Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases

Boxi Cao[1,3], **Hongyu Lin**[1*], **Xianpei Han**[1,2*], **Le Sun**[1,2]
Lingyong Yan[1,3], **Meng Liao**[4], **Tong Xue**[4], **Jin Xu**[4]
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Data Quality Team, WeChat, Tencent Inc., China
{boxi2020,hongyu,xianpei,sunle,lingyong2014}@iscas.ac.cn
{maricoliao,xavierxue,jinxxu}@tencent.com

Abstract

Previous literatures show that pre-trained masked language models (MLMs) such as BERT can achieve competitive factual knowledge extraction performance on some datasets, indicating that MLMs can potentially be a reliable knowledge source. In this paper, we conduct a rigorous study to explore the underlying predicting mechanisms of MLMs over different extraction paradigms. By investigating the behaviors of MLMs, we find that previous decent performance mainly owes to the biased prompts which overfit dataset artifacts. Furthermore, incorporating illustrative cases and external contexts improve knowledge prediction mainly due to entity type guidance and golden answer leakage. Our findings shed light on the underlying predicting mechanisms of MLMs, and strongly question the previous conclusion that current MLMs can potentially serve as reliable factual knowledge bases[1].

1 Introduction

Recently, pre-trained language models (Peters et al., 2018; Devlin et al., 2019; Brown et al., 2020) have achieved promising performance on many NLP tasks. Apart from utilizing the universal representations from pre-trained models in downstream tasks, some literatures have shown the potential of pre-trained masked language models (e.g., BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019b)) to be factual knowledge bases (Petroni et al., 2019; Bouraoui et al., 2020; Jiang et al., 2020b; Shin et al., 2020; Jiang et al., 2020a; Wang et al., 2020; Kassner and Schütze, 2020a; Kassner et al., 2020). For example, to extract the birthplace of *Steve Jobs*, we can query MLMs like BERT with "*Steve Jobs was born in [MASK]*", where *Steve Jobs* is the subject of the fact, *"was born in"* is a prompt string for the relation "`place-of-birth`" and [MASK] is a placeholder for the object to predict. Then MLMs are expected to predict the correct answer "*California*" at the [MASK] position based on its internal knowledge. To help MLMs better extract knowledge, the query may also be enriched with external information like illustrative cases (*e.g., (Obama, Hawaii)*) (Brown et al., 2020) or external context (*e.g., Jobs lives in California*) (Petroni et al., 2020). Some literatures have shown that such paradigms can achieve decent performance on some benchmarks like LAMA (Petroni et al., 2019).

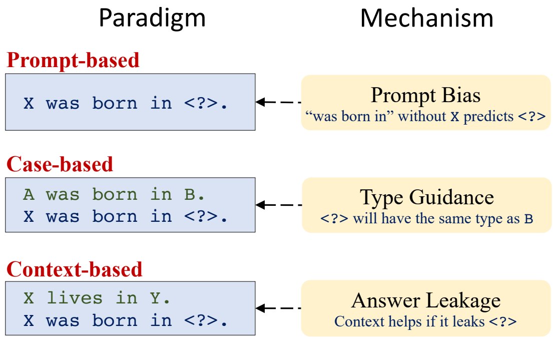


Figure 1: This paper explores three different kinds of factual knowledge extraction paradigms from MLMs, and reveal the underlying predicting mechanisms behind them.

Despite some reported success, currently there is no rigorous study looking deeply into the underlying mechanisms behind these achievements. Besides, it is also unclear whether such achievements depend on certain conditions (e.g., datasets, domains, relations). The absence of such kind of studies undermines our trust in the predictions of MLMs. We could neither determine whether the predictions are reliable nor explain why MLMs make a specific prediction, and therefore significantly limits MLMs' further applications and improvements.

[*] Corresponding Authors

[1] We openly release the source code and data at https://github.com/c-box/LANKA

*Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing*, pages 1860–1874
August 1–6, 2021. ©2021 Association for Computational Linguistics

To this end, this paper conducts a thorough study on whether MLMs could be reliable factual knowledge bases. Throughout our investigations, we analyze the behaviors of MLMs, figure out the critical factors for MLMs to achieve decent performance, and demonstrate how different kinds of external information influence MLMs' predictions. Specifically, we investigate factual knowledge extraction from MLMs[2] over three representative factual knowledge extraction paradigms, as shown in Figure 1:

- **Prompt-based retrieval** (Petroni et al., 2019; Jiang et al., 2020b; Shin et al., 2020), which queries MLM for object answer only given the subject and the corresponding relation prompt as input, e.g., *"Jobs was born in [MASK]."*

- **Case-based analogy** (Brown et al., 2020; Madotto et al., 2020; Gao et al., 2020), which enhances the prompt-based retrieval with several illustrative cases, e.g., *"Obama was born in Hawaii. [SEP] Jobs was born in [MASK]."*

- **Context-based inference** (Petroni et al., 2020; Bian et al., 2021), which augments the prompt-based retrieval with external relevant contexts, e.g., *"Jobs lives in California. [SEP] Jobs was born in [MASK]."*

Surprisingly, the main conclusions of this paper somewhat diverge from previous findings in published literatures, which are summarized in Figure 1. For prompt-based paradigm (§ 3), we find that the prediction distribution of MLMs is significantly prompt-biased. Specifically, we find that prompt-based retrieval generates similar predictions on totally different datasets. And predictions are spuriously correlated with the applied prompts, rather than the facts we want to extract. Therefore, previous decent performance mainly stems from the prompt over-fitting the dataset answer distribution, rather than MLMs' knowledge extraction ability. Our findings strongly question the conclusions of previous literatures, and demonstrate that current MLMs can not serve as reliable knowledge bases when using prompt-based retrieval paradigm.

