Dissecting Paraphrases: The Impact of Prompt Syntax and Supplementary Information on Knowledge Retrieval from Pretrained Language Models

Stephan Linzbach ♢♠, Dimitar Dimitrov♢, Laura Kallmeyer♠,
Kilian Evang♠, Hajira Jabeen♢*, and Stefan Dietze♢♠*
♢ GESIS - Leibniz Institute for the Social Sciences, name.surname@gesis.org
♠ Heinrich Heine University, name.surname@hhu.de

Abstract

Pre-trained Language Models (PLMs) are known to contain various kinds of knowledge. One method to infer relational knowledge is through the use of cloze-style prompts, where a model is tasked to predict missing subjects or objects. Typically, designing these prompts is a tedious task because small differences in syntax or semantics can have a substantial impact on knowledge retrieval performance. Simultaneously, evaluating the impact of either prompt syntax or information is challenging due to their interdependence. We designed ConPARE-LAMA – a dedicated probe, consisting of 34 million distinct prompts that facilitate comparison across minimal paraphrases. These paraphrases follow a unified meta-template enabling the controlled variation of syntax and semantics across arbitrary relations. ConPARE-LAMA enables insights into the independent impact of either syntactical form or semantic information of paraphrases on the knowledge retrieval performance of PLMs. Extensive knowledge retrieval experiments using our probe reveal that prompts following clausal syntax have several desirable properties in comparison to appositive syntax: i) they are more useful when querying PLMs with a combination of supplementary information, ii) knowledge is more consistently recalled across different combinations of supplementary information, and iii) they decrease response uncertainty when retrieving known facts. In addition, range information can boost knowledge retrieval performance more than domain information, even though domain information is more reliably helpful across syntactic forms.

1 Introduction

Symbolic knowledge bases provide relational knowledge and are widely used for tasks like question-answering. However, they rely on costly manual or automated, often supervised, information extraction pipelines to retrieve and represent relational knowledge. Relational knowledge refers to knowledge about relations between entities, e.g. 'Paris', 'capitalOf', 'France', where 'capitalOf' is the *relation*, 'Paris' is the *subject*, and 'France' is the *object*. Previous research on relational knowledge retrieval (*rKR*) from pre-trained language models (PLMs) (Petroni et al., 2019; Sung et al., 2021) has demonstrated that relational knowledge can be retrieved directly from the parameters of a PLM. This finding has led to a plethora of research concerned with knowledge retrieval and reasoning capacities of PLMs (Petroni et al., 2019; Sung et al., 2021; Zhong et al., 2021; Elazar et al., 2021; Jiang et al., 2019), where *rKR* performance is seen as an indicator of PLM's capacities to understand and reason. Several benchmarks have been proposed that aim at measuring *rKR* performance as the ability of a PLM to predict masked objects as part of cloze-style prompts (Petroni et al., 2019; Kalo and Fichtel, 2022; Kassner et al., 2021). It was found that some types of supplementary information (sInf) are helpful to PLMs (Cao et al., 2021; Petroni et al., 2020; Chen et al., 2022) while other types deteriorate knowledge retrieval performance (Pandia and Ettinger, 2021; Kassner and Schütze, 2020), and that PLMs primarily rely on memorization, hence, low-frequency examples are less well remembered (Ravichander et al., 2020). Additionally, prior works have shown that *rKR* through prompts is inconsistent across different paraphrases (Elazar et al., 2021; Heinzerling and Inui, 2020).

Paraphrasing a prompt may introduce a variety of changes, including **semantic** ones that change the information content of the prompt, i.e., *domain* information ('Paris is a city and is the capital of [MASK]') or *range* information ('Paris is the capital of [MASK], which is a country.'), as well as **syntactic** ones that merely change the form in which the same content is expressed i.e., *clausal* ('Paris

* Corresponding Author

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 3645–3655

is a city and is the capital of [MASK]') or an *appositive* syntax ('The city Paris is the capital of [MASK]').

Previous works have not evaluated the combined impact of syntax and semantics or struggled to control all involved variables (Linzbach et al., 2023; Elazar et al., 2021; Heinzerling and Inui, 2020).

Thus, dedicated probes are required that can control the effects of different syntactic and semantic realisations on *rKR*.

