UniArk: Improving Generalisation and Consistency for Factual Knowledge Extraction through Debiasing

Anonymous ACL submission

Abstract

In recent years, several works have investigated the potential of language models as knowledge bases as well as the existence of severe biases when extracting factual knowledge. In this work, we point out the inherent misalignment between pre-training and downstream tuning objectives in language models for probing knowledge under a probabilistic view and hypothesize that simultaneously debiasing these objectives can be the key to generalisation over unseen prompts. We propose an adapter-based framework **UniArk** for generalised and consistent factual knowledge extraction through simple and parameter-free methods. Extensive experiments show that UniArk can significantly improve the model's out-of-domain generalisation as well as being consistent under various prompts. Additionally, we construct a large-scale and diverse dataset **ParaTrex** for measuring the inconsistency and out-of-domain generation of models. Further, ParaTrex offers a reference method for constructing paraphrased datasets using large language models[1].

1 Introduction

Pre-trained Language Models (LMs) have been widely adopted in the NLP field. A key reason for the uptake of LMs is their capability to store knowledge in the parameters learned through pre-training (Liu et al., 2023a). Many works have looked at how to treat LMs as knowledge bases by measuring and extracting factual knowledge directly from them. LAMA (Petroni et al., 2019) is the first benchmark for measuring the extracted factual knowledge from LMs. In LAMA, factual knowledge is represented as triples (*subject, relation, object*) and is extracted through manually designed prompt templates. For example, to answer the query (*Barack Obama, place of birth, ?*), we query LMs using the prompt: "*Barack Obama was born in [MASK]*".

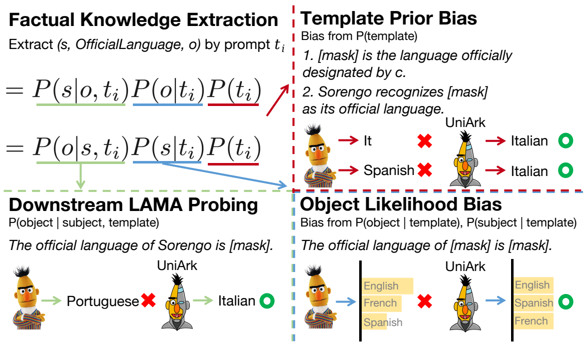

Figure 1: Illustration of the inherent objectives' bias from the template prior and template verbalization, with a comparison to our UniArk framework.

Many subsequent works have searched for optimal prompting strategies in order to improve the accuracy of the extraction (Shin et al., 2020; Li and Liang, 2021; Liu et al., 2023b; Li et al., 2022). However, they did not consider cases with different paraphrased prompt templates due to the limitation of LAMA, which only provides one prompt template for each relation. On the contrary, Elazar et al. (2021) and Newman et al. (2022) focused on the consistency between predictions from semantically similar prompts without optimizing for accuracy. In light of this, in this work we investigate how to improve both accuracy and consistency for unseen prompt templates, i.e. out-of-domain generalisation. We perform a probabilistic decomposition of the factual knowledge retrieval objective $P(subject, object|relation)$, cf. Fig. 1, and find a misalignment between the pre-training and tuning objects. This exposes two biases: $P(subject|template), P(object|template)$ (bias from object likelihood) and $P(template)$ (bias from template prior) as shown in Fig.1. Object likelihood bias refers to the likelihood of a predicted object given template-only prompts, such as

[1]Code and data will be released upon acceptance. ParaTrex datasets are submitted together with this paper.

"*The official language of [MASK] is [MASK]*", being biased. The biased object likelihood has been shown to positively correlate with the predictions from subject-given prompts and negatively influence the performance of factual extraction (Wang et al., 2023; Cao et al., 2021). Template prior bias is defined as the inconsistency among outputs from prompt paraphrases due to the domination of specific verbalizations during pre-training.

Therefore, we propose **UniArk**, a parameter-free unifying framework for optimizing both accuracy and consistency, through debiasing. The key idea behind each debiasing module is to equalize the probability distribution for the decomposed source bias term. To this end, we choose adapter tuning as our base tuning method, which is widely accepted as a modular parameter-efficient way of tuning and an effective way of debiasing (Kumar et al., 2023; Lauscher et al., 2021). However, to the best of our knowledge, we are the first to investigate adapter-tuning in factual knowledge probing tasks.

To evaluate the performance under unseen prompt templates, a paraphrased benchmark of the LAMA dataset is needed. We argue that the existing dataset ParaRel (Elazar et al., 2021) is both small in scale and not lexically diverse enough, as it is constructed based on rule-based methods such as swapping specific phrases. Therefore, we propose the dataset **ParaTrex** which is constructed using the large language model GPT-3.5. ParaTrex provides a more complex and substantially larger paraphrasing dataset. We provide both automatic evaluation and human evaluation statistics to show its high quality. Our main contributions are:

- We point out the misalignment between the pre-training and tuning objectives in a probabilistic view for factual probing, exposing the bias under a unified view as well as showing the possibility of improving generalisation via holistic debiasing.
- We construct ParaTrex, a comprehensive benchmark for out-of-domain generalisation measurements. We provide a thorough evaluation of ParaTrex.
- We propose a simple and parameter-free method based on an adapter-tuning framework for knowledge probing tasks. Extensive experiments show the effectiveness of our methods in improving the generalisation performance of knowledge probing and mitigating biases.

2 Objective Decomposition

We start with the objective for factual probing, showing that it is equivalent to the mask language modeling goals. We then decompose the probability representation of the task to show its misalignment with the tuning objectives, thus targeting two key components of the biased terms: the object likelihood and the template prior. We introduce several metrics for measuring these biased objectives.

Let $\mathcal{R} = \{r_1, r_2, \dots, r_{n_r}\}$, $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ and $\mathcal{O} = \{o_1, o_2, \dots, o_n\}$ respectively be sets of relations, subjects and objects. Given a relation r_j, *factual knowledge extraction* aims to extract factual knowledge triples (s_i, r_j, o_k) within LMs $\mathcal{M}$. Mathematically, we model $P(s_i, o_k|r_j)$ (the probability of subject-object pairs for a specific given relation). In practice, we query $\mathcal{M}$ with a manually designed prompt template t from the relation r_j. For instance, the template "*The capital of [X] is [Y]*" is constructed from the relation "*Capital*". Note that a specific relation can be mapped to different semantically similar prompt templates $\mathcal{T} = \{t_1, t_2, \dots, t_{n_t}\}$. We predict o_k through maximizing $P_{\mathcal{M}}(o_k|s_i, t_m)$. To position the inherent misalignment when modeling the object probability, we use the following probability decomposition of the task objective:

$$
\begin{aligned}
&P(s, o|r) && (1)\\
&= \sum_{t_i \in \mathcal{T}} P(s, o, t_i) && (2)\\
&= \sum_{t_i \in \mathcal{T}} P(s, o|t_i)P(t_i) && (3)\\
&= \sum_{t_i \in \mathcal{T}} P(s|o, t_i)P(o|t_i)P(t_i) && (4)\\
&= \sum_{t_i \in \mathcal{T}} P(o|s, t_i)P(s|t_i)P(t_i) && (5)
\end{aligned}
$$

