Decoding Prompt Syntax: Analysing its Impact on Knowledge Retrieval in Large Language Models

Stephan Linzbach
Stephan.Linzbach@gesis.org
GESIS - Leibniz Institute for Social Sciences
Germany

Tim Tressel
Heinrich Heine University
Düsseldorf, Germany
Tim.Tressel@hhu.de

Laura Kallmeyer
Heinrich Heine University
Düsseldorf, Germany
Laura.Kallmeyer@hhu.de

Stefan Dietze
GESIS - Leibniz Institute for Social Sciences,
Heinrich Heine University
Germany
Stefan.Dietze@gesis.org

Hajira Jabeen
GESIS - Leibniz Institute for Social Sciences
Germany
Hajira.Jabeen@gesis.org

ABSTRACT

Large Language Models (LLMs), with their advanced architectures and training on massive language datasets, contain unexplored knowledge. One method to infer this knowledge is through the use of cloze-style prompts. Typically, these prompts are manually designed because the phrasing of these prompts impacts the knowledge retrieval performance, even if the LLM encodes the desired information. In this paper, we study the impact of prompt syntax on the knowledge retrieval capacity of LLMs. We use a template-based approach to paraphrase simple prompts into prompts with a more complex grammatical structure. We then analyse the LLM performance for these structurally different but semantically equivalent prompts. Our study reveals that simple prompts work better than complex forms of sentences. The performance across the syntactical variations for simple relations (1:1) remains best, with a marginal decrease across different typologies. These results reinforce that simple prompt structures are more effective for knowledge retrieval in LLMs and motivate future research into the impact of prompt syntax on various tasks.

CCS CONCEPTS

• **Computing methodologies → Information extraction**; *Language resources*; *Lexical semantics.*

KEYWORDS

Large Language models, BERT, Syntax aware prompt, Knowledge retrieval

ACM Reference Format:
Stephan Linzbach, Tim Tressel, Laura Kallmeyer, Stefan Dietze, and Hajira Jabeen. 2023. Decoding Prompt Syntax: Analysing its Impact on Knowledge Retrieval in Large Language Models. In *Companion Proceedings of the ACM Web Conference 2023 (WWW '23 Companion), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3543873.3587655

1 INTRODUCTION

Recent advancements in Large Language Models (LLMs) have led to significant progress in various natural language processing tasks such as translation, summarization, and question-answering by providing efficient representations of language in a self-supervised way. In addition to encoding linguistic and syntactic knowledge [7, 9], recent studies have demonstrated [15, 23] that deep LLMs also capture relational knowledge, enabling them to support basic question answering and reasoning tasks. Petroni et al. [22] show that introducing an information retrieval component which captures the relevant context from a LLM for a relation or factual question significantly improves performance in knowledge extraction tasks, where the context may include both specific semantics and syntactic characteristics. Additionally, leveraging of syntactic information while training complex language models is shown to improve representational quality across languages [26, 28]. Taking into account the beneficial influence of additional syntactical information, it seems natural to question the linguistic reliability of contextualized embeddings when testing for relational knowledge retrieval and extraction. This is further stressed by the debate on the extent to which acquired knowledge generalises beyond the statements seen as part of training data [14]. In particular, the generalisation of knowledge across different syntactic transformations of seen relational knowledge defines a desirable property of reliable LLMs. Although an in-depth investigation of the dependencies between syntactical-information and relational knowledge in LLMs seems promising, it remains underexplored. Based on these insights, we investigate the hypothesis that syntactic structure plays a role in the inference of knowledge from language models. In this paper, we

WWW '23 Companion, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9419-2/23/04…$15.00
https://doi.org/10.1145/3543873.3587655

Typological-transformation	Template
simple	The capital of [S] is [O] . [S] maintains diplomatic relations with [O].
compound	[S] is a country and it's capital is [O]. [S] maintains diplomatic relations with countries and [O] is one of them.
complex	[S] is the country, who's capital is [O]. [S] is a country that maintains diplomatic relations with [O].
compound-complex	[O] is a city and it is the city that is the capital of [S]. [S] is a country that maintains diplomatic relations with [O].

