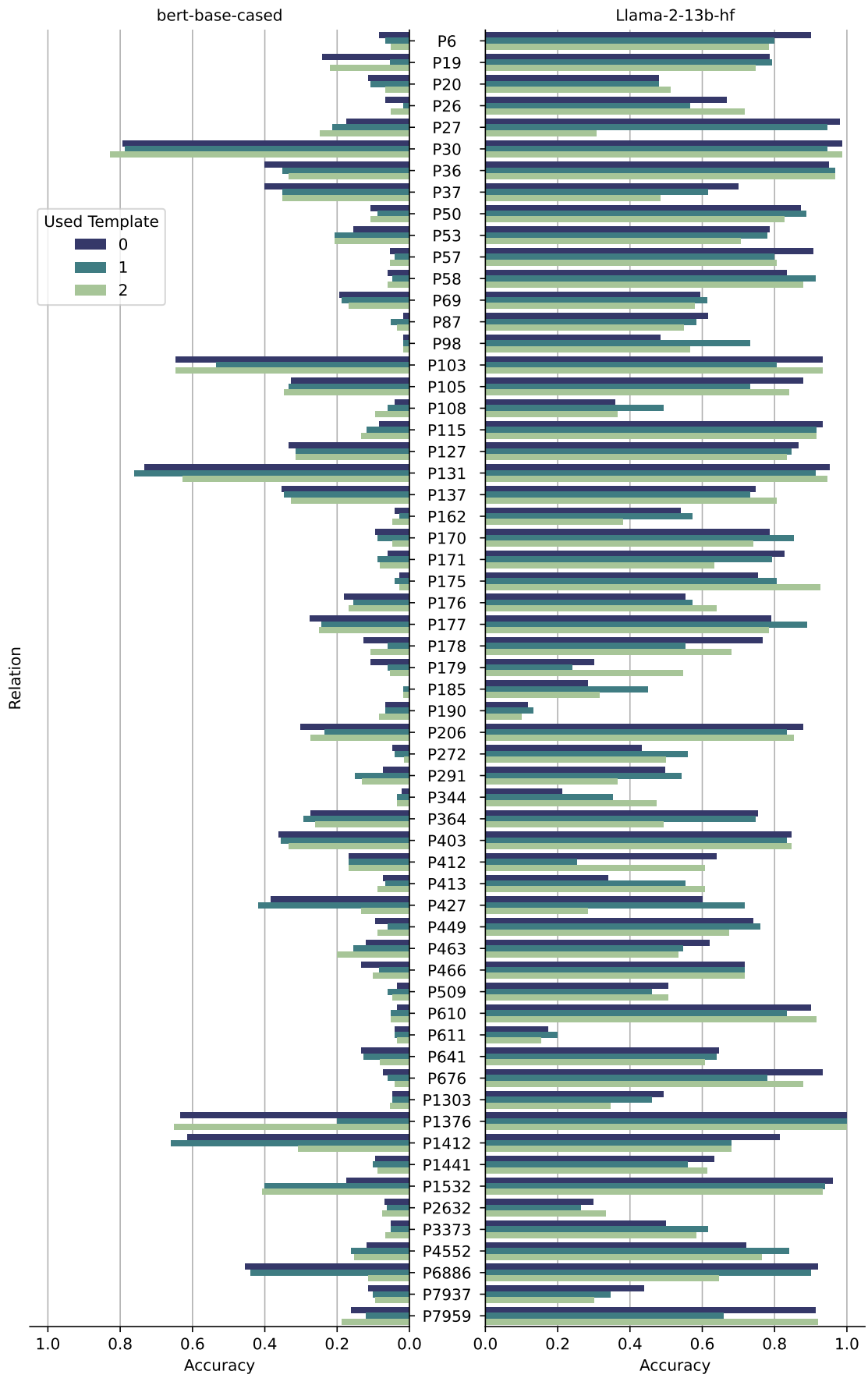
Figure 14: Accuracy of two models on each of the BEAR relations.