BELIEFs: Bias-resilient, Multifaceted Evaluation 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 assess PLMs' capability to understand facts more fairly, we introduce a new dataset called MyriadLAMA, along with the evaluation benchmarks BELIEF and its variant BELIEF-ICL to evaluate encoder- and decoder-based PLMs, respectively. MyriadLAMA presents diverse fill-in-the-blank prompts for the same fact, leveraged by BELIEFs not only to mitigate prompt bias during factual knowledge probing by consolidating results from multiple prompts but also to offer a comprehensive evaluation of factual knowledge in PLMs, including accuracy, consistency and reliability. We validate the efficacy of the BELIEFs through comprehensive evaluations of encoder-based and decoder-based PLMs.

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, the 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) and its variant BELIEF-ICL (§5.1), benchmarks designed for bias-resilient evaluation of encoder- and decoder-based PLMs in factual knowledge understanding. The evaluation of BELIEFs is conducted using MyriadLAMA (§2), a new factual knowledge probing dataset. It significantly expands an existing dataset LAMA-UHN (Petroni et al., 2020) by offering multiple prompts for each fact. Specifically, we obtain a wide variety of lexically, syntactically, and semantically diverse prompts from LAMA-UHN by rewriting manually and then rephrasing them using GPT-4, resulting in myriad diverse prompts tied to each fact. BELIEFs then integrate the outputs from diverse prompts offered by MyriadLAMA to evaluate specific knowledge, thereby mitigating the impact of individual prompt bias on evaluation and offering multifaceted evaluation of the robustness and reliability of PLMs in fact prediction.

We applied BELIEFs to various PLMs, including BERT (Devlin et al., 2019) (§4) and Llama2 families (Touvron et al., 2023) (§5.1). Consequently, we confirm that diverse prompts enables i) a bias-resilient factual knowledge probing and ii) a multifaceted evaluation of PLMs' knowledge in terms of robustness and reliability beyond 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 evalua-

tion. 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).

2.1 Dataset construction

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]). LAMA-UHN was originally designed for encoder-based PLMs, which can utilize bidirectional information for mask prediction. The procedure for factual knowledge probing using LAMA-UHN is to first fill in the relational template with the target knowledge triples, replace [Y] with a mask token, and generate **masked prompt** (hereafter, **prompt**). Next, it feeds prompts into PLMs to see if PLMs can correctly predict the "object".

MyriadLAMA generates multiple prompts for each fact by using many relational templates for each "relation" and varying the linguistic expressions of entities ("subject" and "object"). Specifically, we define knowledge triples that neglect the diversity of surface expressions as **unique triples** and distinguish them from **derived triples**, which are knowledge triples that embody 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.). The derived triple can be used to create the masked prompt (*e.g.*, John Lennon was born in [MASK]). The overview of the triple extension method is described below. Please refer to §A.1 for more detailed knowledge triple extension settings.

[1]LAMA-UHN is a subset of LAMA probe (Petroni et al., 2019) and deletes overly helpful entity names that allow name-based reasoning (*e.g.*, Apple Watch is a product of [MASK].), thus enabling more reliable factual knowledge probing.

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

Paraphrasing 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 entailment expressions and diverse syntactic patterns like statements and question-answer formats to provide semantic and syntactic variations. 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.

2.2 Dataset analysis

In this section, we report the statistics of MyriadLAMA and compare it with other factual knowledge probing dataset, including LAMA-UHN and multi-prompts datasets. Our MyriadLAMA demonstrates superiority in providing more diverse prompts for the same knowledge while maintaining the quality of each prompt.

Statistics We first report the statistics of LAMA-UHN and MyriadLAMA, as shown in Table 1. Due to previous findings that the performance of PLMs in predicting facts is significantly influenced by the number of mask tokens (Zhao et al., 2024), our study focuses exclusively on evaluating derived triples in which the "object" is represented as a single token following tokenization by the WordPiece tokenizer (Devlin et al., 2019).

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. Furthermore,

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

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

	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 1: Statistics of LAMA-UHN and MyriadLAMA.

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 2: Comparison between multi-prompts datasets.

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.

Diversity comparison Given that our study seeks to mitigate the influence of individual prompt bias in evaluations, the availability of a wide range of prompts characterized in both quantity and diversity is crucial. The diversity ensures that different prompts can capture different aspects of the true knowledge distribution.