Table 1: Templates for 'capital of' (1:1) and 'diplomatic relation' (M:N)

present initial experiments to study this hypothesis and share manually created prompts with the community [1]. We have extended T-REx to incorporate different grammatical structures alongside the relations already provided. The key findings of our work are that a simple sentence structure performs better for relational knowledge extraction than complex grammatical constructions. However, the impact of sentence structures is negligible for simpler relations (1:1). Moreover, these relations are easier to extract than complex relations (N:M).

Overall, this paper is organised as follows: Firstly (Related Work), we briefly cover the state of the relevant research for this paper. The main section (Preliminary Experiments) is then divided into four subsections, three of which describe the methodology (Data, Task & metrics, Prompt Engineering), while the final subsection Results presents the performance of different models on our earlier established experimental setting. We conclude our work with a discussion and an outlook (Conclusion and Future Work).

2 RELATED WORK

Since the proposal of transformer-based LLMs which learn representations through Masked Language Modelling (MLM-task) [2], two research fields have emerged: (1) Understanding the knowledge inherent in LLMs [25], and (2) Enhancing the LLMs' inherent knowledge [33].

Understanding the knowledge inherent in LLMs:

Current research generally proposes two different methods to test the self-taught knowledge of LLMs. (a) Prompts that pose knowledge related tasks in a cloze-text format. This research direction is heavily influenced by the LAMA-probe proposed by [23], a cloze-text data-set that encodes simple relational facts about real world entities. E.g. the prompt 'Where was Dante born [MASK]?' is paired with 'Florence'. Using BERT [2] for predicting missing tokens, the authors show that BERT already carried a surprisingly high amount of relational knowledge. Following Petroni et al's [15, 23] findings, Heinzerling et al. [8] focus on entity representations, storage capacity and paraphrased queries. However, they draw a more critical picture of storage and query capabilities of these models. Moreover, Roberts et al. [24] investigate how much knowledge can be stored in model parameters. To approximate the storage capacities, they over-fit the model on knowledge triples. Since then, many probing-suites have been published to understand the impact of memorization and knowledge types (KMIR [6], KAMEL [13]). Furthermore, the performance improvements which were achieved through fine-tuning LLMs on the provided prompts were investigated. For this, an archive with different prompts as well as train, validation, and test-splits for the T-REx subset of the LAMA-probe called LPAQA [12] was created. (b) In addition to prompts, probing tasks are often used to investigate the knowledge encoded in LLMs. This method uses auxiliary classifications with features derived from the frozen network to understand inherent information. For the example of transformer-based language models, probing tasks can be solved by using the output representations [29], the attention information [1] or the information change across the different layers [11, 29]. The information derivable from those features has been used to understand several aspects of the contextualization of the representation [29], the syntactic truthfulness of the attention mechanism [1], and the workflow of the layer-wise processing [11].

Enhancing the LLMs inherent knowledge:

Various types of information are used to enhance the model's inherent knowledge. Approaches range from enhancing lexical word relations [18], in-context semantic abstractions [19], sentiment sensitivity [17, 30], and entity centred information [5, 21] to improving any knowledge type [20, 31]. Knowledge enhancement approaches also differ in their infusion technique. Proposals that stay the closest to pure language modelling only change the probabilities of the corruption task in a way in which it teaches stance- [16], or entity-knowledge [27]. Another infusion strategy tries to enhance the model by simultaneously teaching a secondary learning objective. This is applied to entity- [32], sentiment- [30] and general linguistic knowledge [20].

In this paper, we focus on *understanding the knowledge inherent in LLMs*. In particular, we aim to study the impact of syntactical differences while treating LLMs as a black box model. In comparison to Heinzerling et al. [8], we test paraphrasing motivated by linguistics. Additionally, we open the field for new probing-tasks [29], i.e. how sentence processing [11] impacts knowledge inference. Thus, we gain insight into information encoding and potential directions for knowledge enhancement strategies.

