BELIEF: A Bias-free Estimation of Language Models in Factual Knowledge Understanding

Anonymous ACL submission

Abstract

The fill-in-the-blank prompts are widely used to evaluate how well pre-trained language models (PLMs) capture real-world factual knowledge. However, the prompt-based evaluation results vary significantly depending on the linguistic expressions of the prompts, even for the same knowledge. To evaluate the ability of PLMs in understanding facts more fairly and from diverse perspectives, we propose a new factual knowledge probing dataset - MyriadLAMA, and the evaluation benchmark, BELIEF. MyriadLAMA is a fill-in-the-blank dataset that contains numerous lexically, syntactically, and semantically diverse prompts for each fact. BELIEF is a method to mitigate prompt bias in evaluating knowledge in PLMs by aggregating results from multiple prompts for each fact. Based on MyriadLAMA, BELIF enables the comprehensive evaluation of factual knowledge in PLMs from multiple perspectives, covering the perspective of consistency and reliability. Our experiments confirm the effectiveness of the BELIEF through the evaluation of the factual knowledge of multiple PLMs such as BERT.

1 Introduction

Pre-trained language models (PLMs) are considered to be utilized as the knowledge base as they implicitly acquire and retain factual knowledge during the pre-training process. The research about evaluating the ability of PLMs in understanding facts, known as **factual knowledge probing**, is increasingly gathering attention. The LAMA probe dataset (Petroni et al., 2019) uses masked prompts (*e.g.*, John Lennon was born in [MASK].) to probe the presence of facts in PLMs. By measuring the accuracy of predicted mask tokens, LAMA probe can quantitatively gauge the PLMs' knowledge.

However, while effective, the LAMA probe relies on a single masked prompt to verify the presence of specific fact. This makes the results significantly affected by minor variations in the prompt's linguistic expression (Kassner and Schütze, 2020; Misra et al., 2020; Ravichander et al., 2020). Some studies have observed that prompts possess specific bias and using different prompt sets can significantly change the accuracy (Elazar et al., 2021; Jiang et al., 2020). As PLMs are expected to handle a wide variety of user inquiries, even for the same fact, accuracy measurement based on a single-prompt is not sufficient to make accurate evaluation. This facilitates the need to establish a more reliable and effective factual knowledge probing method.

Our study introduces BELIEF (§ 3), a benchmark designed for bias-free evaluation of PLMs in factual knowledge understanding. BELIEF is conducted based on MyriadLAMA (§ 2), a novel factual knowledge probing dataset that offers a variety of prompts for each fact, which is constructed by extending an existing dataset (Petroni et al., 2020). Specifically, we semi-automatically construct a wide variety of lexically, syntactically, and semantically diverse prompts from LAMA-UHN by manual rewritten and using GPT-4, resulting in multiple diverse prompts tied to each fact. BELIEF then integrates multiple output distributions from diverse prompts offered by MyriadLAMA to evaluate a factual knowledge, thereby mitigating the impact of individual prompt bias on evaluation. Moreover, BELIEF enables evaluation of the robustness and reliability of PLMs in fact prediction.

In experiments (§ 4), we apply BELIEF to multiple BERT models trained in different sizes and with different loss functions. Consequently, we confirm that employing multiple prompts yields a more unbiased evaluation in factual knowledge probing than relying solely on single prompt. Moreover, we assessed the PLMs' robustness and reliability in predicting factual knowledge. PLMs show performance variations across metrics, underscoring the importance of evaluating knowledge from perspectives beyond mere accuracy.

2 MyriadLAMA Dataset

In this section, we describe MyriadLAMA, the factual knowledge probing dataset that offers various prompts for each fact to support unbiased evaluation. To mitigate the impact of prompt bias in evaluation, we argue that integrating predictions from diverse prompts is important, as it can offset the bias in specific prompts. Although multiple knowledge probing datasets providing multiple prompts for each fact have been proposed, these datasets lack diversity in expressing facts, making them insufficient to provide a balanced and comprehensive evaluation (Elazar et al., 2021; Jiang et al., 2020).

In this study, we build MyriadLAMA by semi-automatically extending the existing fact probe LAMA-UHN (Petroni et al., 2020). LAMA-UHN[1] comprises single prompts corresponding to each fact extracted from Wikipedia, where each fact consists of **knowledge triples** ⟨subject, relation, object⟩ (*e.g.*, ⟨Tokyo, Capital, Japan⟩). A single template expression is provided for each "relation" (hereafter, **relational template**, *e.g.*, [X] is the capital of [Y]). The basic procedure for factual knowledge probing using LAMA-UHN is to first fill in the relational template with the target knowledge triples, replace [Y] with the [MASK] token, and generate **masked prompt** (hereafter, **prompt**). Next, it inputs the prompt into the PLMs to see if PLMs can correctly predict the "object" token.