Our main contributions include:

Controlled Paraphrasing. We apply a meta-template that streamlines prompt engineering across arbitrary relations while enabling control over syntactic form and semantic content (§3). In terms of syntactic form, in English single-sentence prompts, information must be added either in the form of a noun phrase appositive or as an additional clause. The meta-template covers prompts with clausal (as compound, complex, or compound-complex) and appositive syntax. Additionally, it includes placeholders for sInf as domain or range information or both. In our case, automated prompt construction is enabled by fetching domain/range information (as sInf) for given relations from established knowledge bases such as Wikidata. Given this method, we can compare paraphrases focused on particular combinations of semantics and syntax while controlling for variables not under investigation. **Probe and benchmark (**ConPare-LAMA**).** We introduce ConPare-LAMA[1], a novel **Con**trolled **Par**aphrasing Prob**e** for **LAMA** (§3.4), that is to the best of our knowledge the first *rKR* probe that facilitates extensive experiments by controlling for both syntax and semantics of prompts and is the largest *rKR* probe so far publicly released. We investigate a set of 60 relations, derived from the established LAMA-probe, ensuring wide comparability with other knowledge retrieval research. More specifically, we utilise the TREx, GoogleRE and ConceptNet corpora from LAMA as described in Petroni et al. (2019). For each relation, using our meta-template, we generate prompts where sInf is added to the subject, the object, both, or neither. The sInf is realized with different syntax, resulting in a total of seven prompts per relation. Varying the sInf obtained from Wikidata for each such prompts results in roughly 34 million prompts (TREx: 7 mio, ConceptNet: 26 mio, GoogleRE: 1 mio) unique prompts contained in ConPare-LAMA. **Experiments and findings.** We conduct experiments on the base versions of three well-established PLMs, i.e., BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019b) and Luke (Yamada et al., 2020) to advance the understanding of *rKR* from PLMs. In particular, we investigate the following research questions. [RQ 1:] What is the impact of prompt syntax and information on *rKR* performance? [RQ 2:] What is the impact of syntax on the PLMs' ability to efficiently combine sInf in prompts?; [RQ 3:] How consistent are the answers of PLMs when comparing the set of correctly retrieved samples for different prompt syntax and information content?; [RQ 4:] How does prompt syntax impact the response uncertainty of PLMs? We find that all models perform better on prompts using sInf through clausal syntax, on all investigated corpora, as compared to information added via appositives. BERT achieves the best performance in this case. In addition, knowledge retrieved through prompts that rely on clausal syntax when adding sInf is more consistent given assumptions about the a priori knowledge available in the prompt. Underlining these findings is our observation that the uncertainty of models for responses to *known facts* decreases when adding sInf through clausal syntax, which is not the case for appositive syntax.

2 Related Work

In this section, we provide the necessary background and discuss prior work.

Knowledge in PLMs. Since the proposal of transformer-based PLMs (Devlin et al., 2018), huge efforts have been spent to analyse the knowledge encoded in the learned representations (Rogers et al., 2021). All conceivable types of knowledge are tested. (i) Syntactic and general linguistic knowledge (Ettinger, 2020; Hewitt and Manning, 2019; Liu et al., 2019a; Htut et al., 2019; Goldberg, 2019; Tenney et al., 2019; Clark et al., 2019), where Swayamdipta et al. (2019) show that PLMs do not benefit from shallow syntactical features. Reif et al. (2019) show correlations between the gold standard dependency tree and the attention mechanisms in PLMs. Hewitt and Manning (2019) show for BERT (Devlin et al., 2018) that syntax trees are consistently embedded by the neural network. (ii) Performance in knowledge driven tasks (Bosselut et al., 2019; Radford et al., 2019; Da and Kasai, 2019;

[1] https://github.com/Stephan-Linzbach/ConPare-LAMA

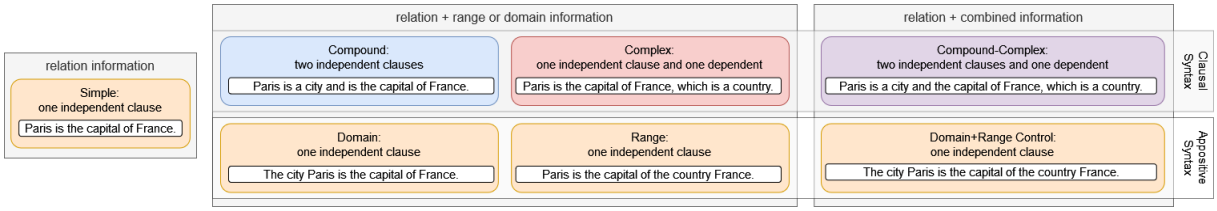


Figure 1: Relationship between prompt types (simple (orange), compound (blue), complex (red), compound-complex (purple)), syntactic forms (clausal and appositive), and sInf combinations (relation, relation+range or domain, relation+combined) used to study the influence of syntax on knowledge retrieval.