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. MyriadLAMA introduces diversity into prompts by using various subject expressions and relational templates. On average, MyriadLAMA provides 2.47 expressions for each subject. In addition, we measure the diversity of relational templates from three aspects, as shown below:

Lexicon: We utilize the Jaccard distance of words in templates to gauge lexicon diversity.

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 embeddings given by $\text{BERT}_{\text{large}}$.

As shown in Table 2, 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 distance supervision to discover new templates. Such method often results in problematic templates that inadequately describe the relationships between subjects and objects. For example, for relation P937 ([X] used to work in [Y].), the mined templates in LPAQA includes templates like: "[X] to meet [Y].", that significantly deviate from the original semantic meaning. In contrast, every prompt in MyriadLAMA can precisely describe the correct relationship. Refer to §A.2 for further ablation analysis on MyriadLAMA.

3 BELIEF Benchmark

In this section, we propose the benchmark BELIEF for bias-resilient evaluation of encoder-based 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).

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 single prompts. 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 $p_t^i \in P_t$ as the token $a_t^i \in C_t$ 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 addition, when the output distribution corresponding to mask token of arbitrary prompt p is $\mathcal{O} = \{(w_j, o_j) | \sum_j o_j = 1\}$, the prediction result is defined as the token $\hat{w} = \text{argmax}_{w_j,(w_j,o_j)\in\mathcal{O}} o_j$.

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. Specifically, we randomly select one prompt for each "subject-relation" pair $t \in T$ to collect the set of prompts $P = \{p_1, ..., p_{|T|}\}$. By feeding prompts P to PLMs, we can calculate accuracy based on their predictions. We repeat this process to collect a set of accuracies, which is then used to measure both the average and fluctuation.

Average accuracy In BELIEF, accuracy metrics include Acc@K, which measures the proportion of prompts with the correct token predicted within the top-k output probabilities. Considering top-k tokens allows for a more flexible evaluation, as relying solely on the top-1 token may capture only limited aspects of the PLMs' output distribution. 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. For each sample prompts set, we calculate Acc@K and MRR as follows:

$$\text{Acc@}K = \frac{\sum_t^{|P|} \mathbb{1}[\text{rank}(a_t, \mathcal{O}_t) \leq K]}{|P|} \quad (1)$$

$$\text{MRR} = \frac{1}{|P|} \sum_t^{|P|} \frac{1}{\text{rank}(a_t, \mathcal{O}_t)} \quad (2)$$

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

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 $|V_*| = N$. The final average accuracy is calculated as the mean value of V_*.

Fluctuation of accuracy: 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 in P_t. Specifically, we compute the degree of match between the prediction result $\hat{w}_t^i$ 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 the mask token. Subsequently, all of the prompts are arranged in descending order based on confidence and segmented into M bins ($P^{(1)}$, $P^{(2)}$, ..., $P^{(M)}$), with the same amount of data points in each bin. For each bin i, we compute the average accuracy $\overline{\text{Acc@}K}^{(i)}$ and average confidence $\overline{o_{max}}^{(i)}$. In our work, we use $M = 10$ for all the experiments. 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:

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

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	.1098	.220/.288
$\text{BERT}_{\text{large}}$	**.2454/.5509/.3456**	.1118/.3069/.1777	.1800/.3228/.2157	.0231/.0396/.0274	**.1119**	.218/.290
BERT_{wwm}	.2448/.5248/.3380	**.1367/.3497/.2085**	.1777/**.3044/.2063**	**.0219/.0366/.0256**	.1021	**.116/.164**

Table 3: Evaluation results of BERT and its variants via BELIEF.

4 Encoder-Based PLMs Evaluation

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

4.1 Experiment setup

We evaluate BERT families, including $\text{BERT}_{\text{base}}$,[4] $\text{BERT}_{\text{large}}$,[5] and BERT_{wwm},[6] $\text{BERT}_{\text{base}}$ and the other two models have 110M and 340M paramters, respectively. 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), we set a large sample number ($N = 50,000$) to provide stable and accurate evaluation results. In each of the N trials, we share the same template for facts with the same relation. We also employ consistent seeds for prompt sampling for different PLMs to ensure fair comparison.

4.2 Results and analysis

Vulnerability of single prompt-based evaluation As shown in Table 3, 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.