For case-based paradigm (§ 4), we find that the illustrative cases mainly provide a "type guidance" for MLMs. To show this, we propose a novel algorithm to induce the object type of each relation based on Wikidata[3] taxonomy. According to the induced types, we find that the performance gain brought by illustrative cases mainly owes to the improvement on recognizing object type. By contrast, it cannot help MLMs select the correct answer from the entities with the same type: the rank of answer within its entity type is changed randomly after introducing illustrative cases. That is to say, under the case-based paradigm, although MLMs can effectively analogize between entities with the same type, they still cannot well identify the exact target object based on their internal knowledge and the provided illustrative cases.

For context-based paradigm (§ 5), we find that context can help the factual knowledge extraction mainly because it explicitly or implicitly leaks the correct answer. Specifically, the knowledge extraction performance improvement mainly happens when the introduced context contains the answer. Furthermore, when we mask the answer in the context, the performance still significantly improves as long as MLMs can correctly reconstruct the masked answer in the remaining context. In other words, in these instances, the context itself servers as a delegator of the masked answer, and therefore MLMs can still obtain sufficient implicit answer evidence even the answer doesn't explicitly appear.

All the above findings demonstrate that current MLMs are not reliable in factual knowledge extraction. Furthermore, this paper sheds some light on the underlying predicting mechanisms of MLMs, which can potentially benefit many future studies.

2 Related Work

The great success of Pre-trained Language Models (PLMs) raises the question of whether PLMs can be directly used as reliable knowledge bases. Petroni et al. (2019) propose the LAMA benchmark, which probes knowledge in PLMs using prompt-based retrieval. Jiang et al. (2020a) build a multilingual knowledge probing benchmark based on LAMA. There are many studies focus on probing specific knowledge in PLMs, such as linguistic knowledge (Lin et al., 2019; Tenney et al., 2019; Liu et al., 2019a; Htut et al., 2019; Hewitt and Manning, 2019; Goldberg, 2019; Warstadt et al., 2019),

[2] This paper shows the experimental results on BERT-large because previous work has shown that it can achieve the best performance on factual knowledge extraction among all MLMs. In the Appendix, we also report the experimental results on RoBERTa-large, which also reach the main conclusions reported in the paper.

[3] www.wikidata.org

semantic knowledge (Tenney et al., 2019; Wallace et al., 2019; Ettinger, 2020) and world knowledge (Davison et al., 2019; Bouraoui et al., 2020; Forbes et al., 2019; Zhou et al., 2019; Roberts et al., 2020; Lin et al., 2020; Tamborrino et al., 2020). Recently, some studies doubt the reliability of PLMs as knowledge base by discovering the the spurious correlation to surface forms (McCoy et al., 2019; Poerner et al., 2020; Shwartz et al., 2020), and their sensitivity to "negation" and "mispriming" (Kassner and Schütze, 2020b).

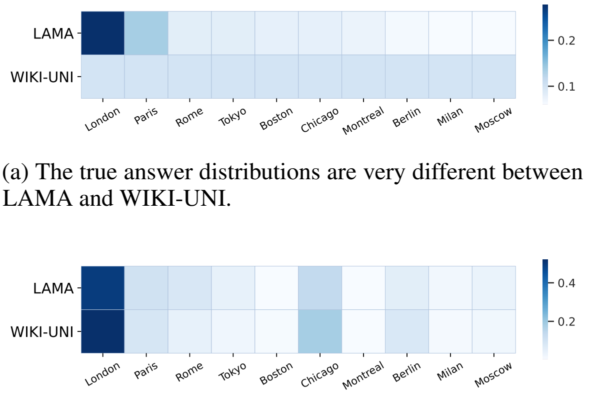
Currently, there are three main paradigms for knowledge extraction from PLMs: prompt-based retrieval (Schick and Schütze, 2021; Li and Liang, 2021), case-based analogy (Schick and Schütze, 2020a,b), and context-based inference. For prompt-based retrieval, current studies focus on seeking better prompts by either mining from corpus (Jiang et al., 2020b) or learning using labeled data (Shin et al., 2020). For case-based analogy, current studies mostly focus on whether good cases will lead to good few-shot abilities, and many tasks are tried (Brown et al., 2020; Madotto et al., 2020; Gao et al., 2020). For context-based inference, current studies focus on enhancing the prediction by seeking more informative contexts, e.g., for knowledge extraction (Petroni et al., 2020) and CommonsenseQA (Bian et al., 2021). However, there is no previous work which focuses on systematically study the underlying predicting mechanisms of MLMs on these paradigms.

3 Prompt-based Retrieval

The prompt-based retrieval extracts factual knowledge by querying MLMs with (subject, prompt, [MASK]). For example, to extract the "`place-of-birth`" of *Steve Jobs*, we could query BERT with "*Steve Jobs was born in [MASK].*" and the predicted "*California*" would be regarded as the answer. We consider three kinds of prompts: the manually prompts T_{man} created by Petroni et al. (2019), the mining-based prompts T_{mine} by Jiang et al. (2020b) and the automatically searched prompts T_{auto} from Shin et al. (2020).

3.1 Overall Conclusion

Conclusion 1. *Prompt-based retrieval is prompt-biased. As a result, previous decent performance actually measures how well the applied prompts fit the dataset answer distribution, rather than the factual knowledge extraction ability from MLMs.*



(a) The true answer distributions are very different between LAMA and WIKI-UNI.

(b) However, the prediction distribution made by MLMs on them are still very similar.

Figure 2: An illustration example of the vastly different answer distributions but similar prediction distributions on LAMA and WIKI-UNI on "`place-of-birth`" relation.