Talmor et al., 2020; Warstadt et al., 2019; Sung et al., 2021; Gao et al., 2022). (iii) Investigation of relational knowledge inherent in PLMs through the LAMA-probe (Petroni et al., 2019). This probe contains knowledge-targeted cloze-style prompts for querying of different models. The idea, behind the LAMA-probe motivated researchers to investigate multilingual knowledge (Kassner et al., 2021), knowledge about entities with more complex naming (Kalo and Fichtel, 2022), and the impact of prompt phrasing on retrieval performance (Jiang et al., 2020).

Consistency of PLMs. Research concerned with the consistency analyses the answer space of PLMs queried for the same fact through various cloze-style prompts (Heinzerling and Inui, 2020; Elazar et al., 2021). Testing the consistency of PLMs regarding negation and mispriming Kassner and Schütze (2020) showed that PLMs are mostly insensitive to the notion of negation and distracted by mispriming. The latter finding was additionally strengthened by Pandia and Ettinger (2021), Misra et al. (2020), and lately confirmed for LLMs (Shi et al., 2023) where irrelevant context was used to distract 'code-davinci-002' from the GPT3 family. In comparison to Jiang et al. (2020) that used paraphrases to investigate peak knowledge, and research that investigated semantic perturbation (Kassner and Schütze, 2020; Misra et al., 2020; Pandia and Ettinger, 2021), Elazar et al. (2021) introduced the PARAREL probe with which they investigate consistency of language models across prompt paraphrases. They conclude that the models have a generally low consistency. Considering, the justified distrust in PLM *rKR* performance Petroni et al. (2020) and Cao et al. (2021) try to understand the impact of helpful information in prompts. They find that a wide array of sInf helps the models to increase retrieval performance. Thus, motivating the research in the field of prompt engineering Hu et al. (2021), KnowPrompt (Chen et al., 2022), and an Ontology based proposal by Ye et al. (2022).

Our research. Our work builds upon the ideas proposed by the LAMA-probe. However, we study the influence of prompt paraphrases by controlling for syntax and semantic change on *rKR* performance, not the general capacity of PLMs (Zhong et al., 2021; Petroni et al., 2019). Whereas Cao et al. (2021) investigate the impact of an array of sInf on the models' performance, we study how sInf is differently incorporated depending on syntax. In contrast to Elazar et al. (2021) we measure consistency of PLMs in different syntactic and semantic scenarios.

3 Controlled Paraphrasing

We propose CONPARE-LAMA (**Con**trolled **Par**arphrasing Prob**e** for **LAMA**) to investigate how syntax and semantics of paraphrases impact knowledge retrieval performance of PLMs. We hypothesize that certain syntactic forms facilitate correct interpretation, while certain semantic additions are more useful than others. As we are working with relations, we found that a natural addition of information would be domain and range type constraints. Using this as sInf, we hardly change the semantics of the original task of *rKR*. Furthermore, we restricted our work to single-sentence English prompts. We can classify all such sentences as either *clausal* (cf. Fig. 1, upper row) or *appositive* syntax (cf. Fig. 1, lower row). Additionally, we are interested in the sentence's overall shape, as reflected by the classification of sentences by traditional grammars as *simple* (orange), *compound* (blue), *complex* (red), or *compound-complex* (purple) (cf. Huddleston, 1984). We introduce our probe in three steps: first we *control prompt syntax and semantic effects* (§3.1) by selecting prompt sentence types and sInf, then we describe the *meta-template definition* (§3.2), and lastly we show the automatic *template instantiation* (§3.3).

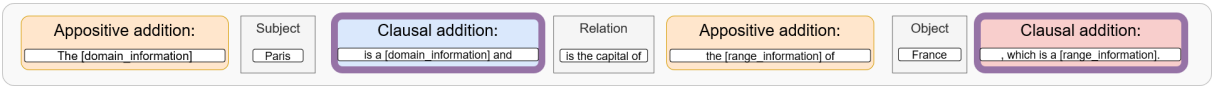


Figure 2: Meta-template that facilitates comparable prompt creation for various relations and information demands.

Relation	LAMA	Ours
Occupation	[S] is a [MASK] by profession.	[S] has occupation [MASK].
Native Language	The native language of [S] is [MASK].	[S] natively speaks [MASK].
HasProperty	The [world] today is getting more and more [MASK].	[S] can be described as [MASK].

Table 1: LAMA prompts with high syntactical variation vs. our schematic prompt style.

3.1 Dissecting Paraphrasing

Previous work has shown that subtle changes in the phrasing of the prompt can have a substantial impact on model predictions (Jiang et al., 2020; Elazar et al., 2021). However, changes of the semantic information in a prompt are not yet studied independently of their syntactic realisation and vice versa. We aim to isolate the effects of syntax and semantic change on *rKR*. For this, we control the the syntactic realisation of the prompt while varying the sInf. As we test several syntactic realisation we can also observe the impact of semantic change.