Below, we examine how BERT models perceive facts, with $\text{BERT}_{\text{large}}$ as an example. First, the accuracy fluctuation presented in Table 3 demonstrates variances. The high stdev and low Consist@1 also indicate that using different prompts for evaluation yields significantly varied predictions. Moreover, we observe that even $\text{BERT}_{\text{large}}$ exhibits higher accuracy than BERT_{wwm} in LAMA-UHN, the relationship is reversed in BELIEF. Similarly, the MRR gain of $\text{BERT}_{\text{large}}$ over $\text{BERT}_{\text{base}}$ is less prominent in MyriadLAMA. These discrepancies underscore the unreliability of knowledge probing using single prompts.

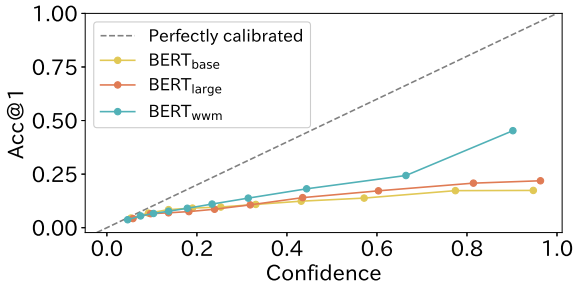

Figure 1: Overconf@1 of BERT and is variants.

Finally, Figure 1 illustrates the relationship between confidence and Acc@1 of $\text{BERT}_{\text{large}}$. The figure indicates that $\text{BERT}_{\text{large}}$ exhibits low accuracy even for prompts with high confidence. Additionally, expanding the token range (K) leads to further deterioration in overconfidence, as detailed in Table 3. These results underscore PLMs' tendency towards overconfidence in predictions.

Comparison between PLMs From Table 3, 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 accuracy, though it has low consistency. This implies a possible trade-off between attaining high accuracy and maintaining consistent prediction across diverse prompts. Furthermore, BERT_{wwm} demonstrated superior abilities on reliability, as can be also seen in Figure 1.

[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.

5 Decoder-Based PLMs Evaluation

We extend the benchmark BELIEF to incorporate decoder-based large language models (LLMs). Due to different nature of decoder-based and encoder-based PLMs, the fill-in-the-blank style dataset is not suited to evaluate LLMs' abilities in factual knowledge understanding. To comprehensively evaluate these models, we introduce a modified version of BELIEF employing in-context learning (ICL), termed BELIEF-ICL (§5.1). Finally, we conduct a thorough evaluation and analysis of several LLMs based on the BELIEF-ICL (§5.3).

5.1 BELIEF-ICL

Evaluating factual knowledge directly using BELIEF poses several challenges for recent decoder-based LLMs. Unlike the encoder-based models, which can predict `[MASK]` tokens based on all surrounding contexts, decoder-based models encounter difficulties with prompts containing mask tokens in the middle of sentences in MyriadLAMA. Furthermore, while encoder models allow for specifying the number of answer tokens by setting masks, locating answers precisely in decoder-based models proves challenging due to their auto-regressive generation process.

In-context learning settings We utilize the in-context learning ability of LLMs to solve the challenges. The in-context learning ability allows LLMs to perform complex tasks during inference using task-specific prompts (Brown et al., 2020). Each prompt contains three components: the **task instruction**, **few-shot learning context** and the target **knowledge prompt**.

In this study, we develop two prompt types to fit different relational templates as follows.

1) QA task: Initially, we define the question-answer (QA) prompts utilizing the QA-style templates available in MyriadLAMA.[8] For the QA prompt, we employ the few-shot prompt comprising random QA pairs, following the format outlined in InstructGPT (Ouyang et al., 2022). Given that all objects in MyriadLAMA are intended to be matched with single words, we prepend the instruction "`Answer each question in one word.`"

2) MP prompt: We introduce the mask prediction (MP) prompt style, which is accessible for all templates. The task instruction is formulated as "`Predict the [MASK] in each sentence in one word.`" For prompts of the few-shot examples and questions, we adhere to the same conventions as BELIEF, replacing the object placeholder with "`[MASK]`" within the template.

Evaluation setup The evaluation of accuracy and its fluctuation, consistency and reliability mostly follow the instruction in §3. However, unlike encoder-based PLMs where we can pre-define the set of candidate answers using the single mask token in prompts, LLMs pose challenges in measuring the matching between two sequences due to the diverse and autoregressive generation. To mitigate the impact of diverse generation, we normalize all the generated sequences and object entities through tokenization and lemmatization.