MyriadLAMA generates multiple prompts for each fact by providing multiple relational templates for each "relation" and varying the linguistic expressions of entities ("subject" and "object"). Specifically, we define knowledge triples that neglect diversity of surface expressions as **unique triples** and distinguish them from **derived triples**, which are knowledge triples that embodies the diverse entity expressions and relational templates in each unique triple. For example, the unique triple ⟨E_{John Lennon}, R_{born-in}, E_{United Kingdom}⟩ could correspond to multiple derived triples (⟨John Lennon, born in, UK⟩, ⟨John Lennon, birthplace, United Kingdom⟩), etc.). One derived triple can generate one masked prompt (e.g., John Lennon was born in [MASK]). The overview of the triple extension method is described below.

Extension of entities The knowledge triples in LAMA-UHN constitute a subset of the Wikipedia-based T-REx knowledge base (Elsahar et al., 2018), selectively including only certain objects for "subject-relation" pairs. MyriadLAMA extends the unique triples by searching T-REx using "subject-relation" as the key to include other allowed objects. For example, if LAMA-UHN contains only E_{guitar} for instruments that "John Lennon" can play, we can extend the unique triple to include E_{piano}. We also extend the expressions of the entities using aliases obtained from Wikidata.[2] For example, the entity E_{United Kingdom} can be represented as expressions of "United Kingdom," "UK" or "Britain."

Paraphrase of relational templates MyriadLAMA creates a great variety of relational templates by a semi-automatic process. Firstly, we manually generate five distinct templates for each relation. They incorporate semantic and syntactic variations, including entailment expressions and diverse syntactic patterns like statements and question-answer formats. Next, to enhance quantity and lexical diversity, we automatically paraphrase each manually created template 19 times using the GPT-4 API.[3] Finally, all templates undergo manual verification by human reviewers, yielding a total of 4100 templates covering 41 relations.

Please refer to appendix for detailed knowledge triple extension settings (§ A.1) and the statistics of MyriadLAMA (§ A.2). Additionally, to prove the superiority and validity of MyriadLAMA, we compare it with existing multi-prompt probing datasets and conduct comprehensive evaluations (§ A.3).

3 BELIEF benchmark

In this section, we propose the benchmark BELIEF for bias-free evaluation of PLMs in fact understanding. BELIEF employs the numerous prompts from MyriadLAMA (§ 2) for a fairer and comprehensive factual knowledge probing. Beyond merely assessing the amount of facts stored in PLMs (accuracy), BELIEF further aids in evaluating the consistency and reliability of PLMs in fact prediction. In the following sections, we first outline the formulation (§ 3.1), then introduce the metrics proposed in BELIEF (§ 3.2-3.4).

[1] LAMA-UHN is a subset of LAMA probe (Petroni et al., 2019) that includes the subject as a string (*e.g.*, Apple Watch is a product of [MASK].) or a prompt that predicts the native language as a target from the subject of a person's name (*e.g.*, The native language of Jean Marais is [MASK].)), which is a more appropriate dataset for evaluating factual knowledge than LAMA.

[2] https://www.wikidata.org/wiki/Wikidata:Data_access

[3] OpenAI: gpt-4-1106-preview

3.1 Preliminary

MyriadLAMA encompasses one-to-many relations and diverse linguistic expressions referring to the same "object," allowing for several "object" tokens to be the correct response to a single prompt. For instance, with the subject E_{John Lennon} and the relation R_{born-in}, acceptable tokens could include "UK" and "Britain." Consequently, we consider the fact to be present, if the model's predicted token matches any of the correct tokens, regardless of which correct answer is predicted.

We denote the "subject-relation" pairs in MyriadLAMA as T, the set of prompts for a given "subject-relation" pair $t \in T$ as P_t, and the corresponding set of correct "object" tokens for t as C_t. We determine the correct answer for the i-th prompt in P_t as the token $a_t^i \in C_t^i$ that the PLM predicts with the highest probability. This token a_t^i, regarded as the "golden object," is then used for the following evaluation of the prompt $p_t^i \in P_t$. In addition, when the output distribution corresponding to [MASK] of the prompt p_t^i is $\mathcal{O}_i^t = \{(w_j, o_j) | \sum_j o_j = 1\}$, the prediction result is defined as the token $\hat{w}_i^t = \text{argmax}_{w_j,(w_j,o_j)\in\mathcal{O}_i^t} o_j$.

3.2 Accuracy and its fluctuations

In evaluating the prediction accuracy of the "object" for a given "subject-relation" pair, BELIEF aggregates results from multiple prompts, which mitigates the impact of individual prompt biases. This approach ensures accuracy less influenced by single-prompt bias. Furthermore, we consider top-k tokens to enable a more flexible evaluation, as relying solely on the top-1 token may only capture limited aspects of the PLMs' output distribution.