Since $\mathcal{T}$ is defined as the set of templates relevant to the relation r, we can drop r in Eq. (2). We observe that the factual knowledge extraction goal $P(s, o|r)$ is equivalent to Eq. (2), which is approximated by the masked language modeling objective of LMs. After being decomposed, this objective function is influenced by five terms: $P(s|o, t_i)$, $P(o|s, t_i)$, $P(o|t_i)$, $P(s|t_i)$ and $P(t_i)$ (Eq. (4) and Eq. (5)). We note that sometimes we can rewrite object by subject since we might be interested in extracting the reversal relations, e.g. (*United Kingdom, capital, London*) and (*London, capital of, United Kingdom*). The subject and object might therefore be substitutable for different relations on the same text corpus. We therefore treat $P(s|o, t_i)$, $P(o|s, t_i)$ and $P(o|t_i)$, $P(s|t_i)$ as the same in the remaining context. The first two terms coincide with our tuning objectives but additional terms are exposed, indicating that the objectives between pre-training and downstream tuning are not aligned. We refer to these additional terms as *biased objectives*. $P(o|t_i)$, $P(s|t_i)$ show the bias from the object likelihood given a specific prompt template, and $P(t_i)$ gives an insight into the bias from the template prior.

2.1 Bias from the Object Likelihood

We define the *object likelihood* as $P(o|t)$. For $t_k \in \mathcal{T}$, we then define the bias from the object likelihood as $P(o_i|t_k) \neq P(o_j|t_k)$ for all $o_i, o_j \in \mathcal{O}$. That means that given only the prompt template without the subject, the object predicted by an LM is biased. This is also inline with the object bias defined in Wang et al. (2023). To measure this bias, we propose the *counterfactual hitting rate* (CT_hit1). This measures the accuracy of outputs from the prompt-only inputs, which should be close to 0 due to the lack of subjects. We measure the bias from object likelihood on 4 types of popular tuning methods. Table 1 shows the average CT_hit1 over all 41 relations in the LAMA dataset, where LAMA refers to do inference with the provided prompt in LAMA without tuning. Here we observe a clear increase in the hitting rate and entropy by comparing LAMA with other tuning methods, suggesting that after tuning, the model becomes stronger at guessing the correct answer from the likelihood of the object over the templates.

To show the influence of the object likelihood bias over the accuracy of the prediction, we also report the Pearson correlation coefficient (R) between the rank of grounding truth label over subject-given and subject-masked prompts over all samples in LAMA. In Table 1, we can observe a positive correlation between object likelihood and subject-given predictions. Moreover, greater positive correlations are observed for the wrong cases. This implies that some of the inaccurate predictions can be attributed to the bias from the object likelihood.

2.2 Bias from Template Prior

The bias from the *template prior* is defined as the inconsistency among different verbalizations with semantically similar prompt templates. Inconsistency problems have been widely discussed in previous works, e.g. (Elazar et al., 2021; Newman et al., 2022). This bias towards seen prompt templates $P(t_i)$ comes from unequal appearances of different prompts t_i during pre-training. This will influence the quality of factual probing since the appearance of a specific prompt t_i will weigh up $P(t_i)$, which results in learning better to predict $P(s, o|t_i)$ under this verbalization and neglecting other ones when being optimized. More importantly, this bias may be neglected in datasets such as LAMA where only one prompt template is used for tuning and testing. This motivates us to construct a more diverse and complex dataset for measuring the inconsistency as well as to propose a self-augmentation strategy aimed at averaging the biased template prior.

	CT_hit1	R	R (×)
LAMA	5.23	0.322	0.353
P-tuning	15.91	0.709	0.753
Adapter	12.77	0.341	0.376
Fine-tuning	13.11	0.228	0.284

Table 1: Counterfactual hitting rates for prompt-only inputs and correlations (R) between the rank from outputs with and without given subject among all predictions and incorrect predictions.

	ParaRel	ParaTrex
# Relations	39	40
# Patterns	329	1526
Min # patterns per rel.	1	26
Max # patterns per rel.	20	47
Avg # patterns per rel.	8.23	38.15
Avg lexical per rel	5.73	8.46

Table 2: Statistics of the ParaRel and ParaTrex datasets.

3 The ParaTrex Resource

We introduce the **ParaTrex** resource, which is a large-scale and comprehensive paraphrasing dataset used for measuring both inconsistency and the generalisation capability of models on different unseen inputs. ParaTrex comprises 1526 paraphrases from 40 relations[2], with an average of 38.15 templates per relation. The statistics of the dataset are provided in Table 2, with comparison to the ParaRel dataset (Elazar et al., 2021).

Data Construction We construct ParaTrex, a paraphrased version of the LAMA dataset, using the following steps: (1) We begin with the patterns provided by LAMA. Each relation has one prompt

[2]Like ParaRel (Elazar et al., 2021), we omit one relation hard for generating paraphrases: "[X] is a [Y]"

template called *base-pattern*. For example, the base pattern of relation "*capital of*" is *"[X] is the capital of [Y]."* (2) For each relation, to make the generation more specific, we extract its base pattern and its corresponding Wikidata (Vrandečić and Krötzsch, 2014) provided in the LAMA description. For instance, for the relation *CapitalOf*, "*country, state, department, canton or other administrative division of which the municipality is the governmental seat*". (3) We formulate a manually crafted prompt directing GPT-3.5-turbo to produce a total of 40 paraphrases. This includes 5 succinct paraphrases, each comprising no more than 7 words, as well as 5 extended paraphrases, each encompassing more than 15 words. More details of the paraphrase generation process can be found in Appendix A.1. (4) Through human inspection, we remove inappropriate paraphrases characterized by excessive ambiguity or similarity to preceding generations. (5) We iteratively execute Steps (3) and (4) until satisfying answers are achieved. We have at least 25 paraphrases: 5 short, 5 long, with the rest being medium length. Furthermore, we introduce a random division of our paraphrases into two distinct sets: a training set comprising 50% of the entire dataset, and a test set constituting the remaining 50%. The out-of-domain set encompasses all long and short paraphrases, aiming at simulating the situation where individuals seek to extract specific knowledge by inputting a concise or exceptionally long query. We provide an example in Appendix A.2.

Evaluation We evaluate the quality of ParaTrex using two automatic metrics and human evaluation. A detailed description of the evaluation can be found in Appendix A.3. We next discuss the most salient points. We measure the diversity of the paraphrases through the average pairwise BLEU scores (Papineni et al., 2002) of paraphrases among each relation. The results show that the 1-4 gram BLEU scores of ParaTrex are consistently lower than those of ParaRel, suggesting that ParaTrex datasets are lexically and syntactically more diverse. To evaluate the quality, we report the cosine-similarity between the paraphrase and the raw template using a paraphrase version of sentence-bert (Reimers and Gurevych, 2019). We observe a clear difference between the randomly chosen paraphrase and the generated paraphrase, proving that the quality of paraphrasing is acceptable. Human evaluation from NLP graduate students for ParaTrex also shows a 96.88% precision and 92% recall respectively, indicating the high quality of ParaTrex datasets.