Prompt sentence types. Our prompt construction starts from the *simple* type (orange box in the upper left of Fig. 1). It consists of one main clause that encodes the basic knowledge triple ('Paris is the capital of France'). Contrast this with the *clausal* types that add supplementary domain information in the form of a coordinated clause (blue, 'Paris is a city and is the capital of France') or range information in the form of a subordinated clause (red, 'Paris is the capital of France, which is a country'), or both (purple, 'Paris is a city and is the capital of France, which is a country'). Following Huddleston (1984), we call these clausal prompt types *compound*, *complex*, and *compound-complex*, respectively. We used coordination (compound) for subjects and subordination (complex) for objects because these forms were deemed to be the smallest natural-sounding clausal additions that contain the respective domain/range information (i.e., using coodination for object would require duplicating the masked object *Paris is the capital of [MASK] and [MASK] is a country*. Analogously, prompts with *appositive* syntax are used to add the equivalent sInf (orange boxes in the lower row).

Supplementary Information. As we intend to assess the slot-filling performance on prompts with syntax going beyond the simple syntax used in LAMA-probe, we use sInf (e.g. city, European country etc.) to construct complex or compound prompt types. To fetch this information, we utilise domain and range type constraints for the objects and subjects in respective relations. For example, in Wikidata the relation 'capital' (P:36[2]) has the domain type restriction[3]: 'area', 'geographic region', and 'fictional planet' etc., and the range type restriction[4]: 'political territorial entity', 'fictional city', and 'capital city' etc. This enables us to dynamically generate paraphrases with sInf.

3.2 Meta-Template Definition

The LAMA-probe tests the assumption that PLMs can function as knowledge bases by manually crafting prompts. These prompts are written to achieve reasonable retrieval performance with no focus on syntactic features. Furthermore, prompts for several corpora are written to query a single knowledge triple. We, however, are interested in *syntactic comparability* between all prompts. Therefore, we introduce a meta-template (cf. Fig. 2). In our meta-template, the grey boxes (i.e., subject, relation, object) are mandatory, while the colored boxes (orange, blue, red) are prompt type specific additions. Purple indicates the combination of blue (compound) and red (complex) to form a sentence following the compound-complex typology (analogue for appositive). Applying this meta-template avoids confounding effects of relation-specific syntactic forms on retrieval performance. We manually crafted a natural-language encoding of each relation that fits the meta-template. A brief comparison of ConPare-LAMA and LAMA-probe can be seen in Tab. 1. Note that the prompt for relation ('HasProperty') is uniquely written for the triple ('world', 'HasProperty', 'complicated') in the LAMA-probe. Moreover, our meta-template assumes that we can use the same template for all triples of one relation and that relation and object

[2]https://www.wikidata.org/wiki/Property:P36
[3]https://www.wikidata.org/wiki/Q21503250
[4]https://www.wikidata.org/wiki/Q21510865

text always remain in the same main clause.

3.3 Template Instantiation

To instantiate our meta-template, we propose two completion strategies for selecting supplementary (range/domain) information:

Quality Completion. - choosing the information that leads the model to predict the *right* token with the highest probability.

Confidence Completion. - choosing the information that leads the model to predict *any* token with the highest probability.

3.4 ConPare-LAMA

We adapt the LAMA-probe with our controlled probe design to introduce CONPARE-LAMA (**Con**trolled **Par**aphrasing Prob**e** for **LAMA**). Domain and range type constraints depend on relations. Hence, CONPARE-LAMA contains only triple-based LAMA probe corpora (i.e. TREx, GoogleRE, ConceptNet). All corpora are reduced to a comparable size, meaning only triples where the object is in the token vocabulary of all studied models are considered (cf. CONPARE-LAMA statistics in Tab. 2). We manually expressed all used relations in natural language statements, in a minimal fashion to fit the meta-template. For the 41 Wikidata relations available in TREx, we queried the domain[5] and range[6] type constrains from Wikidata. For five relations there is no sInf available. However, we manually define one type constraint for the range and the domain for those five relations to ensure that all prompts can be instantiated. To get the domain and range information for GoogleRE, we translated the given relations to their Wikidata counter-part. For the 16 relations in ConceptNet, we mapped our manually chosen type constraints to their noun concept in ConceptNet. From those concepts we inferred domain and range type constraints using all concepts connected to the seed concept via 'related to' or 'defined by' relations.