Accuracy: The accuracy metrics in BELIEF include Acc@K, indicating the proportion of prompts with the correct token predicted in the top-k output probability. We also include Mean Reciprocal Rank (MRR), which considers the rank of the correct answer, offering a more detailed understanding of the model's performance across all ranks.

$$\text{Acc@}K = \frac{\sum_{t\in T} \sum_{i}^{|P_t|} \mathbb{1}[\text{rank}(a_t^i, \mathcal{O}_i^t) \leq K]}{\sum_{t\in T} |P_t|} \quad (1)$$

$$\text{MRR} = \frac{1}{\sum_{t\in T} |P_t|} \sum_{t\in T} \sum_{i}^{|P_t|} \frac{1}{\text{rank}(a_t^i, \mathcal{O}_i^t)} \quad (2)$$

where $\text{rank}(a_t^i, \mathcal{O}_i^t)$ denotes the rank of the "golden object" a_t^i within the output probability distribution $\mathcal{O}_i^t$, and $\mathbb{1}[x]$ is an indicator function returning 1 if x is true, and 0 otherwise.

Fluctuation of accuracy: Next, we evaluate the fluctuations of accuracy based on different prompt set. For each "subject-relation" pair t, we select a corresponding prompt randomly and calculate the accuracy (as per Eq. 1 and Eq. 2 where $|P_t| = 1, \forall t \in T$). Then we repeat this process N times to obtain the set of accuracies, which are denoted as $V_{\text{Acc@}K}$ and V_{MRR}, where $|V_*| = N$ (in the experiments, $N = 50000$). For V_*, we can evaluate the fluctuation of accuracies by the range and the standard deviation as following:

$$\text{range} = \max(V_*) - \min(V_*) \quad (3)$$

$$\text{stdev} = \sqrt{\frac{1}{N}\sum_{v_i\in V_*}^{N} (v_i - \frac{1}{N}\sum_{v_i\in V_*}^{N} v_i)^2} \quad (4)$$

where V_* could be either $V_{\text{Acc@}K}$ or V_{MRR}.

3.3 Consistency

For each "subject-relation" pair t, we assess the PLM's consistency in predicting the "object" across different prompts P_t. Specifically, we compute and average the degree of match between the prediction result $\hat{w}_i^t$ for a given prompt p_t^i and the prediction results $\hat{w}_t^j$ for other prompts $p_t^j \in P_t$ (where $j \neq i$), across all "subject-relation" pairs in T (Elazar et al., 2021; Fierro and Søgaard, 2022):

$$\text{Consist@1} = \frac{1}{|T|} \sum_{t\in T} \frac{\sum_{i,j:i\neq j,i,j\leq|P_t|} \mathbb{1}[\hat{w}_i^t = \hat{w}_j^t]}{\frac{1}{2}|P_t|(|P_t|-1)} \quad (5)$$

3.4 Reliability

The reliability of PLMs reflects the extent to which we can trust the predictions they provide. This encompasses not only the prediction accuracy but also the correctness of the confidence assigned to those predictions. In our study, we use diverse prompts from MyriadLAMA to assess PLMs' overconfidence levels in making fact prediction. The overconfidence calculation draws from the expected error calibration metric (Desai and Durrett, 2020). Specially, we measure the difference between true prediction accuracy and models' confidence to their predicted tokens. For each prompt, we first acquire the maximum probability (hereafter, **confidence**) from the output distribution for [MASK]. Subsequently, all of the prompts are arranged in descending order based on confidence and segmented into

PLMs	Accuracy (Acc@1/Acc@10/MRR) ↑		Accuracy fluctuation (Acc@1/Acc@10/MRR) ↓		Consistency ↑	Reliability ↓
	LAMA-UHN	MyriadLAMA	range	stdev	Consist@1	Overconf@K (k=1,10)
$\text{BERT}_{\text{base}}$	.2403/.5377/.1767	.1051/.2941/.1696	**.1714**/.3121/.2183	.0224/.0404/.0270	.167	.220/.288
$\text{BERT}_{\text{large}}$	**.2454/.5509/.3456**	.1118/.3069/.1777	.1800/.3228/.2157	.0231/.0396/.0274	**.180**	.218/.290
BERT_{wwm}	.2448/.5248/.3380	**.1367/.3497/.2085**	.1777/**.3044/.2063**	**.0219/.0366/.0256**	.084	**.116/.164**

Table 1: Evaluation results of BERT and its variants based on BELIEF.

M bins ($P^{(1)}$, $P^{(2)}$, ..., $P^{(M)}$). For each bin i, we compute the average accuracy $\overline{\text{Acc@}K}^{(i)}$ and average confidence $\overline{o_{max}}^{(i)}$. Finally, the PLM's overconfidence in predicting the "object" is assessed by averaging differences between average confidence and accuracy across all bins, as shown below:

$$\text{Overconf@}K = \sum_{i=1}^{M} \frac{|P^{(i)}|}{M} (\overline{o_{max}}^{(i)} - \overline{\text{Acc@}K}^{(i)}) \quad (6)$$

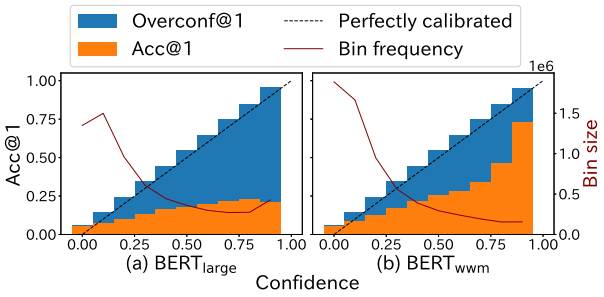

Figure 1: Overconfidence of $\text{BERT}_{\text{large}}$ and BERT_{wwm}.