4 Methodology

Based on the probability decomposition in Section 2, we hypothesize that mitigating the misalignment between the tuning and pre-training objectives is the key to improving both the accuracy and consistency of models on unseen prompts. To this end, the core idea behind UniArk is to equalize the probability of biased parts through an additional loss and template augmentation. We discuss below the three main components of **UniArk**.

Adapters We use adapter-tuning (Houlsby et al., 2019) as it is better suited for debiasing settings (Kumar et al., 2023) and internal knowledge protections than other popular parameter-efficient fine-tuning methods. Moreover, we want to evaluate and thus fill in the vacancy of adapter-tuning on the factual knowledge extraction tasks. Note that for factual probing, it is common to tune a model for each relation. Due to the cost of storage when the relations scale up, we therefore do not choose full parameter fine-tuning as the basis of our framework. The basic idea is to insert an adapter into our base language models and freeze all other parameters. Specifically, for each output $\mathbf{h}^n \in \mathbb{R}^d$ in the n-th transformer layer, our adapters perform the following transformation:

$$\mathbf{h}^{n+1} = \text{GELU}(\mathbf{h}^n \mathbf{W}_d)\mathbf{W}_u + \mathbf{h} \qquad (6)$$

where GELU (Hendrycks and Gimpel, 2016) is a non-linear activate function, $\mathbf{W}_d \in \mathbb{R}^{d\times k}$ and $\mathbf{W}_u \in \mathbb{R}^{k\times d}$ are two learnable parameter matrices in adapters. They are used for first down-projecting the hidden states into dimension $k < d$, and then projecting them back to d-dimension spaces, with k a hyperparameter.

Object likelihood Bias Mitigation As discussed in Section 2.1, to mitigate the object likelihood bias, the output distribution should ideally satisfy: for all $o_i, o_j \in \mathcal{O}, s_i, s_j \in \mathcal{S}$ and $t_k \in \mathcal{T}$, we have that $P(o_i|t_k) = P(o_j|t_k), P(s_i|t_k) = P(s_j|t_k)$. In other words, the retrieved likelihood distribution should be close to a uniform distribution from the subject-masked and object-masked inputs. To this end, we introduce an addition max entropy loss L_{me} weighted by hyperparameter λ_{me} over subject-masked prompts and object-masked prompts. This loss maximizes the entropy over top retrieved candidates to encourage the model to assign equal probability to each relevant candidate. We perform an object filtering process to remove stopwords like "*and*". We choose to max the entropy of only the top k words because, based on our empirical observation, they include most of the relevant candidates. Formally, given the output probability of object $i : p(i), i = 1, 2, \ldots, k$ and the stopwords set S, the max entropy loss is:

$$\mathcal{L}_{me} = -\lambda_{me} \sum_{i=1,\ i \notin S}^{k} p(i) \log_2(p(i)) \quad (7)$$

We note that unlike MeCoD (Wang et al., 2023), our method does not bring any additional parameters and focuses on equalizing the likelihood for all potential candidates while MeCoD performs neural object selecting and does contrastive learning over the selected objects. This suggests that our method is lighter than MeCoD. We also generalise MeCoD since we consider both subject-masked and object-masked prompts, guided by our objective decompositions.

Template prior Bias Mitigation To alleviate the template prior bias, we propose a novel self-data augmentation method to mitigate the influence of $P(t_i)$ by weighted averaging them. We augment our raw data with prefixes "*It is true that*" and "*It is false that*" and encourage the model's self-consistency by weighted averaging their output distribution to make final predictions. Specifically, the output probability $P(o_i|s,t)$ for object candidate i and the masked language model (MLM) loss L_{mlm} are calculated as:

$$P(o_i|s,t) = softmax\Big(\sum_{t_j \in \mathcal{T}^*} w_j P(o_i|s,t_j)\Big) \quad (8)$$

$$\mathcal{L}_{mlm} = -\sum_{i=1}^{n_{vocab}} y_i \log P(o_i|s,t) \quad (9)$$

where $\mathcal{T}^* = \{t, t_{\text{true}}, t_{\text{false}}\}$ is the set of augmented prompt templates and the weight $\sum_j w_j = 1$ is a hyperparameter balancing the weight for each template. Note that we set $w_{\text{true}} = -w_{\text{false}}$ since the prompts "*It is true that*" and "*It is false that*" give opposite predictions.

5 Experiments

Dataset We use LAMA-TREx (Petroni et al., 2019) as our main training dataset, with the same train-test splits as in (Liu et al., 2023b). This dataset comprises 41 relations and 29,500 testing triples. To test the generalising ability and consistency for different prompt templates, we test the model on two additional paraphrased datasets: our ParaTrex and ParaRel (Elazar et al., 2021). Since in both datasets N-M relations are omitted when measuring consistency, because it can be hard to measure consistency among several correct answers, 25 relations remained after filtering those.

Evaluation Metrics We evaluate the performance of models on three aspects: quality of extraction, object likelihood bias, and template prior bias. (1) For measuring the quality, we evaluate the macro F1 score for each relation over LAMA (LM), ParaTrex (PT), and ParaRel (PR) to test its performance in in-domain settings and generalisation on out-of-domain prompt templates. (2) To test the bias from the object likelihood, we report the hitting rate of the candidates from the counterfactual subject-masked prompt (CT_hit1). Additionally, we report the KL-divergence (KLD) between the subject-masked prompt and the original prompt to show the influence of the prompt template on the likelihood distribution of the final retrieved candidates. (3) For the template prior bias, we measure the consistency of paraphrases in both ParaTrex and ParaRel. Following Elazar et al. (2021) and Newman et al. (2022), the *consistency* is calculated as the ratio of consistent predictions from different paraphrases with all the paraphrases permutations. We also measure consistency between the unique raw prompt template from LAMA and the paraphrased templates. We refer to this consistency as *raw_cst* while consistency between all permutations as *all_cst*. The previous consistency measures do not consider strict factual accuracy. Thus, we also measure the consistency over factual correct predictions, called *acc_cst*. Formal definitions of *raw_cst*, *all_cst* and *acc_cst* are in App. B.1.

Baselines We split our experiments into two settings: soft and manual prompts. In the soft prompt setting, we choose P-tuning (Liu et al., 2023b), which is a popular prompt-tuning method in knowledge probing tasks and the SoTA MeCoD (Wang et al., 2023) as baselines. We compare them with the adapter tuning to explore its performance. Note that we cannot measure the consistency over paraphrases here since the whole prompt template is learned through training. For the manual prompt setting, we take the manual prompt without tuning (LAMA) and adapter tuning as baselines. Additionally, we re-implement MeCoD as MeCoD (OI) through adapter tuning as it is originally based on P-tuning. App. B.2 provides more training details.

Significance Test To test the significance of any improvements or deterioration, we perform the following tests between our UniArk and the adapters baseline: (1) Paired T-Test and Wilcoxon Sign Test for a fixed seed among results across all relations and (2) T-test among the averaged values of all relations after running UniArk with three different seeds. See detailed results in the Appendix B.3.