4 Experiments

We use base models of three different PLMs for evaluation: BERT (Devlin et al., 2018) as it is a well established model and it has already been assessed using the LAMA probe, RoBERTa (Liu et al., 2019b) as it is shown to be superior in performance on a range of downstream tasks when compared to BERT, and Luke (Yamada et al., 2020) as it uses entity word cross-attention to enhance the knowledge of a base RoBERTa model. However, to keep it comparable, we only input the tokenized text without entity span information. If not mentioned explicitly we complete the prompts that either add domain or range information with *Quality Completion* and reuse this information to populate the prompt that encodes both kinds of information. This is done to ensure best possible performance per prompt and triple. We use the P@1 metric to measure *rKR* performance following Petroni et al. (2019). We only conduct our experiment on base models as consistency concerning paraphrases only marginally increases from base to large configurations of the PLMs (Elazar et al., 2021).

Corpus	Grouping	#Relations	#Facts	Dom	Rng
TREx	1:1	2	651	6.5	5.5
	N:1	23	18682	9.5	6.4
	N:M	16	10190	15.6	10.1
	Total	41	29523	11.6	7.7
GoogleRE	death place	1	649	10	10
	birth place	1	2404	7	8
	birth date	1	1565	16	1
	Total	3	4618	11	6.3
ConceptNet	Total	16	22739	13.2	12.6

Table 2: CONPARE-LAMA corpora statistics with mean number of available object domain (Dom) and range (Rng) types per relation as defined in Wikidata.

4.1 Results

Impact of syntax and semantic on knowledge retrieval performance (RQ1). We organize the process of paraphrasing across two dimensions, the semantic dimension (i,e., changing sInf), and the syntactic dimension (i.e., changing the sentence type) where we consider *clausal* (compound=Cpnd, complex=Cplx) and *appositive* syntax (Appo). Table 3 offers three perspectives on the impact of paraphrasing (from top to bottom) (1) observing performance change across semantic paraphrases (Relation, Relation+Domain Information, Relation+Range Information), (2) the impact of paraphrasing per model (BERT, RoBerta, Luke) and (3) performance change for syntactic paraphrasing of semantically equivalent content (Simple, Cpnd, Cplx, Appo). We can measure the impact of semantic change by comparing the best performance per semantic category per model. We observe that sInf through prompts using *clausal* (compound=Cpnd,

[5]https://www.wikidata.org/wiki/Q21503250
[6]https://www.wikidata.org/wiki/Q21510865

Corpus	Grouping	Relation			Relation + Domain Information						Relation + Range Information					
		BERT	RoB	Luke	BERT		RoBERTa		Luke		BERT		RoBERTa		Luke	
		Simple			Cpnd	Appo	Cpnd	Appo	Cpnd	Appo	Cplx	Appo	Cplx	Appo	Cplx	Appo
TREx	1:1	.4439	.3118	.3394	.6405	.5760	.5238	.6098	.5330	.5944	.6205	.6052	.5775	.5238	.5668	.5176
	N:1	.2876	.1956	.2240	.3329	.3086	.2706	.2316	.2792	.2500	.3656	.2919	.3547	.2722	.3652	.2496
	N:M	.2517	.2205	.2401	.3217	.3261	.2986	.3166	.3168	.3105	.3488	.1871	.2979	.2121	.3015	.1879
	Total	**.2786**	.2067	.2321	**.3358**	.3205	.2859	.2693	.2978	.2785	**.3654**	.2627	.3400	.2570	.3477	.2342
GoogleRE	birth-date	.0	.0012	.0191	.0	.0	.0178	.0364	.0319	.0428	.0044	.0	.0083	.0031	.0031	.0031
	birth-place	.1738	.1156	.0183	.2129	.1855	.0994	.0715	.0590	.0345	.2254	.2029	.1863	.1730	.2104	.1988
	death-place	.1479	.0061	.0015	.1479	.1263	.0061	.0077	.0154	.0077	.1571	.1771	.1587	.1510	.1879	.1448
	Total	**.1113**	.0614	.0162	**.1316**	.1143	.0586	.0506	.0437	.0335	**.1409**	.1305	.1221	.1123	.1370	.1249
ConceptNet	Total	**.0229**	.0226	.0230	.0455	.0390	**.0512**	.0463	.0494	.0507	.0495	.0084	**.0586**	.0098	.0484	.0088

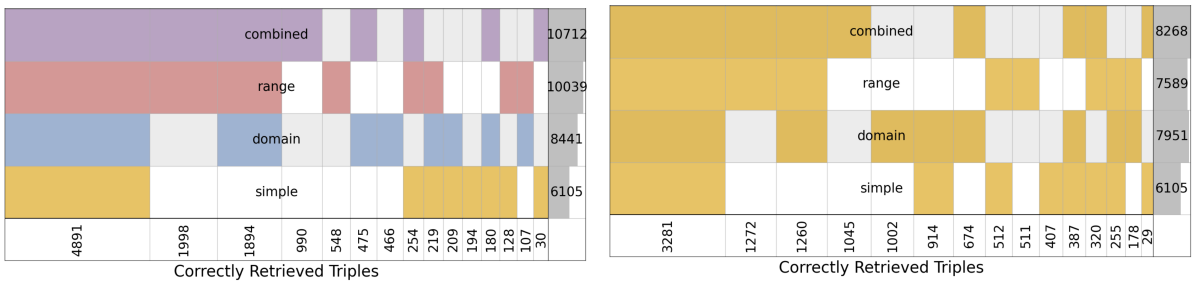
Table 3: Performance (P@1) when querying with the base typologies and respective appositive. Underline indicates per row and model winner of either the clausal or the appositive prompt. Bold indicates the best performance across all models per corpus and available information.