When evaluating matching, we check if the normalized sequence contains any of the candidate object entities. We only report the accuracy (Acc@1) and overconfidence (Overconf@1) of the greedy generation with the highest probability for LLMs and ignoring metrics where $k > 1$, as determining the rank of generated answers poses significant challenges. For consistency calculation, we evaluate the matching between two sequences bidirectionally to ensure better coverage. To gauge the confidence of the prompt's greedy generation, we employ multinomial sampling decoding strategy, [9] repeating the process 100 times. We then determine the confidence level by calculating the percentage of generations that match the greedy generation.

5.2 Experiment setup

We apply BELIEF-ICL to three Llama2 models with different parameter sizes: 7b,[10] 13b,[11] and 70b.[12] To examine the effectiveness of in-context learning settings, we adopt eight patterns of ICL prompts by combining two task instructions (QA, MP) and four types of contexts as follows. i) **zero-shot**: no context; ii) **4-random**: sampling 4 facts from all relations as the few-shot learning examples; iii) **4-relation**: sampling 4 facts from the same relation but with random templates; iv) **4-template**: sampling 4 facts from the same relations and the

[8] MyriadLAMA provides 20 QA-style templates for each relation, offering not only syntactical diversity but also accommodating causal language modeling in decoder-based PLMs. Each QA-style prompt follows a format in which the subject and relation form the question, and the object serves as the answer, such as "`Who developed [X]? [Y].`"

[9] Multinomial sampling selects a next token according to the probability over the entire vocabulary given by the model.

[10] https://huggingface.co/meta-llama/Llama-2-7b

[11] https://huggingface.co/meta-llama/Llama-2-13b

[12] https://huggingface.co/meta-llama/Llama-2-70b

ICL settings	Fact prediction (Acc@1)		Single-word generation	
	QA	MP	QA	MP
zero-shot	**.5862**	.5326	**.5820**	.5713
4-random	**.6077**	.5973	.8144	**.8376**
4-relation	**.6754**	.6715	.9243	**.9288**
4-template	**.6881**	.6812	.9237	**.9284**

Table 4: Instruction following rate on Llama2-7b.

	PLMs	Acc@1	Fluctuation		Consist @1	Overconf @1
			range	stdev		
BERT	$\text{BERT}_{\text{base}}$	.1095	.1534	.0217	.1682	.2154
	$\text{BERT}_{\text{large}}$	.1102	.1574	.0220	.1713	.2052
	BERT_{wwm}	.1364	.1517	.0208	.1524	.1000
Llama2-7b	zero-shot	.4323	.1962	.0248	.1923	-.0922
	4-random	.5350	.1743	.0234	.2786	-.0920
	4-relation	.6485	.0737	.0103	.3939	**-.0913**
	4-template	**.6711**	**.0282**	**.0036**	**.4158**	-.0920

Table 5: Evaluation on BERTs and Llama2-7b.

same template. In the few-shot learning settings, we ensure that the probed fact is excluded in the context. Refer to §A.4 for examples of prompts.

Given the limited number of templates (20 for each relation), QA-style prompts represents one-fifth of the full prompts in MyriadLAMA. Additionally, owing to the high computational cost of LLMs' inference, we select only five manually rewritten templates to represents in MP-style prompts. Due to the considerable computational cost involved in sampling 100 generations when calculating Overconf@1, we opt for an efficient approach. Specifically, we sample 10,000 prompts from 10,000 unique subject-relation pairs and utilize them for the calculation. For other evaluation scenarios, we adhere to the same settings outlined in §4.1.

5.3 Results and analysis

Does ICL prompting adhere to instructions? Our initial investigation focuses on evaluating the effectiveness of the proposed ICL settings in adhering to instructions, from two perspectives: predicting facts and generating single-word answers. The latter is crucial as the target objects in MyriadLAMA primarily consist of single-word entities. To ensure a fair comparison between QA- and MP-style ICL, we conduct evaluations using shared templates in both settings, namely, manually created templates, with only one for each relation.

We evaluate the abilities of fact prediction and single-word generation individually on Llama2-7b using Acc@1 and single-word generation rate metrics. As demonstrated in Table 4, Llama2-7b exhibits a remarkable capability to comprehend instructions for answering questions and generating single-word answers. We observe that the QA-style instruction performs better on both perspectives when no context is provided, possibly due to the decoder-based PLMs' ability to generate in a casual manner. However, this gap diminishes with the use of few-shot examples. Moreover, by comparing zero-shot and few-shot ICL, we conclude that while instructions alone (few-shot) achieve a comparable compliance rate, incorporating additional and explicit contexts significantly enhances prompt adherence to instructions.