Specifically, we conduct studies and find that 1) Prompt-based retrieval will generate similar responses given quite different datasets. To show this, we construct a new dataset from Wikidata – WIKI-UNI, which have a totally different answer distribution from the widely-used LAMA[4] dataset (Petroni et al., 2019). However, we find that the prediction distributions on WIKI-UNI and LAMA are highly correlated, and this spurious correlation holds across different prompts. Such results reveal that there is just a weak correlation between the predictions of MLMs and the factual answer distribution of the dataset. 2) The prediction distribution is dominated by the prompt, i.e., the prediction distribution using only (prompt, [MASK]) is highly correlated to the prediction distribution using (subject, prompt, [MASK]). This indicates that it is the applied prompts, rather than the actual facts, determine the predictions of MLMs. 3) The performance of the prompt can be predicted by the divergence between the prompt-only distribution and the answer distribution of the dataset. All these findings reveal that previous decent performance in this field actually measures the degree of prompt-dataset fitness, rather than the universal factual knowledge extraction ability.

3.2 Different Answers, Similar Predictions

Finding 1. *Prompt-based retrieval will generate similar responses to quite different datasets.*

A reliable knowledge extractor should generate

[4] Since we focus on factual knowledge, we use the T-REx (Elsahar et al., 2018) subset of the LAMA benchmark.

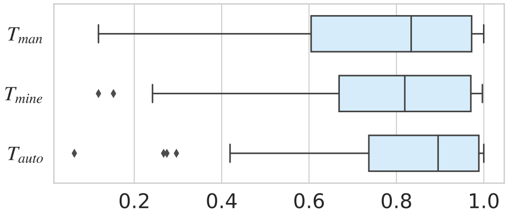


Figure 3: Correlations of the prediction distributions on LAMA and WIKI-UNI. Even these two datasets have totally different answer distributions, MLMs still make highly correlated predictions.

Distribution	Datasets	Top1	Top3	Top5	Precision
Answer	LAMA	22.04	39.37	48.03	-
	WIKI-UNI	1.68	5.03	7.78	-
Prediction	LAMA	31.09	49.21	57.93	30.36
	WIKI-UNI	27.12	44.19	52.18	16.47

Table 1: Average percentage of instances being covered by top-k answers or predictions. For answer distribution, top-5 objects in LAMA cover 6.2 times of instances than that in WIKI-UNI, however, for prediction distribution, they are almost the same. As a result, the precision is significantly dropped in WIKI-UNI.

different responses to different knowledge queries. To verify whether MLMs meet this standard, we manually construct a new dataset – WIKI-UNI, which has a comparable size but totally different answer distribution to LAMA, and then compare the prediction distributions on them. For a fair comparison, we follow the construction criteria of LAMA: we use the same 41 relations, filter out the queries whose objects are not in the MLMs' vocabulary. Compared with LAMA, the major difference is that WIKI-UNI has a uniform answer distribution, i.e., for each relation, we keep the same number of instances for each object. Please refer to Appendix for more construction details. Figure 2a shows the answer distributions of LAMA and WIKI-UNI on relation "`place-of-birth`". We can see that the answers in LAMA are highly concentrated on the head object entities, while the answers in WIKI-UNI follow a uniform distribution.

Given LAMA and WIKI-UNI, we investigate the predicting behaviors of MLMs. Surprisingly, the prediction distributions on these two totally different datasets are highly correlated. Figure 2b shows an example. We can see that the prediction distribution on WIKI-UNI is very similar to that on LAMA. And these two distributions are both close to the answer distribution of LAMA but far away from the answer distribution of WIKI-UNI.

To investigate whether this spurious correlation is a common phenomenon, we analyze the Pearson correlation coefficient between prediction distributions on LAMA and WIKI-UNI across different relations and three kinds of prompts. The boxplot in Figure 3 shows the very significant correlation between the prediction distributions on LAMA and WIKI-UNI: on all three kinds of prompts, the correlation coefficients exceed 0.8 in more than half of relations. These results demonstrate that prompt-based retrieval will lead to very similar prediction distributions even when test sets have vastly different answer distributions.

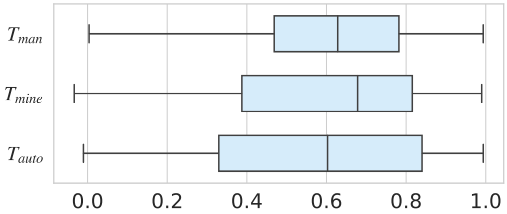


Figure 4: Correlations between the prompt-only distribution and prediction distribution on WIKI-UNI. MLMs make correlated predictions w. or w/o. subjects.

Furthermore, we find that the prediction distribution obviously doesn't correspond to the answer distribution of WIKI-UNI. From Table 1, we can see that on average, the top-5 answers of each relation in WIKI-UNI cover only 7.78% instances. By contrast, the top-5 predictions of each relation in WIKI-UNI cover more than 52% instances, which is close to the answer distribution and prediction distribution on LAMA. As a result, the performance on WIKI-UNI (mean P@1: 16.47) is significantly worse than that on LAMA (mean P@1: 30.36). In conclusion, the facts of a dataset cannot explain the predictions of MLMs, and therefore previous evaluations of the MLMs' ability on factual knowledge extraction are unreliable.

3.3 Prompts Dominates Predictions

Finding 2. *The prediction distribution is severely prompt-biased.*

To investigate the underlying factors of the predicting behavior of MLMs, we compare the prompt-only prediction distribution using only (prompt, [MASK]) and the full prediction distribution using (subject, prompt, [MASK]). To obtain the prompt-only distribution, we mask the subject and then use ([MASK], prompt, [MASK]) to query MLMs (*e.g., [MASK] was born in [MASK]*). Because there is no subject information in the input, MLMs can only depend on applied prompt's information to make the prediction at the second [MASK]. Therefore, we regard the probability distribution at the second [MASK] symbol as the prompt-only distribution.

After that, we analyze the correlations between the prompt-only distribution and the prediction distribution on WIKI-UNI dataset. Figure 4 shows the boxplot. On all three kinds of prompts, correlation coefficients between the prompt-only distribution and the prediction distribution on WIKI-UNI exceed 0.6 in more than half of relations. This demonstrates that in these relations, the prompt-only distribution dominates the prediction distribution. Combining with the findings in Section 3.2, we can summarize that the prompt-based retrieval is mainly based on *guided guessing*, i.e., the predictions are generated by sampling from the prompt-biased distribution guided by the moderate impact of subjects.