complex=Cplx) syntax increases the performance for all three models on all corpora. For the *appositive* syntax (Appo), this is generally true, with a few exceptions, though the increase is relatively less across relations than with clausal syntax, except for Luke on ConceptNet. Furthermore, we can see that the performance given domain information is more stable across different syntaxes when compared to the addition of range information. However, adding range information has the most potential to increase performance. When we compare this across different models we observe a common trend in all of them. Lastly, we analyze the impact of syntactic paraphrasing. In general *clausal* prompts outperform their *appositive* counterpart. In particular for TREx, the average performance gain from sInf is weaker for compound prompts ($\approx 2\%$) than for complex prompts ($\approx 10\%$) when compared to their respective appositive counterparts. Expanding upon the findings presented by Petroni et al. (2020) and Cao et al. (2021), we show that supplementing range and domain information is helpful for *rKR*.

Impact of prompt syntax on efficient information combination (RQ2). We investigate if the performance differences between causal and appositive prompt syntax persist when we query with prompts that carry both range and domain information. Furthermore, we intend to assess if such prompts help the PLM to uncover synergies or act as noise. For this reason, we define a relative performance interval from the already observed *rKR* performance on compound and complex prompts (analog for appositive). The low-end of performance is marked by choosing the answer with the highest confidence. The high-end of performance is estimated by choosing the answer to a prompt that signals the highest probability for the correct token. The results of this experiment are displayed in Figure 4. A model capable of combining information given a prompt (purple or orange) would be between the bounds (black-part). If, given a prompt with both sInf, a model performs below the lower bound (grey-part), it is not able to recover the knowledge it previously displayed, inferring that the encoding of the prompt was noisy. As the PLM's response is determined by probability, exceeding the black part is only possible in highly unlikely data constellations. From Fig. 4, we can observe that appositive syntax is less often in the expected boundaries. Additionally, neither clausal nor appositive syntax enable all PLMs to reliably combine the information such that it performs above the lower bound (in the black-part). However, RoBERTa is able to combine information reliably for clausal syntax. In comparison, the behaviour is less reliable for appositive prompts. All models have a similar peak performance on the complete TREx corpus. These findings expand the conclusion of Pandia and Ettinger (2021) that even potentially helpful information could be detrimental for *rKR* performance.

Impact of prompt syntax on knowledge consistency with regards to different levels of available information (RQ3). This analysis is only conducted on the TREx corpus as the sample size of correctly predicted samples given simple prompts is sufficiently large to support conclusive insights (GoogleRE: 277, ConceptNet: 513). We display the results as a multi-set Venn diagram[7] for RoBERTa in Figure 3 and report the numbers for the remaining models in the text. We can read the diagram from no sInf (bottom) to high sInf (top) content. Each row indicates the correctly predicted triples given a specific prompt/syntax type. Each column represents an intersection (subset) of triples

[7] https://github.com/gecko984/supervenn



(a) Clausal prompts

(b) Appositive prompts

Figure 3: Knowledge consistency for sInf added through (a) clausal and (b) appositive prompts for all intersections of correctly predicted triples by RoBERTa on the TREx corpus.

known to different prompt types. Every intersection (column) has two properties: the size of the respective intersection (bottom row), and which prompt types intersect (coloring per cell, white means no intersection). Lastly, the total cardinality (correct triples) for each prompt type is aggregated at the end of each row. Note that along the rows an upside-down staircase pattern should emerge indicating that less sInf offers worse performance and more sInf offers more information while retaining the already retrieved knowledge.

We make the following observations: (i) Looking at the leftmost column of Fig. 3b, for RoBERTa (R), only 54% (BERT (B): 51%, Luke (L): 51%) of the triples correctly retrieved by the simple prompt are also correctly retrieved by all three appositive prompts ($3281/6105 \approx 54\%$). We achieve much higher consistency with clausal prompts (R: $4891/6105 \approx 80\%$, B: 79%, L: 78%, Fig. 3a). (ii) The number of knowledge triples retrieved only through the simple prompt is twice as big for appositive (R: 407, B: 476, L: 529) vs. clausal prompts (R: 194, B: 368, L: 256), implying that information is more distracting when added with appositive syntax. R recalls 86% (B: 85%, L: 86%) of all triples known with the simple prompt through the compound-complex prompt. The consistency is worse for the appositive syntax, where R recalls 62% (B: 58%, L: 62%) with the combined appositive prompt, which equals a decrease in recall of 24% compared to the clausal case (B: 27%, L: 24%). We can see that, in general, the appositive syntax performs worse and is less consistent when compared to clausal prompts. Our results agree with Elazar et al. (2021) that there are indeed substantial differences in consistency and performance between different paraphrases.