4 Experiments

In this section, we use BELIEF to evaluate multiple PLMs, comparing its effectiveness with LAMA-UHN and uncovering insights hidden by single-prompt-based evaluations.

4.1 Experiment setup

We evaluate BERT and its variants, including $\text{BERT}_{\text{base}}$ (bert-base-uncased[4]), $\text{BERT}_{\text{large}}$ (bert-large-uncased[5]) and BERT_{wwm} (bert-large-uncased-whole-word-masking[6]). $\text{BERT}_{\text{large}}$ and BERT_{wwm} have 340M parameters, and are about three times larger than $\text{BERT}_{\text{base}}$ which has 110M parameters. BERT_{wwm} differs from $\text{BERT}_{\text{large}}$ in the approach of masking[7] during pre-training.

To calculate the fluctuations of accuracy (§ 3.2), it is necessary to collect multiple accuracy measures. To achieve this, we sample one prompt for each "subject-relation" pair and calculate one accuracy over all pairs. Then, by repeating this process N times, we can obtain N accuracies. About the details of our experiments, in each of the N trials, we share the same template for facts with the same relation. Additional, we set large N ($N = 50,000$) to provide stable evaluation results, and employ consistent seeds for prompt sampling for different PLMs to ensure fair comparison.

4.2 Results

Vulnerability of single prompt-based evaluation As shown in Table 1, we note significant fluctuations in accuracy among BERT and its variants. Additionally, all PLMs exhibit low prediction consistency and tend to display overconfidence in their predictions regarding facts. Below, we examine how BERT models process factual knowledge, with $\text{BERT}_{\text{large}}$ as an example.

First, the accuracy fluctuation presented in Tab 1 demonstrates significant variances of accuracy, with the average accuracy (Acc@1) at a modest 0.1118, peaking at 0.2084, and dipping to a low of 0.0284. Furthermore, the high standard deviation and low consistency (Consist@1) indicate that using different prompts for evaluation yields significantly varied prediction results. Specifically, when considering evaluation accuracy on LAMA-UHN, $\text{BERT}_{\text{large}}$ shows better accuracy than BERT_{wwm}, but this is reversed in the evaluation by BELIEF. Also, $\text{BERT}_{\text{large}}$ and BERT_{wwm} exhibited notably higher MRR compared to $\text{BERT}_{\text{base}}$, although this difference was less prominent in MyriadLAMA. These difference highlight the low trustworthiness of the single-prompt-based factual knowledge probing. Finally, we demonstrate the correspondences between the confidence and accuracy (Acc@1) of $\text{BERT}_{\text{large}}$ in the Figure 1 (a). The figure shows that

[4] https://huggingface.co/bert-base-uncased
[5] https://huggingface.co/bert-large-uncased
[6] https://huggingface.co/bert-large-uncased-whole-word-masking
[7] BERT_{wwm} masks all tokens corresponding to a single word at the same time, while $\text{BERT}_{\text{large}}$ and $\text{BERT}_{\text{base}}$ allow for partial tokens in one word to be masked.

$\text{BERT}_{\text{large}}$ shows low accuracy even for prompts with high confidence, indicating an overconfidence in its predictions. Furthermore, as shown in Table 1, increasing tokens range (K) leads to a deterioration in overconfidence.

Comparison between PLMs From Table 1, we can observe that $\text{BERT}_{\text{large}}$ outperforms $\text{BERT}_{\text{base}}$ in terms of both accuracy, consistency and reliability metrics. Moreover, BERT_{wwm} shows better performance in metrics other than consistency. This indicates that both parameter size and learning strategy, such as masking methods, are crucial for knowledge acquisition. We can also observe that BERT_{wwm} generally outperforms others with less fluctuation in prediction accuracy, though it has low consistency in prediction. This implies a possible trade-off between attaining high accuracy and maintaining consistent prediction across diverse prompts. Furthermore, BERT_{wwm} also demonstrated superior abilities in terms of reliability. As shown in Figure 1 (b), unlike $\text{BERT}_{\text{large}}$, BERT_{wwm} shows a significant correlation between confidence and actual accuracy, suggesting that its predictions are more reliable.

5 Related work

Prompt-based factual knowledge probing The LAMA probe was first proposed to evaluate the potential of using PLMs as knowledge bases using the the clozed query (prompt) (Petroni et al., 2019). It drived research of optimizing prompts that can retrieve more facts from PLMs (Shin et al., 2020; Zhong et al., 2021; Qin and Eisner, 2021; Li et al., 2022b). On the contrary, some studies questioned the validity of prompt-based factual knowledge probing, as using different prompts for the same fact could result in inconsistent predictions, making PLMs difficult to provide reliable and consistent answers (Jiang et al., 2020; Elazar et al., 2021).