5.1 Quantitative Results

Table 3 presents results for knowledge retrieval quality together with object likelihood bias on BERT-large (Devlin et al., 2019) and RoBERTa-large (Liu et al., 2019). Table 4 shows results for template prior bias. The best value is marked in bold and the second best value is marked in italics.

Main Results For probing quality, we find that with the appropriate tuning methods, models with manual prompts outperform those with soft prompting. This shows the necessity of tuning parameters within the models rather than within the input embeddings. Among all vanilla tuning methods, Adapters demonstrate a remarkable capability for in-domain knowledge and object likelihood bias. They outperform fine-tuning over 0.01 (4%) on the in-domain F1-score, with also less object likelihood bias than P-tuning and fine-tuning. However, it is still shown to be under severe biases and performs poorly on the out-of-domain prompts. With our proposed framework UniArk for mitigating both biased objectives, we significantly improve the generalisation ability to probe knowledge on the unseen prompts. Various significance tests prove the improvements in the out-of-domain generalisations and two bias mitigations over adapters and MeCoD baselines. The in-domain quality is also shown not harmed. Indeed, UniArk outperforms the current SoTA MeCoD in both in-domain and out-of-domain prompt templates.

Adapters versus Other Tuning Methods To better understand the capabilities of the adapter-tuning method on factual knowledge extraction, we compare it with manual prompts (LAMA), P-tuning (PT), and fine-tuning (FT). We do not consider other parameter-efficient fine-tuning methods, such as prefix-tuning (Li and Liang, 2021), since they are shown to be less powerful than P-tuning (Liu et al., 2023b; Wang et al., 2023). Table 3 shows that the adapter-tuning performs consistently better than all other parameter-efficient fine-tuning methods in the F1 score when tuning on the in-domain settings. This strongly suggests that tuning methods such as adapters, which modify the inner transformer layers instead of only embedding layers without changing the initial parameters, may do better in extracting the knowledge hard encoded within the parameters in LMs. However, there exists a substantial difference in performance between in-domain and out-of-domain settings. Indeed, we observe a big gap in F1 scores, suggesting that those parameter-efficient tuning methods tend to be biased on the given prompt template.

Bias Mitigation and Quality Improvements In Table 3, we observe that with our proposed framework UniArk, both object likelihood bias and prompt prior bias are effectively mitigated. The counterfactual hitting rate drops to nearly 0. This means the model can no longer guess the correct answers given only templates. The sharp rise of KL-divergence also indicates that the model tends to predict a distribution diverging substantially from the object likelihood under prompt templates. Both metrics show that the model no longer suffers from being influenced by the object likelihood. Additionally, in Table 4, the consistency over all paraphrased datasets increases significantly, showing the effectiveness of our prior bias mitigation module. At the same time, we can respectively observe improvements of 7% (22.12 to 23.68), 4% (23.78 to 24.7), and 13% (24.69 to 27.99), 4% (27.34 to 28.48) of out-of-domain F1 score in UniArk compared with the adapters baseline for RoBERTa and BERT on ParaTrex and ParaRel. This validates our hypothesis that mitigating the two decomposed bias terms helps generalisation to unseen prompts. Besides, we also provide a scaling study in App. B.4, where we show that UniArk has significant improvement on both base and larger models.

5.2 Ablation Studies

We take adapter-tuning as a baseline and perform ablation studies to clarify the source of performance improvement. The results in Table 5 demonstrate that our max entropy (ME) module plays a prominent role in relieving object likelihood bias while our self-augmenting (Aug) module makes the main contribution to mitigating prompt preference bias. Both modules increase the F1 scores of extraction quality, showing the help of bias mitigation for improving the out-of-domain generalisation.

Method	BERT-Large					RoBERTa-Large				
	OOD		ID	OL Bias		OOD		ID	OL Bias	
	PT_F1	PR_F1	LM_F1	CT_hit1	KLD	PT_F1	PR_F1	LM_F1	CT_hit1	KLD
P-tuning			29.94	15.91	3.34			19.36	17.13	2.06
+MeCoD	-		29.33	1.02	*8.48*	-		23.13	5.67	5.39
+Adapters			31.21	14.00	3.40			27.70	14.72	3.47
LAMA	14.21	16.00	20.68	4.19	3.57	8.34	9.19	12.37	5.23	1.83
Adapters	24.69	27.34	*32.10*	12.77	5.54	*22.12*	*23.78*	**29.74**	16.88	3.40
+MeCoD (OI)	*25.64*	*27.58*	31.79	*0.13*	7.31	21.97	23.34	28.72	*5.00*	*6.13*
+UniArk	**27.99**	**28.48**	**32.14**	**0.04**	**11.66**	**23.68**	**24.70**	*29.29*	**3.65**	**10.24**
Fine-tune	28.50	29.27	30.85	13.11	8.07	25.05	25.53	27.85	12.23	6.11

Table 3: Main results for out-of-main (OOD), in-domain (ID) performance, and object likelihood bias (OL Bias) on LAMA (averaged over all relations). The underlines represent the significance after three significance tests.

Model	Method	ParaTrex			ParaRel		
		raw	all	acc	raw	all	acc
Roberta -large	LAMA	23.9	20.6	6.9	33.0	28.3	10.4
	Adapters	61.9	55.2	34.1	66.9	60.4	37.3
	+ MeCoD (OI)	61.7	54.8	34.6	67.9	61.2	38.1
	+ UniArk	**63.8**	**59.0**	**36.2**	**69.1**	**63.4**	**38.5**
BERT -large	LAMA	33.6	28.3	15.8	54.9	46.6	25.0
	Adapters	60.9	53.4	39.1	72.1	65.2	45.8
	+ MeCoD (OI)	63.4	56.5	41.2	73.5	67.3	47.2
	+ UniArk	**69.1**	**62.9**	**44.7**	**76.7**	**71.3**	**49.4**

Table 4: Main results for template prior bias (TP bias) measured by consistency on ParaTrex and ParaRel. Significantly improved results are underlined.

Method	Quality		OL Bias		TP Bias	
	PT	PR	CT_hit1	KLD	PT	PR
UniArk	**28.0**	**28.5**	**0.0**	11.7	**62.9**	**71.3**
w/o ME	26.9	28.4	13.2	5.5	60.8	70.5
w/o Aug	25.3	27.3	**0.0**	**12.3**	56.0	66.3
w/o ME & Aug	24.7	27.3	16.9	3.4	55.2	60.4

Table 5: Ablation study on BERT, we report F1 score for extraction quality; and all_consistency for template prior bias on ParaTrex (PT) and ParaRel (PR)

We emphasize that our ME module contributes to improving consistency and our Aug module brings an improvement on the prompt preference bias as well. This exhibits a synergizing effect of both modules on mitigating both biases, further highlighting the necessity of simultaneously alleviating biases within a unified framework. This effect is probably because, as we equalize the object likelihood over templates, the model is forced to treat the prompt templates as the same, which also weakens the favor of specific templates and thus increases the consistency over unseen prompts. Meanwhile, augmenting the templates forces the model to estimate the object likelihood over various cases, and averaging this likelihood distribution contributes to a more unbiased object likelihood.