Note that among a minor part of relations, the correlations between the prompt-only distribution and the prediction distribution are relatively low. We find that the main reason is the type selectional preference provided by the subject entities, and Section 4 will further discuss the impact of this type-guidance mechanism for MLMs.

3.4 Better Prompts are Over-Fitting

Finding 3. *"Better" prompts are the prompts fitting the answer distribution better, rather than the prompts with better retrieval ability.*

Some previous literatures attempt to find better prompts for factual knowledge extraction from MLMs. However, as we mentioned above, the prompt itself will lead to a biased prediction distribution. This raises our concern that whether the found better prompts are really with better knowledge extraction ability, or the better performance just come from the over-fitting between the prompt-only distribution and the answer distribution of the test set.

To answer this question, we evaluate the KL divergence between the prompt-only distribution and the answer distribution of LAMA on different kinds of prompts. The results are shown in Table 2. We find that the KL divergence is a strong indicator of the performance of a prompt, i.e., the smaller the KL divergence between the prompt-only distribution and the answer distribution of the test set is, the better performance the prompt achieve. Furthermore, Table 3 shows several comparisons between different kinds of prompts and their performance on LAMA. We can easily observe that the better-performed prompts are actually over-fitting the dataset, rather than better capturing the underlying semantic of the relation. As a result, previous prompt searching studies are actually optimized on the spurious prompt-dataset compatibility, rather than the universal factual knowledge extraction ability.

Prompt	Precision	KL divergence
T_{man}	30.36	12.27
T_{mine}	39.49	10.40
T_{auto}	40.36	10.27

Table 2: The smaller KL divergence between the prompt-only distribution and golden answer distribution of LAMA, the better performance of the prompt.

Relation	Prompt	Source	Prec.	KL.
`citizenship`	x is y citizen	T_{man}	0.00	24.67
	x returned to y	T_{mine}	43.58	6.32
`work location`	x used to work in y	T_{man}	11.01	19.07
	x was born in y	T_{mine}	40.25	2.21
`instance of`	x is a y	T_{man}	30.15	22.98
	x is a small y	T_{mine}	52.60	13.98

Table 3: Examples of prompts that can achieve significant improvements on LAMA. We can see that the better performance actually stems from over-fitting: the better prompts are not prompts with a stronger semantic association to the relation.

4 Case-based Analogy

The case-based analogy enhances the prompt-based paradigm with several illustrative cases. For example, if we want to know the "`place-of-birth`" of *Steve Jobs*, we would first sample cases such as (*Obama*, `place-of-birth`, *Hawaii*), and combine them with the original query. In this way, we will use "*Obama was born in Hawaii. [SEP] Steve Jobs was born in [MASK].*" to query MLMs.

4.1 Overall Conclusion

Conclusion 2. *Illustrative cases guide MLMs to better recognizing object type, rather than better predicting facts.*

To show this, we first design an effective algorithm to induce the type of an entity set based on Wikidata taxonomy, which can identify the object type of a relation. According to the induced types, we find that the benefits of illustrative cases mainly stem from the promotion of object type recognition. In other words, case-based analogy guides MLMs with better type prediction ability but contributes

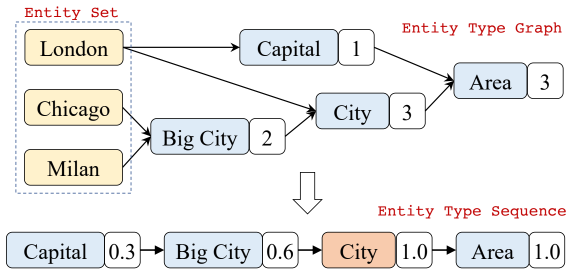


Figure 5: Illustration of our type induction algorithm. The numbers on the right of each type indicate how many entities does the type cover. The type of an entity set is the finest grained type in the type graph that can cover a sufficient number of the instances in the entity set, which is *City* in the example.

little to the entity prediction ability. In the following, we first illustrate our type inducing algorithm, and then explain how we reach the conclusion.

4.2 Entity Set Type Induction

To induce the object type of a relation, we first collect all its objects in LAMA and form an entity set. Then we induce the type of an entity set by designing a simple but effective algorithm. The main intuition behind our algorithm is that a representative type should be the finest grained type that can cover a sufficient number of the instances in the entity set. Figure 5 shows an example of our algorithm. Given a set of entities in Wikidata, we first construct an entity type graph (ETG) by recursively introducing all ancestor entity types according to the `instance-of` and `subclass-of` relations. For the example in Figure 5, *Chicago* is in the entity set and is an `instance-of` *Big City*. *Big City* is a `subclass-of` *City*. As a result, *Chicago*, *Big City* and *City* will all be introduced into ETG. Then we apply topological sorting (Cook, 1985) to ETG to obtain a *Fine-to-Coarse entity type sequence*. Finally, based on the sequence, we select the first type which covers more than 80% of entities in the entity set (e.g., *City* in Figure 5). Table 4 illustrates several induced types, and please refer to the Appendix for details.

4.3 Cases Help Type Recognition

Finding 4. *Illustrative cases help MLMs to better recognize the type of objects, and therefore improve factual knowledge extraction.*

For case-based analogy, the first thing we want to know is whether illustrative cases can improve the knowledge extraction performance. To this end, for each (subject, relation) query in LAMA, we randomly sample 10 illustrative cases. To avoid answer leakage, we ensure the objects of these cases don't contain the golden answer of the query. Then we use (cases, subject, prompt, [MASK]) as the analogous query to MLMs.