Presence of prompt bias The subsequent studies contributed to understanding the reason behinds the inconsistency problem. They observed that PLMs often make correct predictions relying on prompt biases rather than truly capturing the facts (Cao et al., 2021). The prompt bias could come from the overfitting of prompts to dataset artifacts (Poerner et al., 2020; Cao et al., 2021), fact distribution leakage, or the domain overlap between pre-trained corpora and probing datasets (Zhong et al., 2021; Youssef et al., 2023; Li et al., 2022a; Cao et al., 2022). Additionally, some studies quantitatively assessed prediction consistency by evaluating diverse prompts for each fact, akin to our work (Elazar et al., 2021; Jiang et al., 2020). However, these studies often use prompts of low quality and limited diversity, making them insufficient for robustly evaluating PLMs' understanding of facts.

Bias-free factual knowledge probing Several studies have proposed the prompt debiasing methods to facilitate accurate evaluation of PLMs' understanding of facts (Zhao et al., 2021; Dong et al., 2022; Wang et al., 2023; Yoshikawa and Okazaki, 2023; Newman et al., 2021). Their approaches are orthogonal to our proposed method of diversifying prompts to alleviate the influence of individual prompt bias. Additionally, some studies mitigated individual prompt biases by aggregating multiple output distributions derived from prompt paraphrases (Jiang et al., 2020; Qin and Eisner, 2021; Kamoda et al., 2023). Although these methods employ multiple prompts akin to ours, our approach distinguishes itself by obtaining output for each prompt, enabling multifaceted evaluation encompassing accuracy, consistency, and reliability.

6 Conclusion

Our study introduces a novel benchmark for probing factual knowledge, namely BELIEF, with the aim of robustly assessing the factual knowledge of PLMs. We also construct a new factual knowledge probing dataset - MyriadLAMA, which offers diverse prompts for each fact. Based on MyriadLAMA, BELIEF proposes various evaluation metrics such as accuracy, consistency, and reliability, facilitating a comprehensive evaluation of PLMs' understanding of factual knowledge. By applying BELIEF to assess BERT and its variants, we uncover the limitations of current single-prompt-based knowledge probing methods and reveal performance variations among different PLMs, which were previously overlooked in prior research. This underscores the effectiveness of BELIEF in providing the accurate assessment of PLMs' capabilities in understanding fact.

MyriadLAMA contains an extensive amount of prompts, which leads to high evaluation costs. In the future, we aim to extract a diverse yet robust subset from MyriadLAMA to enable more efficient evaluation of factual knowledge. Ultimately, we will commit to making MyriadLAMA publicly accessible once all preparations are finalized.

References

Boxi Cao, Hongyu Lin, Xianpei Han, Fangchao Liu, and Le Sun. 2022. Can prompt probe pretrained language models? understanding the invisible risks from a causal view. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5796–5808, Dublin, Ireland. Association for Computational Linguistics.

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.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, 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.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5937–5947, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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).

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Go Kamoda, Benjamin Heinzerling, Keisuke Sakaguchi, and Kentaro Inui. 2023. Test-time augmentation for factual probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3650–3661, Singapore. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Shaobo Li, Xiaoguang Li, Lifeng Shang, Zhenhua Dong, Chengjie Sun, Bingquan Liu, Zhenzhou Ji, Xin Jiang, and Qun Liu. 2022a. How pre-trained language models capture factual knowledge? a causal-inspired analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1720–1732, Dublin, Ireland. Association for Computational Linguistics.

Yiyuan Li, Tong Che, Yezhen Wang, Zhengbao Jiang, Caiming Xiong, and Snigdha Chaturvedi. 2022b. SPE: Symmetrical prompt enhancement for fact probing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11689–11698, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635, Online. Association for Computational Linguistics.

Benjamin Newman, Prafulla Kumar Choubey, and Nazneen Rajani. 2021. P-adapters: Robustly extracting factual information from language models with diverse prompts. *ArXiv*, abs/2110.07280.

Masanori Oya. 2020. Syntactic similarity of the sentences in a multi-lingual parallel corpus based on the Euclidean distance of their dependency trees. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 225–233, Hanoi, Vietnam. Association for Computational Linguistics.

Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. *ArXiv*, abs/2005.04611.

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.

Guanghui Qin and Jason Eisner. 2021. 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. Association for Computational Linguistics.

Abhilasha Ravichander, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. On the systematicity of probing contextualized word representations: The case of hypernymy in BERT. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 88–102, Barcelona, Spain (Online). Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Yuhang Wang, Dongyuan Lu, Chao Kong, and Jitao Sang. 2023. Towards alleviating the object bias in prompt tuning-based factual knowledge extraction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4420–4432, Toronto, Canada. Association for Computational Linguistics.

Hiyori Yoshikawa and Naoaki Okazaki. 2023. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2017–2028, Dubrovnik, Croatia. Association for Computational Linguistics.

Paul Youssef, Osman Koraş, Meijie Li, Jörg Schlötterer, and Christin Seifert. 2023. Give me the facts! a survey on factual knowledge probing in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15588–15605, Singapore. Association for Computational Linguistics.

Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*.

Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. Tracing the roots of facts in multilingual language models: Independent, shared, and transferred knowledge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, Malta. Association for Computational Linguistics. To appear.

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 Appendix

A.1 Construction of MyriadLAMA

In this appendix, we explain the detailed procedure for generating the derived triples from unique triples in MyriadLAMA. As discussed in § 2, this study first extends the unique triples contained in LAMA-UHN (Petroni et al., 2020) by searching new "objects" from T-REx (Elazar et al., 2021). Next, for the obtained unique triples, we generate derived triples by combining concrete linguistic expressions associated with entities ("subjects" and "objects") and diversify relational templates using both manual labor and LLMs. We describe the detailed procedure as following.

A.1.1 The extension of entities

Extension of unique triples from T-REx LAMA-UHN is a refined subset derived from the LAMA dataset, which LAMA originates from T-REx (Elsahar et al., 2018). T-REx is a large-scale knowledge base containing 11 million real-world knowledge triples, aligned with 3.09 million Wikipedia abstracts, designed to create large-scale alignments between Wikipedia abstracts and Wikidata triples. To achieve this alignment, T-REx employed three distinct aligners—NoSub, AllEnt, and SPO—each offering varying levels of accuracy (0.98, 0.96, and 0.88, respectively) as measured on a test set. Despite the high alignment accuracy of all three aligners, LAMA-UHN selects only the triples aligned by NoSub, the aligner with the highest accuracy. While this choice ensures the high correctness of triples within LAMA, it potentially compromises the ability to fairly assess a PLM's capability in understanding facts, as it may overlook valid answers during evaluation. To address this limitation, we expand the MyriadLAMA dataset by incorporating triples aligned by all three aligners—NoSub, AllEnt, and SPO—found in T-REx, based on the "subject-relation" pairs present in LAMA-UHN. As the result, we increase the number of unique triples from 27,106 to 34,048 as shown in Tab 2.

Extension of entities using aliases Next, we utilize aliases of entities obtained from Wikidata to acquire diverse linguistic expressions (and their paraphrases) for the "subjects" and "objects". Specifically, we used the Wikidata identifiers of entities[8] and the Wikidata API[9] to retrieve the (English) alias expressions of entities. By combining the aliases of "subjects" and "objects" with the relation templates mentioned later, we generate numerous new derived triples. If N "subjects" and M "objects" are given for an unique triple, the number of derived triples according to this unique triple generated from a single relational template is $N \times M$.

A.1.2 Diversification of relation templates

We use a two-step procedure to create new relational templates, to enhance ensure both the quality and quantity. Initially, we manually rewrite relational templates, ensuring that every relation has five templates. Then, we employ the generative LLM (GPT4) to automatically paraphrase 19 additional templates. In total, we produce 100 templates for each relation.

Step 1: Manually rewriting relational templates. The manual rewriting of the relational templates is performed by the first author of this paper. We create new templates by describing the relationship between "subject" and "object" from different perspectives rather than creating templates with absolutely the same meaning with original template. Utilizing the resource provided by Wikidata [10], we not only paraphrase existing templates to generate new ones with diverse lexicons but also devise entailment expressions to encompass various semantic expressions that convey the same relations. These newly created templates are guaranteed to uphold relational equivalence, following the relationship between the "subject" and "object". Taking P20 ([X] died in [Y].)[11] as an example, we create new templates by either changing the sentence pattern or adding type information of object (e.g, [X] resided in [Y] until death). Furthermore, we also create templates without directly using the keywords of the relation (dead/death) but in a entailment way (e.g., [X] spent the last years of life in [Y].) Moreover, we devise a question-answer style template for each relation to enhance syntactic diversity. In this template, the question incorporates the subject and relation information, while the answer corresponds to the object.

Note that, during the paraphrase, we observe that some templates in LAMA-UHN only partially

[8] https://www.wikidata.org/wiki/Wikidata:Identifiers

[9] https://www.wikidata.org/wiki/Special:EntityData/<entity_identifier>.json

[10] https://www.wikidata.org/wiki/Property:<relation_identifier>

[11] https://www.wikidata.org/wiki/Property:P20

express the original meaning of relations defined in Wikidata. These are inappropriate for specific knowledge triples. For example, P136 describes the creative work's genre or an artist's field of work[12], which the type of work includes music, film, literature, etc. However, the original templates of P136 in LAMA-UHN is "[X] plays [Y] music.," which cannot correctly retrieve information on work other than music. For this kinds of template, we abandon the original templates and newly create five templates.