Are LLMs strong factual knowledge learner? To assess the performance gap between encoder- and decoder-based LLMs, we evaluate BERTs by BELIEF using the same templates employed in the MP task and BELIEF-ICL on Llama-7b's MP-style prompts [13]. As shown in Table 5, The discrepancy of Acc@1 between zero-shot and few-shot ICL highlights the significant improvement in LLMs' ability to recall factual knowledge through few-shot learning. Furthermore, the selection of few-shot examples is also critical to model performance. By comparing the three few-shot ICL settings in Table 5, we consistently observe performance improvements across all metrics when using more related and explicit examples.

Decoder-based LLMs show great superiority in understanding factual knowledge than encoder-basde models. As depicted in Table 3, Llama2-7b shows great average accuracy, with even the zero-shot ICL largely outperforming BERT models. LLMs also exhibit minimal fluctuation and high consistency, highlighting their ability to comprehend and unify linguistic nuances within semantically equivalent representations. In Figure 2, LLMs generally exhibit superior calibration and more appropriate confidence levels compared to BERT models (Figure 1), indicating their ability to generate reliable answers and they are well-calibrated, albeit slightly underconfident.

Can Larger models recall factual knowledge better? We examine the evaluation of Llama2 with different sizes (7b, 13b 70b), using the MP-style

[13] We opt here for MP-style prompts for evaluation considering semantic diversity and computational cost. The evaluation results on QA-style prompts are documented in §A.3.

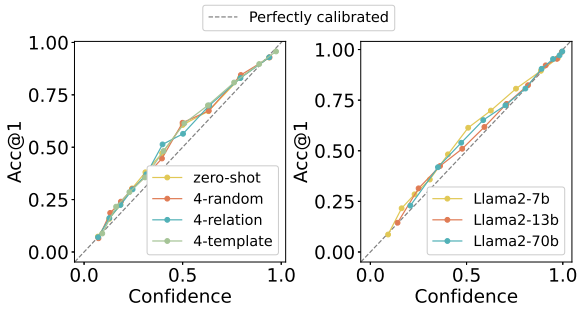

Figure 2: Comparison of Overconf@1 among Llama2 models. (**Left**: Llama2-7b with four ICL types; **Right**: Llama2 models with different sizes.)

PLMs	Acc@1	Fluctuation		Consist @1	Overconf @1
		range	stdev		
Llama2-7b	.6713	.0277	.0036	.4158	-.0923
Llama2-13b	.7095	.0270	.0033	.4314	**-.0662**
Llama2-70b	**.7784**	**.0183**	**.0024**	**.4449**	-.0690

Table 6: Comparison of LLMs with different sizes

4-template ICL setting. As shown in In Table 6, We find that larger LLMs consistently achieve higher accuracy in retrieving factual knowledge, with the 70b model outperforming the 7b models by 10%. Moreover, larger models also demonstrate better robustness in handling different prompts. For reliability, as shown in Regarding reliability, as shown in Figure 2, we find that neither the ICL settings nor the model sizes significantly affect the Overconf@1 measure, indicating that well-calibration is likely an intrinsic nature of decoder-based LLMs.

6 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-resilient 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.

7 Conclusions

Our study introduces novel benchmarks, BELIEF and its variants, BELIEF-ICL, for comprehensive factual knowledge probing across various types of PLMs. Additionally, we present a new dataset, MyriadLAMA, featuring diverse prompts for each fact. Leveraging MyriadLAMA, BELIEFs propose various evaluation metrics, including accuracy, consistency, and reliability, enabling a thorough assessment of PLMs' comprehension of factual knowledge. By conducting extensive evaluation on both encoder-based PLMs and recent LLMs, 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 BELIEFs in providing the accurate assessment of PLMs' capabilities in understanding fact.

8 Limitations

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. Our study focuses on two families of PLMs, which may not fully capture the impact of different language model pretraining paradigms on factual knowledge understanding. To address this limitation, we intend to broaden our evaluation by including a broader array of LLMs, spanning various types of encoder-based and decoder-based PLMs. Ultimately, we will commit to making MyriadLAMA publicly accessible once the paper is accepted.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

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.