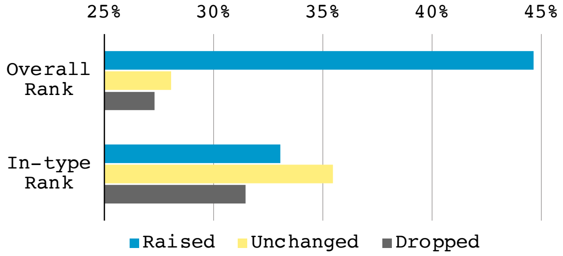


Figure 6: Percentages on the change of overall rank (among all candidates) and the in-type rank (among candidates with the same type) of golden answer. We can see that the illustrative cases mainly raise the overall rank but cannot raise the in-type rank, which means the performance improvements mainly come from better type recognition.

Results show that case-based analogy can significantly improve performance. After introducing illustrative cases, the mean precision increases from 30.36% to 36.23%. Besides, we find that 11.81% instances can benefit from the introduced cases and only 5.94% instances are undermined. This shows that case-based analogy really helps the MLMs to make better predictions.

By analyzing the predicting behaviors, we observe that the main benefit of introducing illustrative cases comes from the better type recognition. To verify this observation, we investigate how the types of predictions changed after introducing the illustrative cases. Table 4 shows the results on relations whose precision improvement is more than 10% after introducing illustrative cases. From the table, it is very obvious that illustrative cases enhance the factual knowledge extraction by improving type prediction: 1) For queries whose predictions are correctly reversed (from wrong to right), the vast majority of them stems from the revised type prediction; 2) Even for queries whose predictions are mistakenly reversed (from right to wrong), the type of the majority of predictions still remains correct. In conclusion, introducing illustrative cases can significantly improve the knowledge extraction ability by recognizing the object type more accurately. That is, adding illustrative cases will provide more guidance for object type.

Relation	Induced Object Type	Precision Δ	Type Prec. Δ	Wrong → Right w/ Type Change	Right → Wrong w/o Type Change
`country of citizenship`	sovereign state	43.37	84.16	100.00	-
`position held`	religious servant	36.88	80.26	91.15	90.00
`religion`	religion	33.20	34.88	100.00	-
`work location`	city	26.10	70.55	85.04	100.00
`instrument`	musical instrument	17.07	55.75	89.08	75.00
`country`	sovereign state	14.30	29.04	88.48	87.93
`employer`	business	12.01	99.22	100.00	-
`continent`	continent	10.87	51.18	96.86	88.24

Table 4: Detailed analysis on relations where the mean precision increased more than 10%. Precision Δ and Type Prec. Δ represents the precision changes on the answer and the type of the answer respectively. "w/ Type Change" and "w/o Type Change" represents the type of prediction changed/unchanged before/after introducing illustrative cases. "-" indicate there is no queries whose predictions are mistakenly reversed.

4.4 Cases do not Help Entity Prediction

Finding 5. *Illustrative cases are of limited help for selecting the answer from entities of the same type.*

To show this, we introduce a new metric referred as *in-type rank*, which is the rank of the correct answer within the entities of the same type for a query. By comparing the in-type rank in prompt-based and case-based paradigm, we can evaluate whether the illustrative cases can actually help better entity prediction apart from better type recognition.

Figure 6 shows the percentages on the change of overall rank (among all candidates) and the in-type rank (among candidates with the same type) of golden answer. Unfortunately, we find that illustrative cases are of limited help for entity prediction: the change of in-type rank is nearly random. The percentages of queries with Raised/Unchanged/Dropped in-type rank are nearly the same: 33.05% VS 35.47% VS 31.47%. Furthermore, we find that the MRR with the type only changed from 0.491 to 0.494, which shows little improvement after introducing illustrative cases. These results show that the raises of overall rank of golden answer are not because of the better prediction inside the same type. In conclusion, illustrative cases cannot well guide the entity prediction, and they mainly benefit the factual knowledge extraction by providing guidance for object type recognition.

5 Context-based Inference

The context-based inference augments the prompt-based paradigm with external contexts. For example, if we want to know the "`place-of-birth`" of *Steve Jobs*, we can use the external context "*Jobs was from California.*", and form a context-enriched query "*Jobs was from California. [SEP] Steve Jobs was born in [MASK].*" to query MLMs. Specifically, we use the same context retrieval method as Petroni et al. (2020): for each instance, given the subject and relation as query, we use the first paragraph of DRQA's (Chen et al., 2017) retrieved document as external contexts.

Answer in context	Prompt-based	Context-based	Δ
Present (45.30%)	34.83	64.13	+29.30
Absent (54.70 %)	25.37	23.26	-2.11

Table 5: Comparison between prompt-based and context-based paradigms grouped by whether the answer presents or absents in the context. We can see that only contexts containing the answer can significantly improve the performance.

5.1 Overall Conclusion

Conclusion 3. *Additional context helps MLMs to predict the answer because they contain the answer, explicitly or implicitly.*

Several studies (Petroni et al., 2020; Bian et al., 2021) show that external context can help knowledge extraction from MLMs. To investigate the underlying mechanism, we evaluate which kinds of information in contexts contribute to the fact prediction, and find that the improvement mainly comes from the answer leakage in context. Furthermore, we find the answers can not only be leaked explicitly, but can also be leaked implicitly if the context provides sufficient information.

5.2 Explicit Answer Leakage Helps

Finding 6. *Explicit answer leakage significantly improves the prediction performance.*

To show this, we split LAMA into two parts ac-

Prompt-based	Context-based	Masked Context-based
30.36	41.44	35.66

Table 6: Overall performance when introducing different kinds of contexts. "Masked Context-based" indicates that we mask the golden answer in contexts, and there is still a significant performance improvement.