5.3 Qualitative Case Studies

To better understand how mitigating the studied biases helps to improve the knowledge extraction results, we perform two specific case studies on randomly selected cases. A detailed analysis can be found in App.B.5. Here we give one example from each biased objective mitigation. For template prior bias (Table 9), although both UniArk and adapter-tuning make a correct prediction "*Finnish*" on the question "*The official language of Vesanto is [mask]*", the answers of adapters may turn to some pronoun such as "*It*" when the templates changed. The UniArk relieves these kinds of errors with the augmented inputs and drops the predictions of predictions for "It" from 861 (7.4%) times to 140 (1.2%) times among all predictions in this relation according to our statistics. For object likelihood bias (Table 10), when it comes to the question "*The official language of Sorengo is [mask]*", the golden truth should be "Italian". However, traditional probing gives "*Portuguese*" as the answer and we found that the rank 2, and rank 3 predictions "*English*" and "*Spanish*" appears in the prediction from the top and third predictions from subject-masked prompt, suggesting that the prediction of a traditional model may be influenced by this object likelihood. In contrast, UniArk, who provides the correct answers, is not influenced by this object "*English*" since the subject-masked likelihood is uniformly distributed.

6 Further Analysis

Using Paraphrased Data for Training To simulate real applications in which paraphrased data is lacking (and for a fair comparison), UniArk is tuned on a single prompt template provided in the LAMA dataset. We try to investigate the following question: What if we use the part of paraphrased data for training? We added a new module called "PARA" following (Elazar et al., 2021), where an additional KL-Divergence loss between the prediction distribution from the LAMA template and the paraphrased template is added. We randomly select 1, 2, and 5 new paraphrased templates to perform experiments. From Table 6, only a subtle improvement can be witnessed after adding new paraphrases to UniArk for training and these improvements also do not scale up with more given paraphrases. This indicates that our proposed self-data augmentation, where no additional paraphrases are provided, is as powerful as using paraphrases to improve generalisability under current frameworks. This result also suggests a potential research direction for incorporating paraphrased data both efficiently and effectively during training.

Method	Quality		OL Bias		TP Bias	
	PT	PR	CT_hit1	KLD	PT	PR
UniArk	28.0	28.5	0.0	11.7	62.9	71.3
+para 1	28.1	28.6	0.0	11.6	63.3	71.8
+para 2	28.3	28.9	0.0	11.5	63.3	71.9
+para 5	28.1	28.6	0.0	11.6	63.2	71.8

Table 6: Results on extraction quality f1, object likelihood bias, and template prior bias consistency using paraphrased data for training

Error Analysis To have a comprehensive understanding of the existing errors in our factual probing framework, we conducted a random sampling of 50 incorrect predictions within the relation P37 "*Official_Languages*" We categorized these errors, documenting the findings in Appendix B.6. In summary, we find that LMs still do not have a comprehensive knowledge of specific cities such as Azad Kashmir. They also make mistakes in predicting pronouns like "*It*" (4 cases), and in spelling (2 cases). Besides, we found 21 (42%) cases where the model makes a feasible answer among several correct answers but is treated wrong because only one of the labels is provided, e.g. Finnish for Turku, suggesting that we may underestimate the knowledge stored in LMs via current metrics.

7 Related Work

Factual Knowledge Extraction There are several works on how to treat LMs as knowledge bases and extract factual knowledge from the weights of an LM. Petroni et al. (2019) is one of the seminal works on this and also introduces the LAMA benchmark for extracting factual knowledge from LMs. To access the knowledge, Li et al. (2022) applies further pre-training (fine-tuning) on LMs. Liu et al. (2023a) suggests that manual prompts offer a promising avenue for directly accessing this knowledge without the need for extra fine-tuning. To search for an optimal prompt, AutoPrompt (Shin et al., 2020) automatically creates a prompt using gradient-based search. Recent works look at soft prompts with continuous learnable prompts. Liu et al. (2023b) proposes P-tuning, making all tokens within prompt templates as learnable soft prompts and showing similar scaling results on larger language models. However, we observe that adapter tuning (Houlsby et al., 2019) has not been applied to this task so far. In this paper, we compare our results within both soft prompt and manual prompt settings, showing that adapter tuning is a promising and robust way of factual knowledge extraction.

Bias study Cao et al. (2022) and Elazar et al. (2021) argue that there exist severe risks and biases under prompt-based knowledge extraction. Therefore, Newman et al. (2022) intends to increase the consistency through asserting a single multiple-layer perception after embedding layers called p-adapters. Wang et al. (2023) propose the contrastive-learning-based framework MeCoD for mitigating the bias. In this paper, we position and decompose the object likelihood bias and template prior bias under a probabilistic view and propose a unified framework for mitigating them, which is a more general case compared with previous studies.

8 Conclusion

In this paper, we revisit the factual probing objectives under a probabilistic view and point out the misalignment between the pre-training and fine-tuning objectives. This motivates our hypothesis that mitigating both template prior and object likelihood bias may improve the generalisability of knowledge-probing models. We introduce ParaTrex, a large and high-quality dataset for measuring the generalisability, and propose a parameter-free method to validate this hypothesis. Experiments show the superiority of our framework and a synergizing effect is found in our modules for alleviating both biases, proving the necessity of a unified framework towards a generalised factual knowledge extraction.

Limitations

We identify the following two limitations related to the methodology and base models. First, in our verbalization bias mitigating module, we perform a naive average between the self-augmenting inputs and the original inputs, following our objective decomposition parts. Although it works effectively, it would be interesting to investigate other methods. Second, the prompt template in LAMA and ParaTrex/ParaRel datasets is designed for masked language modeling instead of next token prediction. We made a scaling study on encoder-only models to show the scalability of our methods, it would be interesting to also construct corresponding datasets for decoder-only large language models, such as Llama2 (Touvron et al., 2023) and perform experiments on them. We leave this for future work.

Ethics Statement

During the construction of the paraphrased dataset ParaTrex, we did not generate any data that is harmful to society and humans, nor include any private personal information within the dataset.

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.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. 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.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023b. Gpt understands, too. *AI Open*.

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

Benjamin Newman, Prafulla Kumar Choubey, and Nazneen Rajani. 2022. P-adapters: Robustly extracting factual information from language models with diverse prompts. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 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 Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

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.

A ParaTrex Details

A.1 ParaTrex: Construction Workflow

Figure 2 provides an illustration of the workflow to generate the ParaTrex datasets using large language models.

A.2 ParaTrex: Exemplary Templates

Table 7 provides a full example of the generated templates in ParaTrex for the relation "P1376": "*Capital_of*".

A.3 ParaTrex: Evaluation Details

We evaluate the quality of ParaTrex using two automatic metrics and human evaluation.

Diversity We test the lexical diversity by reporting the average pairwise BLEU scores of each relation. Specifically, all pair-wise permutations of n templates for each relation are listed, resulting in $n(n-1)$ sentence pairs. Then pair-wise n-gram BLEU score (Papineni et al., 2002) was calculated on these pairs to represent their diversity. The average score of the lower-order n-gram score captures lexical diversity and the average score of the higher-order n-gram score tends to capture the diversity of complex syntactic structures. Fig 3 shows the trend over n-gram average pairwise BLEU scores of all relations. We find that the BLEU scores of ParaTrex perform consistently lower than ParaRel, which depicts that our proposed dataset has a good lexical and syntactical diversity of generated sentences compared with the existing baseline datasets.