3 PRELIMINARY EXPERIMENTS

3.1 Data

In this work, we propose that utilizing cloze-text prompts offers a direct means of studying the impact of syntactic features on knowledge retrieval in language models. Knowledge capturing

[1] https://github.com/Thrasolt/ContextualKnowledgeOfLMs

Model	Simple	Compound	Complex	Compound-complex
BERT-large-cased	**30.22**	16.28	16.99	17.99
BERT-base-cased	**28.19**	12.40	12.95	15.77
BERT-base-multilingual-cased	**19.99**	13.40	13.39	10.97
BERT-large-uncased	**3.38**	1.10	1.20	0.47
BERT-base-uncased	**3.07**	0.75	1.54	0.77
BERT-base-multilingual-uncased	**3.50**	0.60	1.83	0.39

Table 2: Knowledge Retrieval Model Comparison with T-REx Data Set for average top-1 metric in percent (#Triples=34039)

Model	Simple	Compound	Complex	Compound-complex
BERT-large-cased	**58.48**	41.76	43.73	42.82
BERT-base-cased	**57.74**	37.32	39.82	38.96
BERT-base-multilingual-cased	**39.29**	33.36	30.68	32.17
BERT-large-uncased	**11.34**	6.51	6.38	5.14
BERT-base-uncased	**9.33**	6.78	6.07	5.36
BERT-base-multilingual-uncased	**9.40**	4.31	5.29	3.54

Table 3: Knowledge Retrieval Model Comparison with T-REx Data Set for Average top-10 Accuracy in percent (#Triples=34039)

Cardinality	#Triples	#Relations
1:1	937	2
N:1	20006	23
N:M	13096	16
Total	34039	41

Table 4: Properties of the T-REx [23]

prompt-templates were first used in the LAMA-probe [23]. Those templates enable the parsing of a subject and object tokens to form a correct sentence. In the test prompts, the mask token replaces the correct object-token. Thus, the model tries to predict the correct object for a given prefilled prompt. We give an example of such samples here for relations ((1) P36, (2) P108, (3) P136):

Template:
(1) "The capital of [S] is [O] ."
(2) "[S] works for [O] ."
(3) "[S] plays [O] music ."
Prompt:
(1) "The capital of France is [MASK]."
(2) "Tim Cook works for [MASK]."
(3) "Bruno Mars plays [MASK] music."
Parsed:
(1) "The capital of France is Paris."
(2) "Tim Cook works for Apple."
(3) "Bruno Mars plays funk music."

The T-REx subset of the LAMA-Probe relies on the T-REx knowledge base [4] derived from Wikidata triples. The 34039 triples are organized into 41 different Wikidata relations. For each relation, no more than 1000 facts are sub-sampled. All relations have a maximum of 995 and a minimum of 225 facts, with most relations specifying more than 900 facts. The 41 relations cover all possible cardinality types $1:1$, $N:1$ and $N:M$.

3.2 Task & metrics

In this work, we have limited our typological paraphrasing to the T-REx triples (and corresponding relations) of the LAMA-probe. Given a set of four syntactical typologies T, and a set of subject-relation-object triples $< s, r, o >$ named D, we transform D in a set of tuples $D_t = \{< p_s^r, o > \mid < s, r, o > \in D\}$. This is achieved by describing each r through a prompt written with the typology t named p^r. We can use this prompt to parse s so that we obtain p_s^r, o is the cloze-prediction target. Given such a set D_t, we measure the performance of a model m for typology t, by calculating the top-k accuracy for all tuples in D_t.

$$\text{top-k}_t \text{ accuracy} = \frac{\sum_{(p_s^r, o) \in D_t} \mathbb{1}(o \in \text{top-k}_m(p_s^r))}{|D_t|} \tag{1}$$

Where o is the correct label, top-k$_m$ are the k predictions with the highest probability assigned by the model m, $|D_t|$ is the number of samples, and $\mathbb{1}$ is the indication function. In our results, we consider $top\text{-}\{1, 10\}_T$ *accuracy*. This can be noted that $top\text{-}1_{simple}$ *accuracy* is virtually equal to the evaluation conducted by Petroni et al. [23] with the $P@1$ metric.