Answer Reconstructable	Prompt-based	Context-based	Δ
Reconstructable (60.23%)	39.58	60.82	+21.24
Not-reconstructable (39.77 %)	28.84	35.83	+6.99

Table 7: Comparison between prompt-based and context-based paradigms grouped by whether the masked answer in the context can be reconstructed from the remaining context. We can see that contexts can reconstruct the masked answer is more likely to improve the performance.

cording to whether the additional context contains the answer. Table 5 shows the results on these two parts respectively. We can see that the improvements on these two parts diverge significantly. For context containing the answer, context-based inference significantly improves the factual knowledge extraction performance. However, there is even a little performance drop for those instances whose context does not contain the answer. This indicates that the improvement of factual knowledge extraction is mainly due to the explicit existence of the answer in the context.

5.3 Implicit Answer Leakage Helps

Finding 7. *Implicit answer leakage can also significantly improve the prediction performance.*

As we mentioned above, explicit answer leakage significantly helps the answer prediction. The answer-leaked context may explicitly provide the answer or implicitly guide the prediction by providing answer-specific information. To understanding the underlying mechanism, we mask the answer in the context and verify whether it can still achieve the performance gain.

Table 6 shows the results. We find that the performance gain is still very significant after masking the answer. This indicates that the contexts previously containing the answer are still very effective even the answer doesn't explicitly present. To further investigate the reason behind, we split the masked version of answer-leaked instances into two groups by whether MLMs can or cannot correctly reconstruct the masked answer from the remaining context. The results are shown in Table 7. We can see that the performance gain significantly diverges in these two groups: the improvements mainly come from the instances whose answer in context can be reconstructed – we refer to this as *implicit answer leakage*. That is to say, for these instances, the context serves as a sufficient delegator of its answer, and therefore MLMs can obtain sufficient answer evidence even the answer does not explicitly appear. However, for contexts that cannot reconstruct the masked answer, the improvements are relatively minor. In conclusion, the real efficacy of context-based inference comes from the sufficient answer evidence provided by the context, either explicitly or implicitly.

6 Conclusions and Discussions

In this paper, we thoroughly study the underlying mechanisms of MLMs on three representative factual knowledge extraction paradigms. We find that the prompt-based retrieval is severely prompt-biased, illustrative cases enhance MLMs mainly via type guidance, and external contexts help knowledge prediction mostly because they contain the correct answer, explicitly or implicitly. These findings strongly question previous conclusions that current MLMs could serve as reliable factual knowledge bases.

The findings of this paper can benefit the community in many directions. By explaining the underlying predicting mechanisms of MLMs, we provide reliable explanations for many previous knowledge-intensive techniques. For example, our method can explain why and how incorporating external contexts will help knowledge extraction and CommonsenseQA (Talmor et al., 2019). Our findings also reveal why PLM probing datasets may not be reliable and how the evaluation can be promoted by designing de-biased evaluation datasets.

This paper also sheds light on future research directions. For instance, knowing the main benefit of illustrative cases comes from type-guidance, we can enhance many type-centric prediction tasks such as NER (Lample et al., 2016) and factoid QA (Iyyer et al., 2014). Moreover, based on the mechanism of incorporating external contexts, we can better evaluate, seek, and denoise external contexts for different tasks using MLMs. For example, we can assess and select appropriate facts for CommonsenseQA based on whether they can reconstruct the candidate answers.

This paper focuses on masked language models, which have been shown very effective and are widely used. We also want to investigate another representative category of language models – the generative pre-trained models (e.g., GPT2/3 (Radford et al., 2019; Brown et al., 2020)), which have been shown to have quite different mechanisms and we leave it for future work due to page limitation.

Acknowledgments

We sincerely thank all anonymous reviewers for their insightful comments and valuable suggestions. This work is supported by the National Key Research and Development Program of China (No. 2020AAA0106400), the National Natural Science Foundation of China under Grants no. U1936207, and in part by the Youth Innovation Promotion Association CAS(2018141).

References

Ning Bian, Xianpei Han, Bo Chen, and Le Sun. 2021. Benchmarking Knowledge-Enhanced Commonsense Question Answering via Knowledge-to-Text Transformation. *arXiv:2101.00760 [cs]*.

Zied Bouraoui, José Camacho-Collados, and Steven Schockaert. 2020. Inducing relational knowledge from BERT. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7456–7463. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Stephen A. Cook. 1985. A taxonomy of problems with fast parallel algorithms. *Information and Control*, 64(1):2–22.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pre-trained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. 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.

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).

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do Neural Language Representations Learn Physical Commonsense? *arXiv:1908.02899 [cs]*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making Pre-trained Language Models Better Few-shot Learners. *arXiv:2012.15723 [cs]*.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. *arXiv:1901.05287 [cs]*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do Attention Heads in BERT Track Syntactic Dependencies? *arXiv:1911.12246 [cs]*.

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 conference on* *empirical methods in natural language processing (EMNLP)*, pages 633–644.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

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

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 552–564, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020a. BERT-kNN: Adding a kNN search component to pretrained language models for better QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3424–3430, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020b. 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.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270. The Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. *arXiv:2101.00190 [cs]*.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Andrea Madotto, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Language Models as Few-Shot Learner for Task-Oriented Dialogue Systems. *arXiv:2008.06239 [cs]*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Patrick S. H. Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-BERT: Efficient-yet-effective entity embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 803–818, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2020a. Few-Shot Text Generation with Pattern-Exploiting Training. *arXiv:2012.11926 [cs]*.

Timo Schick and Hinrich Schütze. 2020b. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *arXiv:2009.07118 [cs]*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 255–269. 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.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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 4149–4158. Association for Computational Linguistics.

Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. 2020. Pretraining is (almost) all you need: An application to commonsense reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887, Online. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. 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 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Chenguang Wang, Xiao Liu, and Dawn Song. 2020. Language Models are Open Knowledge Graphs. *arXiv:2010.11967 [cs]*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's knowledge of language: Five analysis methods with NPIs. 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 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2019. Evaluating Commonsense in Pre-trained Language Models. *arXiv:1911.11931 [cs]*.