Quality For automatic evaluation, we perform use the current SoTA version *paraphrase-multilingual-mpnet-base-v2* of Sentence-BERT (Reimers and Gurevych, 2019) on the Sentance-BERT leaderboard[3] to evaluate the semantic similarity between the paraphrase and the grounding prompt template provided in the LAMA dataset. We report the average cosine similarity upon all paraphrases for each relation in our dataset and show it in a boxplot (Fig 4). These results show that ParaTrex shares good semantic alignments with the grounding datasets except for two special cases. There are two relations getting scores lower than 0.7. This is because the grounding templates "*[X] plays [Y]*" and "*[X] is located in [Y]*" are missing the information that [Y] refers to musical instruments and continents respectively. This information is included in the description of the dataset, which is also taken into consideration when constructing ParaTrex.

Human Agreement Following Elazar et al. (2021), we randomly picked 82 paraphrases in the ParaTrex dataset and 42 wrong paraphrases by sampling from the paraphrases of wrong relations. We perform human evaluation by asking the evaluators to select candidates that are not the paraphrase of the given inputs. The participants need to pick out the wrong paraphrases. We consider the remaining answers as what they think to be the correct paraphrases of the given inputs. Two examples of questions are shown in Fig 6. Results show that on average among 11 human judgments, human evaluators get 96.88% accuracy in successfully identifying inaccurate paraphrases and a 92% accuracy in selecting the true paraphrases provided by ParaTrex, which shows that our proposed datasets have a satisfying agreement with human beings, thus proving the favorable quality of our datasets.

B Experiments details and further study

B.1 Formal Definitions of Consistency

The *consistency* is calculated as the ratio of consistent predictions from different paraphrases with all the paraphrases permutations Elazar et al. (2021); Newman et al. (2022). Formally, given a set of unordered paraphrase pairs P_i of relation r_i, consisting of n distinct prompts, we have a total of $\frac{1}{2}n(n-1)$ number of permutations. For the j-th sample in the i-th relation, we define the consistency between all paraphrases as:

$$\text{Consistency}_j = \frac{\sum_{p_m,p_n \in P_i} \mathbb{I}[\hat{e}^m_{ij} = \hat{e}^n_{ij}]}{\frac{1}{2}n(n-1)} \tag{10}$$

where $\mathbb{I}$ is the indicator function, $\hat{e}^m_{ij}$ and $\hat{e}^n_{ij}$ refer to the predicted entity by PLMs from prompt p_m and p_n, respectively.

We now give the formal definitions of *raw-consistency* and *all-consistency*. For the reason of simplicity, we consider the combination of the unique raw prompt template from LAMA, and templates from paraphrased LAMA $p_m \in P_i$, getting n combinations in total. The consistency between raw prompts and paraphrased prompts (**Raw-**

[3]https://www.sbert.net/docs/pre-trained_models.html

Templates	inhouse split	paraphrase type
The capital of [Y] is [X] .	test	short paraphrase
[X] is [Y]'s capital .	test	short paraphrase
[X] serves as [Y]'s capital .	test	short paraphrase
[Y]'s capital city is [X] .	test	short paraphrase
[X] acts as [Y]'s capital .	test	short paraphrase
[X] is the administrative division where the municipality of [Y] serves as the capital .	test	long paraphrase
The governmental seat of [Y] is located in [X], which is the capital city .	test	long paraphrase
[X] holds the status of being the capital city and administrative center of [Y] .	test	long paraphrase
The capital of [Y] is none other than [X], where the government operates .	test	long paraphrase
The administrative hub of [Y] is [X], which holds the position of being the capital cit .	test	long paraphrase
[X] is the official capital of [Y] .	test	normal paraphrase
The capital city of [Y] goes by the name of [X] .	test	normal paraphrase
[X] is the designated capital city of [Y] .	test	normal paraphrase
[X] serves as the principal capital city of [Y] .	test	normal paraphrase
[X] is the administrative capital and governmental seat of [Y] .	test	normal paraphrase
[X] is the principal administrative center of [Y] .	test	normal paraphrase
[X] serves as the capital city and governmental hub of [Y] .	test	normal paraphrase
[X] holds the official status of being [Y]'s capital city .	test	normal paraphrase
[X] acts as the administrative capital of [Y] .	test	normal paraphrase
[X] serves as the capital city of [Y] .	test	normal paraphrase
[X] is the primary governing capital and administrative center of [Y] .	test	normal paraphrase
[X] is the primary political center of [Y] .	test	normal paraphrase
[X] holds the title of being [Y]'s capital .	test	normal paraphrase
[X] serves as the seat of government for [Y] .	test	normal paraphrase
[X] is the city that serves as [Y]'s capital .	test	normal paraphrase
The government of [Y] is headquartered in [X], its capital .	test	normal paraphrase
[X] acts as the political center of [Y] .	test	normal paraphrase
[X] holds the official position of being [Y]'s capital .	train	normal paraphrase
[X] serves as the governing center of [Y] .	train	normal paraphrase
The capital city of [Y] is [X] .	train	normal paraphrase
[X] is the administrative center of [Y] .	train	normal paraphrase
The seat of administration in [Y] is [X] .	train	normal paraphrase
The designated capital city of [Y] is [X] .	train	normal paraphrase
The governmental headquarters of [Y] is located in [X] .	train	normal paraphrase
[X] holds the status of being [Y]'s capital .	train	normal paraphrase
The government of [Y] is headquartered in [X] .	train	normal paraphrase
[X] is where the governing body of [Y] is located .	train	normal paraphrase
[X] holds the position of being [Y]'s capital city .	train	normal paraphrase
[X] holds the official governmental seat and capital status of [Y] .	train	normal paraphrase
[X] serves as the governing capital of [Y] .	train	normal paraphrase
The capital city of [Y] is none other than [X] .	train	normal paraphrase
The political center of [Y] is [X] .	train	normal paraphrase
The administrative capital of [Y] is [X] .	train	normal paraphrase
The government headquarters of [Y] can be found in [X] .	train	normal paraphrase
[X] is where the government of [Y] is based .	train	normal paraphrase

Table 7: Example for the relation "*Capital_of*" in ParaTrex. The original prompt template in LAMA is "*[X] is the capital of [Y] .*"