Step 2: Paraphrasing templates using GPT-4 Based on the original relation templates and the relation templates rewritten manually, we further paraphras these relation templates automatically using the GPT4-API (gpt-4-1106-preview[13]) provided by OpenAPI. The instruction for paraphrasing used for GPT-4 generation is:

> *You are a professional tool that can paraphrase sentences into natural sentences that can correctly represent the relationship between [X] and [Y], without repetition. Make the paraphrase as diverse as possible using simple words. Please paraphrase the given sentence 19 times.*

When the duplicated sentence is generated, we remove the duplication and regenerate new templates with the same instruction, until 19 different templates is generated. Furthermore, we observe that GPT-4 occasionally generates relation templates that are semantically inappropriate for specific relationships due to incorrect category information of entities. Consequently, in such instances, we refine the instructions to include the category information of the entities, ensuring accurate representation of the relationship between the subjects and the objects. For example, when paraphrasing the relational template "[X] used to work in [Y]."[14], we additionally add explicit guidance regarding the expected format and semantics of the relation templates to the above instruction, as following.

> *Be aware that [Y] is the geographic location but NOT company or organization, where persons or organizations were actively participating in employment, business or other work.*

[12] https://www.wikidata.org/wiki/Property:P136

[13] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

[14] https://www.wikidata.org/wiki/Property:P937

	LAMA-UHN	MyriadLAMA
Relational templates	41	4100
Unique triples	27,106	34,048
Derived triples	27,106	21,140,500
Subject-relation pairs	24,643	24,643
Prompts	24,643	6,492,800

Table 2: The statistics of LAMA-UHN and MyriadLAMA.

As a result, we can obtain the following paraphrased relational templates for "[X] used to work in [Y].":

- "[X] was formerly employed in [Y]."
- "[X] once worked at [Y]."
- "[Y] was the place where [X] used to be engaged in work."

A.2 The statistics of MyriadLAMA

In this section, we report the statistics of LAMA-UHN and MyriadLAMA, including the number of "subject-relation" pairs, prompts, relational templates, and various types of triples for both LAMA-UHN and MyriadLAMA, as shown in Tab 2.

The numbers presented in Tab 2 exclude elements with objects containing multiple tokens. This exclusion is based on previous findings indicating that the performance of PLMs in predicting facts is significantly influenced by the number of mask tokens (Zhao et al., 2024). Consequently, our study concentrates exclusively on evaluating derived triples in which the "object" is represented as a single token following tokenization by the WordPiece tokenizer utilized by BERT (Devlin et al., 2019) and its variants.

As the result, we increase the number of unique triples from 27,106 to 34,048 by extending object entities for one-to-many relations after searching T-REx. Furthermore, the number of derived triples is increased from 27,106 in LAMA-UHN to 21,140,500, an increase of approximately 778 times, by combining various semi-automatically generated relational templates and the alias expressions for "subject" and "object" entities. As the prompts are generated from derived triples without considering the "object" expressions, the number of generated prompts are less than the number of derived triples, which is increased from 27,106 to 6,492,800.

PLMs	LAMA-UHN	MyriadLAMA		
		Min	Max	Mean
$\text{BERT}_{\text{base}}$	.2403	.0000	.3534	.1103
$\text{BERT}_{\text{large}}$	.2454	.0007	.3728	.1185
BERT_{wwm}	.2448	.0015	.3695	.1453

Table 3: Acc@1 of MyriadLAMA and LAMA-UHN

PLMs	Consist@1↑		Acc@1 range (min/max)	
	Subject	Relation	Subject	Relation
$\text{BERT}_{\text{base}}$	.5745	.1504	.0673/.1441	.0000/.3534
$\text{BERT}_{\text{large}}$	.5497	.1548	.0714/.1554	.0007/.3728
BERT_{wwm}	.5005	.1057	.0831/.1884	.0015/.3695

Table 4: Diversity evaluation of subjects and relation templates

A.3 Evaluation of MyriadLAMA

Given that our proposed knowledge probing method BELIEF seeks to mitigate the influence of individual prompt bias in evaluations, the availability of a wide range of prompts characterized by both quality and diversity is crucial. Quality ensures that the prompts can accurately inquire the target facts, while diversity ensures that multiple prompts can capture different aspects of the true knowledge distribution. In this section, we verify these two properties from three aspects: accuracy (Acc@1), fluctuation of accuracy (range of Acc@1), and prediction consistency (Consist@1).

The quality of MyriadLAMA prompts We evaluate the quality of the relation templates in MyriadLAMA the accuracy measurement based on all the derived prompts evaluated on PLMs. Specifically, for each relation, we evaluate the accuracy (Acc@1) of all relation template separately, and then calculate the minimum, maximum accuracies among all templates for each relation. We then measure the dataset-level minimum/maximum accuracy by micro-averaging the templates set with the minimum/maximum template accuraies (41 templates in each set). Finally, all of the template-specific accuracies are then micro-averaged to compute the average Acc@1. As indicated in Table 3, while the quality of MyriadLAMA's prompts significantly varies, the high-quality prompts are notably superior to those of LAMA-UHN. Although the average accuracy of MyriadLAMA is lower than that of LAMA-UHN, it is considered that this is because MyriadLAMA uses relation templates that have been semi-automatically created, whereas LAMA-UHN uses carefully selected entities and templates.