3.3 Prompt Engineering

The LAMA-probe contains a simple sentence template for each of the 41 relations in the T-REx data. The fitness of these templates is manually improved and tested by Petroni et al. [23]. Therefore, they represent a natural starting point for our syntactically motivated prompt paraphrasing. Expanding this template to different syntactical structures offers insight into the impact of such a transformation on the same knowledge task. We have used four typological transformations, one of which is the same as the LAMA template. Thus, we create three new prompt templates for each of the 41 relations in T-REx. The resulting four templates provide a unique grammatical structure, which is theoretically guided by the research of Rodney Huddleston [10]. In our sentence typology, a *simple* sentence defines

Results	Simple	Compound	Complex	Compound-Complex	#Triples
Total	**30**.22	16.59	16.08	17.87	34039
1:1	**70.65**	69.48	67.56	58.16	937
N:1	**35.07**	18.36	18.18	20.01	20006
N:M	**18.77**	8.61	10.65	11.24	13096

Table 5: top-1 Accuracy in Percent of Bert-Base-Cased on the T-REx Data Set

Results	Simple	Compound	Complex	Compound-Complex	#Triples
Total	**58.48**	42.18	42.75	43.09	34039
1:1	**85.17**	84.95	84.31	81.32	937
N:1	**65.96**	43.49	48.47	45.40	20006
N:M	**43.57**	36.62	29.70	36.18	13096

Table 6: top-10 Accuracy in Percent of Bert-Large-Cased on the T-REx Data Set

a sentence that contains only one main clause (LAMA-probe templates). A sentence that includes two or more independent clauses is known as a *compound* sentence, while a sentence that contains an independent clause and one or more dependent clauses is known as a *complex* sentence. Lastly, a sentence that includes two or more independent clauses and at least one dependent clause is known as a *compound-complex* sentence. Table 1 shows an example of the four templates for the 1:1 relation P36, describing the predicate 'capital of', and the M:N relation P530, describing the 'diplomatic relation with'.

3.4 Results

We applied these template-based-prompts to three BERT variants: BERT-large, BERT-base, and BERT-base-multilingual. We include multilingual BERT to understand the impact of named entity mentions in different languages. Our experiments show that all investigated LLMs perform best on the simple sentence type. Additionally, we discover that the cased models outperform the uncased models by a large margin.

Table 2 shows the top-1 accuracy for each model in percent for all four sentence types. We report slightly worse results for the top-1_{simple} accuracy (BERT-base-cased -3.0, BERT-large-cased -1.1) on the LAMA-probe than in the original paper [23]. In contrast to Petroni et al. [23], we consistently evaluated over the whole vocabulary, which had a notable influence on the reliability of the results for the N:M relations. Specifically, Petroni et al. [23] exclude all other valid entities except for the one they test. Nonetheless, our results are reasonably close, given different reported results on the same data in other works [34]. Generally, the values of the correct tokens are surprisingly high. The best model was able to predict one-third of the masked tokens correctly. However, most comparable results achieved by the cased models are around 15 to 20 percent accurate. Most importantly, the average top-1 accuracy varies significantly between different sentence types. Thus, indicating grammatical structure influences a model's ability to retrieve relational knowledge. This is true for all models under investigation.

From this, we draw four conclusions: First, the BERT-large-cased model outperforms all other models on all four sentence types by at least two and at most four percentage points. Second, there is a chasm in performance between cased and uncased models, as the accuracy of uncased models is comparatively low. Third every model has a higher prediction accuracy when queried with the simple sentence compared to the other three types. Finally, the differences in scores among the non-*simple* sentence types are significantly lower than the variations within the *simple* sentence type. These observations also apply to the results based on the top-10 accuracy, albeit with the expected higher accuracy results, see Table 3.

Table 5 and Table 6 show the average accuracy results for the four sentence types for each of the cardinality relations for top-1 and top-10 for the BERT-large-cased model. Both results show that the *simple* sentence type enables a higher accuracy for all three cardinalities. Additionally, in both sets of results, the performance decreases with increasing cardinality, which is intuitive, as the difficulty level increases with the number of possible subjects and objects. For N:M relations, top-1 is an inappropriate metric, as only one guess is allowed per subject.

The results are the closest for the cardinality 1:1 and furthest apart for N:M, thus implying that the relation extraction works best for simple sentence types and simple relations (1:1). The performance noticeably decreases when either sentence or relation complexity increases. Additionally, the sentence structure (typology) has close to no influence on the top-10 performance for the simple relations (1:1). However, the relations with less mutual information between subject and object co-occurrence (N:1, M:N) show a large decrease in performance for changes in the sentence-typology. Hence, the MLM-task does not incorporate the rules of syntactical change while keeping semantic equivalence.