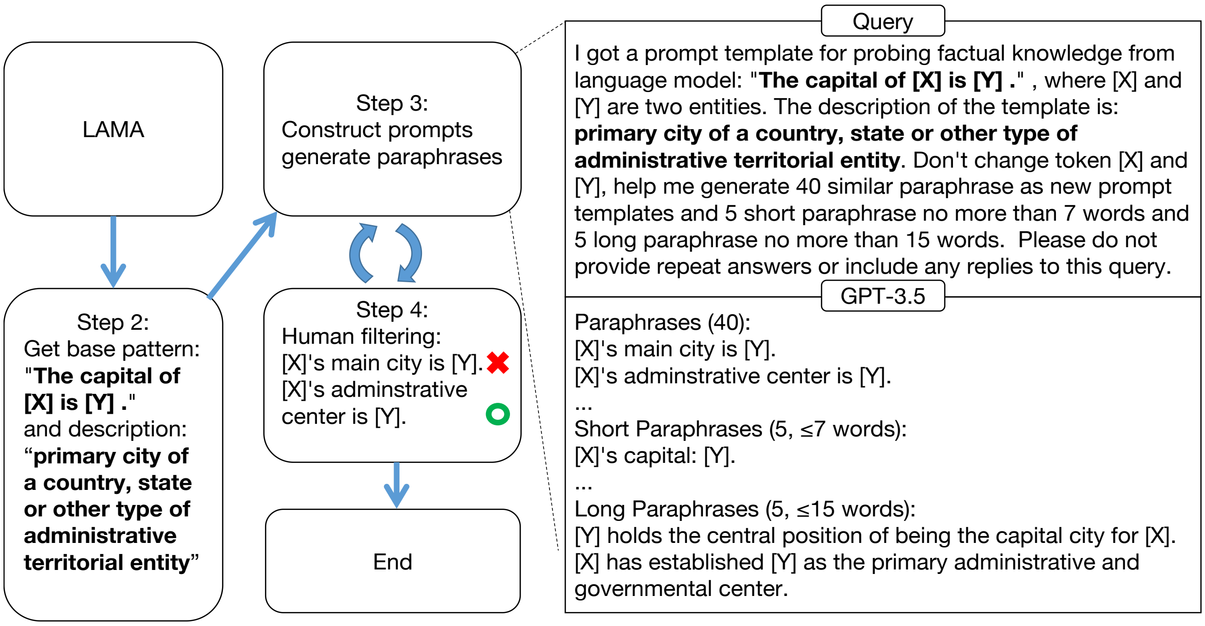

Figure 2: Workflow to generate a paraphrased version of prompt templates in ParaTrex. We exemplify it for the relation 'capital of' in LAMA.

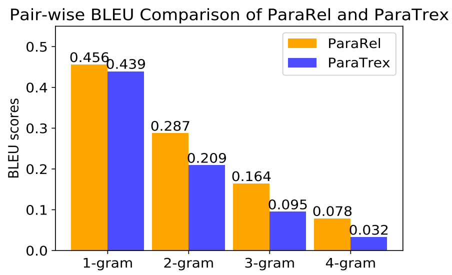

Figure 3: Average pair-wise BLEU between all relations comparison with ParaRel. ParaTrex gets a consistently lower score than ParaRel, representing that the templates in ParaTrex are more lexically and syntactically diverse.

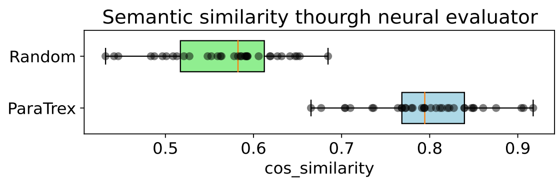

Figure 4: The cosine similarity of the embedding between the grounding template and the paraphrased template. The boxplot shows the comparison between the random paraphrase sampled from other relations and the paraphrase in our dataset for 39 relations.

Consistency) will be degraded to:

$$\text{Raw-Csty}_j = \frac{\sum_{p_m \in P_i, p} \mathbb{I}[\hat{e}_{ij} = \hat{e}^m_{ij}]}{n} \quad (11)$$

Besides, the previous consistency measures only look at the matches between predictions and do not consider strict factual accuracy. However, factual correctness remains a crucial attribute for KBs. Thus, we additionally measure the consistency over factual correct predictions:

$$\text{Acc-Csty}_j = \frac{\sum_{p_m, p_n \in P_i} \mathbb{I}[\hat{e}^m_{ij} = \hat{e}^n_{ij} = e_{ij}]}{\frac{1}{2} n(n-1)}$$

, where e_{ij} is the ground truth entity.

B.2 Training Details

We perform all experiments based on BERT-large and RoBERTa-large on the RTX 2080Ti GPUs, which run for about 1 hour to train on one relation. We set the hyperparameter $\lambda_{me}, \lambda_{kld}$ to be 0.2. w_{true} and w_{false} are set to be simply -1 and 1. For adapters, we take the hidden state to be 256 dimensions. All other hyperparameters (including the random seed) are set as default in (Liu et al., 2023b).

B.3 Significance Test Details

We perform the Paired sample T-test and the Wilcoxon Signed-Ranked test on the results from all 25 relations between adapters and our UniArk to test the significance after performing UniArk. We also apply different seeds (20, 30, 50) and perform a t-test among the average results to test whether the results are significant for different runs. The results of the p-values are shown in Table 8, where cst refers to the consistency, pt, pr, and lm refer to the ParaTrex, ParaRel, and LAMA datasets respectively.

Overall, we can observe that the p-values of all consistency and out-of-domain f1 scores are smaller than 2.5e-2, strongly suggesting that UniArk makes significant improvements over the baseline adapters both with the normally distributed assumption or not. On the contrary, all results in the in-domain f1 scores are bigger than 5e-2, indicating the non-significance of the decrease/increase in in-domain quality. This proves that UniArk makes significant improvements over the out-of-domain generation and both biases while maintaining its performance in the in-domain settings.

B.4 Scaling Study

We want to answer the question of whether the results of UniArk are scalable for models with more parameters. Figure 5 presents comparison results of F1 score, counterfactual accuracy and consistency between BERT-base, BERT-large, RoBERTa-base and RoBERTa large. The results demonstrate that UniArk performs consistently better for both extraction performance and inherent bias. We also observe consistently better results for larger models among all settings. We therefore conclude that (1) The performance for extracting knowledge and bias can be scaled by the size of LMs. (2) The bias mitigation and performance boost from the UniArk framework can also be observed among all sizes of models (3) For bias mitigation, small models are able to be more unbiased and robust through the UniArk framework.

B.5 Details for Qualitative Study

We perform two specific case studies to better understand how mitigating the studied biases helps to improve the knowledge extraction results. Firstly, in Table 10 we present cases showcasing how the models make the incorrect prediction due to the biased object likelihood. PLMs are asked for the official language of a specific item using the prompt:"*The official language of [sub] is [obj].*". The last row shows the results for the vanilla LMs without being tuned and thus suffering from high object likelihood such as *English* and *Spanish*. The logits of objects *English* and *Spanish* of LAMA methods are close, showing that the model is not confident with its predictions and may guess from the object likelihood from templates. The SoTA model MeCoD still gives the wrong answer since they apply an unreliable neural gate to automatically classify which object to be debiased. For instance, MeCoD successfully smooths the high counterfactual logit for the word *English* but causes the model to underfit this object so that it cannot recall the correct object Italian and thus make an incorrect prediction with a high logit. In contrast, UniArk is capable of making accurate predictions with higher logits while having an unbiased prediction distribution under subject-masked inputs, showing that UniArk provides more confident answers without the impact of the prior distribution from prompt templates.