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

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 (Volume 1: Long Papers)*, pages 2088–2102, St. Julian's, Malta. Association for Computational Linguistics.

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 Table 1.

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[14] and the Wikidata API[15] 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 [16], 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].)[17] 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 an-

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

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

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

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

swer corresponds to the object.

Note that, during the paraphrase, we observe that some templates in LAMA-UHN only partially 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[18], 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[19]) 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]."[20], 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.*

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 Ablation analysis 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).

Quality evaluation of MyriadLAMA relational templates 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 7, 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.

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

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

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

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 7: 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 8: Diversity evaluation of subjects and relation 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 Table 8 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.

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. Table 9 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.

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 9: Comparison between prompts generated through manual labor and LLM.

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 Table 9. 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.

A.3 QA-style vs MP-style prompts

In this section, we report the BELIEF-ICL evaluation result based on the QA-style and compare it to MP-style prompts.

As depicted in Figure 10, we observe that QA-style prompts consistently outperform MP-style prompts in accuracy across all four types of context settings. This could be attributed to QA-style prompts offering a more natural way for decoder-based models trained in a casual manner for generation. Additionally, despite QA-style prompts consisting of 20 templates for each relation, which is four times more than the 5 templates used in MP-style prompts, QA-style still exhibits great consistency and smaller fluctuations.

	PLMs	Acc@1	Fluctuation		Consist @1	Overconf @1
			range	stdev		
QA	zero-shot	.5087	.1532	.0196	.1960	-.0909
	4-random	.5606	.1448	.0168	.2910	-.0905
	4-relation	.6670	.0253	.0032	.4393	-.0889
	4-template	**.6780**	**.0221**	**.0025**	**.4411**	**-.0845**
MP	zero-shot	.4323	.1962	.0248	.1923	-.0922
	4-random	.5350	.1743	.0234	.2786	-.0920
	4-relation	.6485	.0737	.0103	.3939	-.0913
	4-template	.6711	.0282	.0036	.4158	-.0920

Table 10: BELIEF-ICL evaluation on Llama2-7b with QA-style and MP-style prompts.

A.4 Examples of in-context learning prompt

In this section, we give prompts of eight patterns introduced in our study. The eight patterns origins from the combination two types of instructions (QA- and MP-style) and four types of context.

A.4.1 MP-style/zero-shot

```
Predict the [MASK] in each sentence in one word.
Q: [MASK] consists of LAUPT.
A:
```

A.4.2 MP-style/4-random

```
Predict the [MASK] in each sentence in one word.
Q: [MASK] is the administrative center of Jiangsu.
A: Nanjing.
Q: Mar del Plata and [MASK] are sister cities that have been developing together.
A: Havana.
Q: Malawi has established diplomatic ties with [MASK].
A: Australia.
Q: Which country is House of Representatives located? [MASK].
A: Libya.
Q: [MASK] consists of LAUPT.
A:
```

A.4.3 MP-style/4-relation

```
Predict the [MASK] in each sentence in one word.
Q: What is the overarching group for Panzer Division Kempf? [MASK].
A: Wehrmacht.
Q: To whom does Mount Bulusan relate? [MASK].
A: Luzon.
Q: Who is responsible for Army National Guard? [MASK].
A: National Guard.
Q: What group is pharmacy a part of? [MASK].
A: biology.
Q: [MASK] consists of environmental factors.
A:
```

A.4.4 MP-style/4-template

```
Predict the [MASK] in each sentence in one word.
Q: [MASK] consists of Panzer Division Kempf.
A: Wehrmacht.
Q: [MASK] consists of Mount Bulusan.
A: Luzon.
Q: [MASK] consists of Army National Guard.
A: National Guard.
Q: [MASK] consists of pharmacy.
A: biology.
Q: [MASK] consists of environmental factors.
A:
```

A.4.5 QA-style prompts

For QA-style prompts, we replace the instruction with "`Answer each question in one word.`" All other settings remain the same as in MP-style prompts. Below, we provide an example of QA-style/4-template prompts.

```
Answer each question in one word.
Q: Which entity does Panzer Division Kempf belong to?
A: Wehrmacht.
Q: Which entity does Mount Bulusan belong to?
A: Luzon.
Q: Which entity does Army National Guard belong to?
A: National Guard.
Q: Which entity does pharmacy belong to?
A: biology.
Q: Which entity does environmental factors belong to?
A:
```