A WIKI-UNI Construction Details

To construct WIKI-UNI, we first collect all the triples which belong to the same 41 relations with LAMA from Wikidata (Vrandečić and Krötzsch, 2014), then we randomly sample 50K triples with a single-token object for each relation. Similar to LAMA, we filter out the instances whose object is not in MLMs' vocabulary. For each relation, we group the instances based on different objects, and indicate f_o as the frequency of each object. We denote the median of f_o with f_m. For groups where $f_o > f_m$, we randomly sample f_m instances, and delete the groups where $f_o < f_m$. Therefore, we acquire a dataset named WIKI-UNI with a uniform answer distribution. There are 70K facts in WIKI-UNI and 34K facts in LAMA. Since BERT and RoBERTa have a different vocabulary, so the datasets for their evaluation are slightly different.

B Results on RoBERTa-large

Our conclusions are similar on BERT-large and RoBERTa-large, therefore, we report the results of BERT-large in the article and results of RoBERTa-large here.

B.1 Promp-based Retrieval

Figure 7 shows the very significant correlation between the prediction distributions on LAMA and WIKI-UNI for RoBERTa-large: on all three kinds of prompts, the Pearson correlation coefficient between these two prediction distributions exceeds 0.9 in most relations. Table 8 shows the percentage of instances that the topk object entities cover for RoBERTa-large.

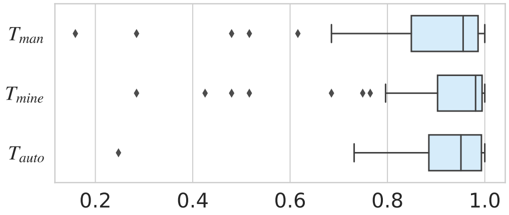


Figure 7: The correlations of the prediction distribution on LAMA and WIKI-UNI for RoBERTa-large.

B.2 Case-based Analogy

Table 9 shows the performance improvement after introducing illustrative cases for RoBERTa-large model, we can see that the illustrative cases could also significantly increase the knowledge extraction performance for RoBERTa-large. Table 14 shows how the entity types of predictions changed after introducing the illustrative cases for RoBERTa-large model, the conclusion is similar with BERT-large. Figure 8 shows the percentage on the change of overall rank and in-type rank for RoBERTa-large model.

Distribution	Datasets	Top1	Top3	Top5	Prec.
Answer	LAMA	23.93	42.02	50.08	-
	WIKI-UNI	1.84	5.53	8.61	-
Prediction	LAMA	37.48	56.85	65.45	23.65
	WIKI-UNI	36.53	55.51	63.58	13.59

Table 8: The percentage of instances that the topk object entities cover for RoBERTa-large. The statistics is different from Table 1 because we filter LAMA with RoBERTa's vocabulary when evaluate RoBERTa-large.

And another finding is that BERT-large has a better type prediction ability than RoBERTa-large, even without illustrative cases. We calculate the overall type precision over prompt-based paradigm (the percentage of predictions that the type is correct). And the type precision for BERT-large is 68% and for RoBERTa-large is only 51%, which partly explains why performance of RoBERTa-large is significantly worse than BERT-large on LAMA dataset.

Enhanced with Cases	Prec.	Better	Worse
No	23.65	-	-
Yes	29.78	14.09	7.96

Table 9: Performance of the case-based analogy paradigm for RoBERTa-large

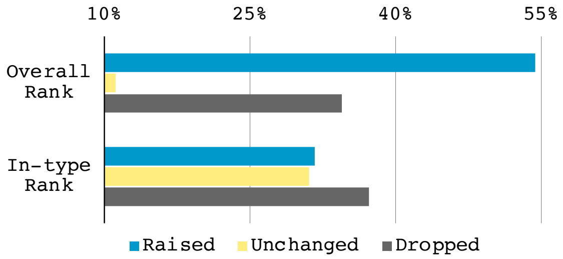


Figure 8: Percentages on the change of overall rank (among all candidates) and the in-type rank (among candidates with the same type) of golden answer of RoBERTa-large model.

B.3 Context-based Inference

Table 10 shows the comparison of contexts group by whether the contexts contain the answer for RoBERTa-large. We can see that for contexts containing the answer, context-based inference significantly improves the factual extraction performance. Meanwhile, there is a performance drop for those instances whose context does not contain the answer. Table 11 shows the overall performance improvements when introducing different external contexts for RoBERTa-large. Table 12 shows the comparison of the masked contexts based on whether they can/cannot reconstruct the masked answer for RoBERTa-large. The improvements mainly comes from the instances whose answer in contexts can be reconstructed.

Answer in context	Prompt-based	Context-based	Δ
Present (46.04%)	27.95	52.05	+24.10
Absent (53.96 %)	18.95	14.72	-4.23

Table 10: Comparison of contexts grouped by whether the answer presents or absents for RoBERTa-large.

Without Contexts	Full Contexts	Masked Contexts
23.65	31.44	24.44

Table 11: The overall performance when introducing different contexts for RoBERTa-large.

Answer Reconstructable	Prompt-based	Context-based	Δ
Reconstructable (61.23%)	30.50	42.37	+11.87
Not-reconstructable (38.77 %)	22.19	22.15	-0.04

Table 12: Comparison of the masked contexts based on whether they can/cannot reconstruct the masked answer for RoBERTa-large.

C Full Version of the Type Prediction Results

Table 13 shows the detailed analysis of all relations using case-based analogy paradigm for BERT-large and Table 14 is the results on RoBERTa-large. Because of the page limit, another finding we didn't mention in the article is that, apart from "type guidance", the illustrative cases could also provide a "surface form guidance" in a few relations (e.g., `part of`, `applies to jurisdiction`, `subclass of`). Specifically, the "surface form" indicate that the object entity name (*e.g., Apple*) is a substring of the subject entity name (*e.g., Apple Watch*). Such phenomenon is also mentioned in Poerner et al. (2020).