Prompt diversity evaluation Next, in order to gauge the diversity of prompts in MyriadLAMA, we examine both the consistency (Consist@1) and the range of accuracy (min/max) across various expressions of subjects or relations, assessed individually. To achieve this, the complete set of prompts was partitioned into multiple subsets, with each subset containing only one expression for each unique subjects or relations. The Acc@1 of the prompts obtained in this manner is then evaluated using different variants of BERT.

The results in Tab 4 indicate that while the accuracy range (min/max) and consistency (Consist@1) caused by aliases of subjects is less pronounced compared to diverse expressions of relational templates, its effect on factual knowledge evaluation remains significant. These findings highlight the vulnerability of factual knowledge evaluation based on single prompts and underscore the significance of harnessing the diversity of prompts within MyriadLAMA for robust assessments.

Comparison between multi-prompts probing datasets We conduct comparison between MyriadLAMA and other multi-prompts probing datasets from the perspective of quantity and diversity. Specially, we measure the average prompts for each "subject-relation" pair as the **quantity** measure.

Then, when evaluating diversity, we ignore the variations stemming from different subject expressions and focus solely on comparing relational templates. we measure the diversity of relational templates from three aspects: lexicon, syntax and semantic. The details are shown below:

Lexicon: We utilize the Jaccard distance of words in templates as a metric to gauge lexicon diversity. Specifically, we begin by generating the set of words following the pipeline of tokenization, stemming, and lemmatization for each sentence. Next, we compute the Jaccard distance of word sets between different templates under the same relation, ultimately averaging these distances.

Syntax: We adopt the syntax distance measure proposed in (Oya, 2020), which calculates the distance between dependency trees.

Semantics: We quantify semantic diversity by calculating the L2 distance of sentence embed-

Dataset	Quantity↑	Diversity↑		
		Lexicon	Syntax	Semantic
PARAREL	7.30	.4860	.1489	11.03
LPAQA	53.27	.5449	.1713	**13.55**
MyriadLAMA	**263.47**	**.6652**	**.2138**	12.69

Table 5: Comparison of multi-prompts probing datasets

PLMs	Average rank of manual prompts based on Acc@1	Consist@1	
		Inner-group	Inter-group
$\text{BERT}_{\text{base}}$	47.40	.2904	.1065
$\text{BERT}_{\text{large}}$	45.64	.2884	.1125
BERT_{wwm}	44.80	.2387	.0630

Table 6: Comparison between prompts generated through manual labor and LLM.

dings, utilizing the representation of the [CLS] token provided by $\text{BERT}_{\text{large}}$ as the sentence embedding.

As shown in Table 5, MyriadLAMA demonstrates a great quantity and diversity comparing to the existing multi-prompt factual probing datasets: LPAQA (Jiang et al., 2020) and PareREL (Elazar et al., 2021). While LPAQA exhibits greater semantic diversity in its measures, this is primarily attributed to its utilization of a mining-based approach for template discovery, which relies on the concept of distance supervision. LPAQA identifies all Wikipedia sentences containing both subjects and objects of a specific relation as candidate templates. However, this approach often results in problematic templates that inadequately describe the relationships between subjects and objects, lacking the precision needed to accurately express relations. For example, for relation P937 ([X] used to work in [Y].), the mined templates in LPAQA includes templates like:

- "[X] was born in [Y]."
- "[X] returned to [Y]."
- "[Y] artist [X]."
- "[X] to meet [Y]."

These prompts significantly deviate from the original semantic meaning. In contrast, every prompt in MyriadLAMA can precisely describe the correct relationship. Despite this, MyriadLAMA still exhibits comparable semantic diversity to LPAQA, indicating its ability to provide diverse semantic prompts while maintaining precision.

Manually rewritten vs. auto-generated templates Upon comparing relational templates generated through manual rewriting and GPT-4 auto-generation, we find that auto-generated templates exhibit comparable quality (accuracy) to manually rewritten templates; they also demonstrate less diversity in acquiring different predictions, aligning with our expectations.

To assess the validity of LLM-generated templates for knowledge probing, we rank the accuracies (Acc@1) of manually created templates against those generated by LLMs. Specifically, for each relation, we rank the 5 manual templates among all 100 templates and calculate the average rank across all manually created templates for all relations. Tab 6 shows the average Acc@1 ranks of manual templates among 100 templates on $\text{BERT}_{\text{base}}$, $\text{BERT}_{\text{large}}$, BERT_{wwm}. They are 47.40, 45.64, and 44.80, respectively. These values closely approximate the average rank of 50, indicating that auto-generated templates can achieve nearly the same accuracy as manually created templates.

Furthermore, we quantify the diversity discrepancy between manually written and auto-generated templates. We categorize the auto-generated templates, including the original ones, as one group, resulting in five groups for each relation, each comprising 20 templates. Subsequently, we evaluate the similarity between templates within the same group and across different groups using the consistency measure (Consist@1), as presented in Tab 6. The consistency among prompts within the same group (inner-group) is notably high, whereas prompts from different groups (inter-group) exhibit less diversity in predictions. This underscores the significance of manual phrase rewriting, which can yield more diverse prompts and facilitate a more comprehensive evaluation.