Impact of prompt syntax on response uncertainty (RQ4). Here, we study the impact of syntax on the response uncertainty in PLMs by plotting the

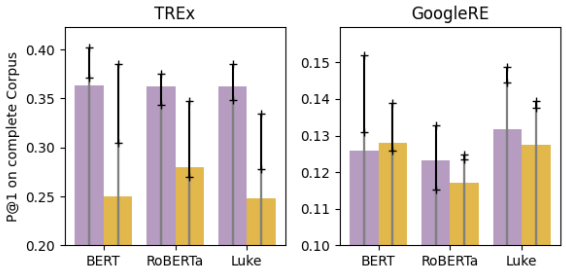


Figure 4: Effect of combined domain+range information using either clausal (purple) or appositive (orange) syntax, compared to expected interval (black). Upper bound is choosing the better answer with either domain or range information, lower bound is choosing the one with the higher confidence.

average binary entropy of the answer distributions with respect to the added information in Figure 5. The binary entropy indicates the average bit-length needed to describe the answer set (i.e., entropy of 3 tells us that we effectively narrowed down the decision to 8 words). We follow Gonen et al. (2022) who showed that information theoretic measures are a descriptor of prompt quality. We conduct our experiment on a subset containing the triples where the respective model retrieves the correct answer in the top 10 predictions for the simple prompt. Given this set is *known* to the model we expect a decrease in uncertainty when sInf is added. The diagrams in the same column show the results for the prompts written in a respective syntax (clausal, appositive), whereas the diagrams in the same row show the results for a respective completion strategy.

We observe that for clausal syntax the *Quality Completion* offers uncertainty decrease with the addition of information, while for the appositive syntax, uncertainty increases as we add more information and this remains true for both completion strategies.

5 Discussion

Jiang et al. (2019) showed that PLMs ability to

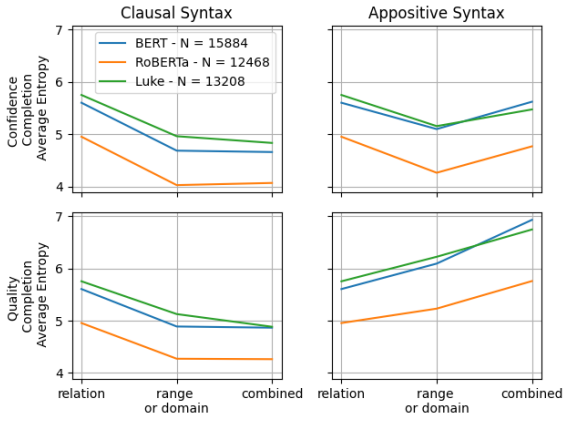


Figure 5: Average binary entropy of response distribution of known subset (correct prediction in top 10) for the TREx corpora with differently completed prompts. Clausal syntax leaves less uncertainty. PLMs even generalize this loss in uncertainty to the combined setting given clausal syntax.

retrieve information depends on the phrasing of the prompt. However, they make no qualitative statements about these phrases and even include non-natural language. In contrast, we classify different paraphrases based on their semantic and syntactic conditioning. We observe, in further detail, that adding supplementary domain and range information to simple prompts offers much less performance gain when realised via appositives than via clauses (RQ1). Adding the type-information in the sentence leaves the *rKR* task mostly unchanged. Thus, in distinction to Cao et al. (2021) and Petroni et al. (2020), we can evaluate the in-sentence information processing. This reveals that appositive syntax adds noise in many cases, thereby lowering retrieval consistency (RQ3). Furthermore, a comparison of results on the combined prompts (RQ2) and the independent prompt (RQ1) shows that information that is helpful in isolation can be distracting when added in conjunction. Thus, we extend the findings from Pandia and Ettinger (2021) showing that it is not only misleading information that obfuscates the usability of a prompt. Lastly, we also show that the consistency of PLMs heavily depends on the syntactic relation between the prompts, which was only broached in Elazar et al. (2021). The lower prevalence of appositives in the PLM training data as well as their semantic function of encoding conventional implications rather than assertions (Potts, 2012), might cause the lower performance for this syntax. Additionally, we observe that the worse-performing appositive prompts (those containing range information) tend to increase the dependency distance between the relation text and the masked object token, suggesting they perturb the models' ability to connect both. This further highlights the fragility of information flow in language representations achieved by PLMs (Ravichander et al., 2020). To counteract this fragility, a specialized training paradigm might be helpful, e.g. as proposed by Elazar et al. (2021).