4 CONCLUSION AND FUTURE WORK

In this paper, we investigate the impact of prompt syntax on the knowledge retrieval performance of LLMs. To achieve this, we expand the well-known and commonly used T-REx subset of the LAMA-probe to support different syntactical structures of prompts. our preliminary results show, that the impact of syntax is only marginal for simple relations (1:1). In general, simple prompts should be the preferred way of querying. Most importantly, we show that

LLMs indeed struggle to generalise knowledge across grammatical structures. This finding highlights the importance of the relationship between syntax and semantics within LLMs as a crossroad of human and machine language representation. Consequently, we will focus on a deeper analysis of the disparities in information coding for typologically different templates. These disparities may be reflected in the attention mechanism [1], the predicted token-distribution [3] or the differences in mask representation among the various typologies per relation.

REFERENCES

[1] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341* (2019).

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[3] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics* 9 (2021), 160–175.

[4] 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)*.

[5] Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. *arXiv preprint arXiv:2004.07202* (2020).

[6] Daniel Gao, Yantao Jia, Lei Li, Chengzhen Fu, Zhicheng Dou, Hao Jiang, Xinyu Zhang, Lei Chen, and Zhao Cao. 2022. KMIR: A Benchmark for Evaluating Knowledge Memorization, Identification and Reasoning Abilities of Language Models. *arXiv preprint arXiv:2202.13529* (2022).

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

[8] Benjamin Heinzerling and Kentaro Inui. 2020. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. *arXiv preprint arXiv:2008.09036* (2020).

[9] Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of ACL*. Association for Computational Linguistics.

[10] Rodney Huddleston. 1984. *Introduction to the Grammar of English*. Cambridge University Press.

[11] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language?. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

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

[13] Jan-Christoph Kalo and Leandra Fichtel. 2022. KAMEL: Knowledge Analysis with Multitoken Entities in Language Models. In *Proceedings of the Conference on Automated Knowledge Base Construction*.

[14] Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are Pretrained Language Models Symbolic Reasoners Over Knowledge? *arXiv preprint arXiv:2006.10413* (2020).

[15] Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Pre-trained Language Models as Symbolic Reasoners over Knowledge?, In Proceedings of the 24th Conference on Computational Natural Language Learning. *CoRR*.

[16] Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*. 4725–4735.

[17] Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2019. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. *arXiv preprint arXiv:1911.02493* (2019).

[18] Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2019. Specializing Unsupervised Pretraining Models for Word-Level Semantic Similarity. arXiv:1909.02339 [cs.CL]

[19] Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2019. Sensebert: Driving some sense into bert. *arXiv preprint arXiv:1908.05646* (2019).

[20] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504* (2019).

[21] Matthew E Peters, Mark Neumann, Robert L Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. 2019. Knowledge enhanced contextual word representations. *arXiv preprint arXiv:1909.04164* (2019).

[22] Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How Context Affects Language Models' Factual Predictions. arXiv:2005.04611 [cs.CL]

[23] 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)*. ACL, Hong Kong, China, 2463–2473. https://doi.org/10.18653/v1/D19-1250

[24] Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? *arXiv preprint arXiv:2002.08910* (2020).

[25] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics* 8 (2021), 842–866.

[26] Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-Informed Self-Attention for Semantic Role Labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 5027–5038. https://doi.org/10.18653/v1/D18-1548

[27] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223* (2019).

[28] Dhanasekar Sundararaman, Vivek Subramanian, Guoyin Wang, Shijing Si, Dinghan Shen, Dong Wang, and Lawrence Carin. 2021. Syntactic Knowledge-Infused Transformer and BERT models. In *CEUR Workshop Proceedings*, Vol. 3052. CEUR Workshop Proceedings.

[29] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316* (2019).

[30] Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. *arXiv preprint arXiv:2005.05635* (2020).

[31] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808* (2020).

[32] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics* 9 (2021), 176–194.

[33] Chaoqi Zhen, Yanlei Shang, Xiangyu Liu, Yifei Li, Yong Chen, and Dell Zhang. 2022. A Survey on Knowledge-Enhanced Pre-trained Language Models. *arXiv preprint arXiv:2212.13428* (2022).

[34] Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual Probing Is [MASK]: Learning vs. Learning to Recall. In *North American Association for Computational Linguistics (NAACL)*.