Relation	Induced Object Type	Precision	Type	Wrong → Right	Right → Wrong
		Δ	Prec. Δ	w/ Type Change	w/o Type Change
named after	physical object	68.06	98.91	99.77	-
country of citizenship	sovereign state	43.37	84.16	100.00	-
position held	religious servant	36.88	80.26	91.15	90.00
religion	religion	33.20	34.88	100.00	-
work location	city	26.10	70.55	85.04	100.00
instrument	musical instrument	17.07	55.75	89.08	75.00
country	sovereign state	14.30	29.04	88.48	87.93
employer	business	12.01	99.22	100.00	-
continent	continent	10.87	51.18	96.86	88.24
languages spoken, written or signed	Indo-European languages	9.91	-0.93	10.56	81.54
applies to jurisdiction	state	8.71	-6.13	7.23	63.64
country of origin	sovereign state	8.36	33.22	71.64	98.28
subclass of	object	7.68	27.28	66.18	87.10
part of	object	7.51	37.66	54.27	97.87
language of work or name	Indo-European languages	6.05	10.95	77.23	77.08
location of formation	city	5.02	66.34	80.77	100.00
has part	abstract object	5.02	27.26	25.33	100.00
genre	series	4.62	17.61	95.45	-
owned by	organization	2.62	11.50	9.57	100.00
instance of	concrete object	2.06	4.34	35.80	96.77
occupation	profession	1.35	-0.53	0.00	100.00
place of death	city	1.26	16.37	68.63	100.00
twinned administrative body	city	0.91	0.80	15.38	75.00
diplomatic relation	sovereign state	0.80	1.11	10.00	100.00
native language	Indo-European languages	0.20	0.62	38.64	92.86
manufacturer	business	-1.02	0.31	33.33	61.29
field of work	knowledge	-1.15	0.00	26.09	90.32
developer	enterprise	-1.52	1.52	4.17	96.97
location	community	-1.57	4.59	3.03	100.00
capital	city	-2.00	0.14	4.55	97.22
position played on team / speciality	position	-4.10	11.03	-	100.00
headquarters location	city	-4.24	0.62	0.00	100.00
official language	Nostratic languages	-5.28	-1.14	5.45	90.57
original language of film or TV show	Nostratic languages	-5.84	-16.71	19.15	43.30
place of birth	city	-6.25	4.34	14.29	100.00
capital of	political territorial entity	-6.84	0.42	-	100.00
shares border with	community	-7.37	2.72	2.22	97.35
record label	record label	-7.93	-22.38	-	0.00
original network	television station	-10.56	0.45	11.36	86.13
located in the administrative territorial entity	community	-12.94	11.69	10.53	99.25
member of	organization	-14.67	16.45	94.74	98.08

Table 13: A detailed analysis of all relations using case-based analogy paradigm for BERT-large, which is corresponding to Table 4 in the article. "-" indicates the number of queries whose predictions are reversed correctly or mistakenly is less than 3.

Relation	Induced Object Type	Precision	Type	Wrong → Right	Right → Wrong
		Δ	Prec. Δ	w/ Type Change	w/o Type Change
religion	religion	56.92	66.36	100.00	-
position held	religious servant	41.86	47.42	99.03	-
country of citizenship	sovereign state	37.16	74.11	100.00	-
member of	organization	31.03	77.83	100.00	-
continent	continent	29.51	87.80	100.00	100.00
instrument	musical instrument	28.26	6.04	94.04	0.00
country of origin	sovereign state	28.18	94.92	99.61	100.00
country	sovereign state	26.64	69.84	95.22	96.55
part of	object	24.57	90.22	96.98	100.00
place of death	city	22.88	95.35	98.95	100.00
instance of	concrete object	14.97	20.53	34.30	97.50
location of formation	city	14.12	99.88	100.00	-
subclass of	object	12.07	26.25	63.31	90.00
capital	city	10.62	36.31	92.19	85.71
named after	physical object	10.25	85.05	100.00	100.00
language of work or name	Indo-European languages	9.10	26.72	89.12	72.17
has part	abstract object	8.79	67.99	77.65	-
work location	city	8.09	12.43	96.95	6.45
languages spoken, written or signed	Indo-European languages	5.09	17.75	54.20	86.90
employer	business	3.97	10.31	19.05	100.00
position played on team / speciality	position	3.26	56.51	71.43	75.00
native language	Indo-European languages	1.09	1.63	28.21	93.10
genre	series	1.05	0.23	75.00	66.67
record label	record label	0.00	-7.55	-	-
place of birth	city	-0.13	41.02	66.67	100.00
twinned administrative body	city	-0.45	1.04	0.00	100.00
headquarters location	city	-1.00	0.00	0.00	100.00
diplomatic relation	sovereign state	-1.16	1.05	25.00	100.00
owned by	organization	-1.45	43.78	64.62	94.59
field of work	knowledge	-2.10	0.69	10.53	96.77
occupation	profession	-2.43	0.00	0.00	100.00
official language	Nostratic languages	-3.11	3.88	18.37	97.40
located in the administrative territorial entity	community	-3.35	45.81	75.93	97.50
original language of film or TV show	Nostratic languages	-5.29	-21.30	15.38	34.29
shares border with	community	-9.82	0.16	0.00	98.86
location	community	-11.49	27.15	41.43	100.00
developer	enterprise	-12.25	6.80	37.50	79.41
original network	television station	-16.46	-15.84	14.29	72.49
applies to jurisdiction	state	-18.38	2.11	35.71	98.00
capital of	political territorial entity	-39.44	7.22	-	100.00
manufacturer	business	-49.63	6.79	44.44	93.82

Table 14: A detailed analysis of all relations using case-based analogy paradigm for RoBERTa-large, which is corresponding to Table 4 in the article. "-" indicates the number of queries whose predictions are reversed correctly or mistakenly is less than 3.