Table 9 presents an example of the consistency study. We provide an instance where adapter-tuning and UniArk are both correct on the original prompts. We randomly sample several paraphrased cases from ParaTrex. The results suggest that the baseline fails to produce correct answers when meeting syntactically and lexically diverse prompt templates. The second and fourth rows of paraphrased prompt templates are examples for the different syntic variants while the first and the last rows of paraphrased templates show more lexically complicated prompts. Our UniArk model gives mostly consistent outputs in those cases, although it may make some mistakes. Additionally, we can observe from the results that UniArk maintains a robust behaviour on outputting language objects instead of stopwords like "*it*". This shows that the UniArk models are more robust on various prompt templates after debiasing.

B.6 Details for the Error Analysis

To have a comprehensive understanding of what kinds of errors UniArk made, we random sample 50 wrong predictions among 4283 error samples in relation P37 "*Official_Languages*". Results are shown in Table 11.

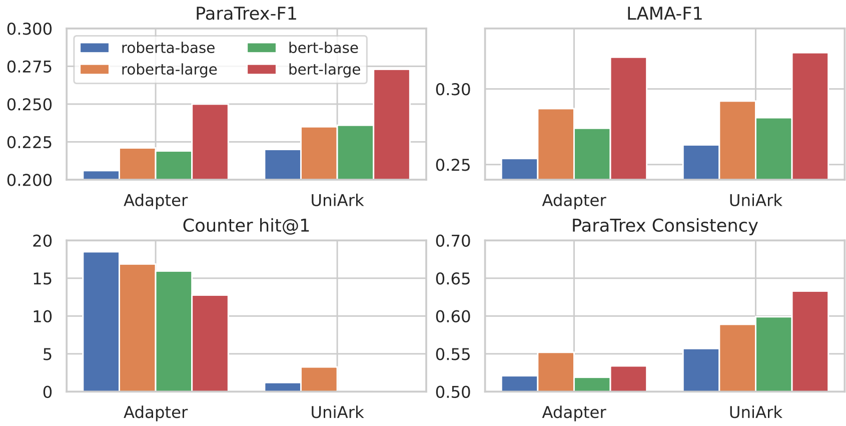

Figure 5: Sscaling results between adapters and UniArk with different scales of models.

Paired T-test	**ood_f1_pt**	**ood_f1_pr**	**all_cst_pt**	**all_cst_pr**	**acc_cst_pt**	**acc_cst_pr**	**id_lm_f1**
BERT	1.36e-04	3.19e-03	1.26e-06	7.82e-06	2.40e-05	6.20e-05	6.26e-01
RoBERTa	7.35e-04	9.39e-03	2.19e-03	1.69e-04	7.28e-03	2.92e-03	4.61e-01
Wil rank Test							
BERT	1.83e-05	3.78e-03	1.19e-07	4.17e-07	2.56e-06	8.34e-07	5.37e-02
RoBERTa	7.50e-05	1.15e-02	2.17e-04	1.51e-05	2.87e-04	3.29e-04	5.65e-02
T-Test							
BERT	1.06e-04	4.80e-03	6.13e-04	5.02e-04	6.09e-05	2.73e-04	5.03e-02
RoBERTa	1.48e-03	1.23e-02	5.21e-03	3.63e-03	1.16e-03	1.09e-03	6.65e-02

Table 8: Significance test between adapter baseline and UniArk over 41 relations for f1 score and 25 relations for consistency (cst) on ParaTrex (pt) and ParaRel (pr).

Inputs (Subject: Vesanto, Object: Finnish)		Predictions	
Type	Prompt template	Adapter-Tuning	UniArk
raw	The official language of [X] is [MASK].	**Finnish**	**Finnish**
paraphrased	[X] designates [MASK] as the official language .	Italian	**Finnish**
	[X] has [MASK] as its official language .	It	**Finnish**
	[MASK] has been declared as the recognized language in [X] .	Finland	**Finnish**
	In [X], [MASK] is acknowledged as the prescribed language by the government.	It	Finland
	The officially recognized language in [X] is [MASK] .	Italian	Italian
	[X] recognizes [MASK] as its official language .	Italian	**Finnish**

Table 9: LM prediction examples from the raw inputs in LAMA and the diverse paraphrased prompts in ParaTrex.

Method	Input	Subject="Sorengo"		
		Top 1	Top 2	Top 3
UniArk	raw	**Italian**	Finnish	Swedish
		0.1213	0.1152	0.1125
	subject masked	Polish	German	Greek
		0.0423	0.0421	0.0421
MeCoD	raw	Finnish	Swedish	Norwegian
		0.1322	0.1232	0.1041
	subject masked	French	Danish	Armenian
		0.1153	0.1051	0.0995
LAMA	raw	Portuguese	English	Spanish
		0.116	0.1146	0.1125
	subject masked	English	French	Spanish
		0.1111	0.1079	0.1016

Table 10: Case study on top-3 objects and their logits extracted by LMs through the original prompt template.

Error Type	N	Example			
		Subject	Prompt	Golden	Prediction
Unknown Case	23	Azad Kashmir	Azad Kashmir bestows official language status upon [Y] .	Urdu	English
Spelling Error	2	Melitopol	[Y] holds the official language designation of Melitopol .	Ukrainian	Ukraine
Pronouns	4	Malax	[Y] is officially recognized as the language of [X] .	Finnish	It
Multiple Correct Answers	21	ASEAN	The designated official language of ASEAN is [Y] .	Thai	Indonesia

Table 11: Types of errors appeared in UniArk on LAMA and ParaTrex test datasets

In the following questions, we provide 1 original input and 3 probable paraphrases. Please choose the sentances you think that are NOT paraphrases of the original inputs. For example, please answer 1-1 if you think the first sentance of the first question is NOT the paraphrase of the orignal sentance. Please answer 1-0 if you think all candidates of the first question are the paraphrase of the question.

Note that there may be several or no answer for a certain question.

You can use translation machine to translate a certain word if you do not understand it. But please write answers based on your own understanding. DO NOT translate the whole sentance and make predictions using automatic machines!

1: Original sentence: "[X] died in [Y] ."
Example: "Otto Brahm died in Berlin . || Nicholas V died in Rome ."
Example [X]: "Otto Brahm || Berlin"
Example [Y]: "Nicholas V || Rome"
Description: "most specific known (e.g. city instead of country, or hospital instead of city) death location of a person, animal or fictional character"
Paraphrase candidates:
1. The final moments of [X] took place in [Y] .
2. [Y] was the means of expression for [X] .
3. [X]'s passing occurred in [Y] .
Ans:

2: Original sentence: "[X] is a subclass of [Y] ."
Example: "quarter note is a subclass of note . || Doublecortin is a subclass of protein ."
Example [X]: "quarter note || note"
Example [Y]: "Doublecortin || protein"
Description: "all instances of these items are instances of those items; this item is a class (subset) of that item. Not to be confused with P31 (instance of)"
Paraphrase candidates:
1. [X] is an offshoot of [Y] .
2. [X] used [Y] as their language of interaction .
3. [X] is grouped within [Y] .
Ans:

Figure 6: Example of the questions for human evaluation