We argue that a knowledge-enhanced pretraining dataset as introduced by Agarwal et al. (2020) would benefit from the integration of our findings. The introduction of multi-hop triples written in prompts that carefully follow a fitting syntax seems promising. We found that BERT is the model that performs on average the best, what is probably caused by the reliable data BERT is pre-trained on. However, peak performance for the TREx corpus of all models is at a similar level. This strengthens the already known fact that RoBERTa and its derivatives (i.e., Luke) learn more general representations of language in comparison to the older BERT model, as their performance gain comes with an increase in information. The findings presented in this paper might be used to add knowledge more reliably in approaches like K-BERT (Liu et al., 2020), KnowPrompt (Chen et al., 2022), or to aid knowledge graph construction as done in KG-BERT (Yao et al., 2019).

6 Conclusion & Future Work

In this paper, we introduced a controlled paraphrasing method and contributed the ConPare-LAMA to advance the investigation of knowledge retrieval from PLMs. Using ConPare-LAMA, we examined the impact of paraphrasing on the knowledge retrieval performance of PLMs, studying both clausal and appositive forms in conjunction with relation-specific sInf from Wikidata.

Our experimental findings reveal substantial variations in how PLMs process information based on prompt syntax. Particularly, we demonstrated that knowledge consistency is enhanced when prompts utilize clausal syntax. At the same time, we observe the vulnerability of language representations in PLMs, especially for appositive phrases. This susceptibility may be attributed to factors such as the prevalence of clausal syntax in the training data, the semantic function, and the syntactic interaction of words with the textual encoding of relations. This interpretation aligns with the conclusion by Jiang et al. (2020) that PLMs carry more knowl-

edge than previously assumed and are highly sensitive to prompt paraphrasing. Although earlier research suggested limited benefits from adding shallow syntactical features, our study found that intentionally applied (clausal) syntax provides increased regularity that can be exploited by contextualized word representations. Therefore, we assume that harnessing these regularities through dedicated syntax-aware pre-training potentially facilitates a more robust knowledge representation.

To gain more conclusive insights into the knowledge retrieval capacities of PLMs, experiments on models trained on non-fictional and factually correct pre-training corpora are crucial. This approach can help distinguish between false pre-training knowledge and wrong retrieval. Exclusively pre-trained PLMs, on non-fictional, peer-reviewed corpora like Wikipedia or scholarly publications, or synthetic corpora (Agarwal et al., 2020), can provide promising insights.

For future work, we plan to expand the template to include more diversity in syntax and knowledge. This involves applying our meta-template to additional relations derived from knowledge bases such as Wikidata and incorporating more complex syntactical structures.

7 Limitations

All our investigations are done on English text and on one token objects. However, the LAMA-probe has already inspired multilingual knowledge retrieval (Kassner et al., 2021), as well as multi-token knowledge retrieval (Kalo and Fichtel, 2022), which are out of the scope of this work. Although our experimental set-up works on a variety of relations, models, and prompt typologies, we have only considered base models. An additional testing of larger models like the 'large' alternatives of BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019b) would provide further insights. Moreover, we only investigated the addition of type information per entity (subject, object) of a knowledge triple. Another variation to our used information could be the use of different types of sInf (e.g., 'Obama is born in 1961 and was born in Hawaii.'). Additionally, one could extend our research to test the addition of more entity-specific information (i.e., 'Obama is a president and was born in 1961 and was born in Hawaii.').

8 Ethical Considerations

Our research is not using any personal data and has no direct ethical implications. However, applying the proposed approach to retrieve knowledge from PLMs might reproduce societal biases encoded in the models (e.g., retrieval performance for male scientists might be higher than for female scientists). Additionally, we strive for the lowest energy footprint by working with the base-types configurations of already pre-trained PLMs.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2020. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. *arXiv preprint arXiv:2010.12688*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. 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. *arXiv preprint arXiv:2106.09231*.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. In *Proceedings of the ACM Web Conference 2022*, pages 2778–2788.

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*.

Jeff Da and Jungo Kasai. 2019. Cracking the contextual commonsense code: Understanding commonsense reasoning aptitude of deep contextual representations. *arXiv preprint arXiv:1910.01157*.

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*.

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.

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.

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*.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. *arXiv preprint arXiv:2212.04037*.

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

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.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2021. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. *arXiv preprint arXiv:2108.02035*.

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

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2019. How can we know what language models know?

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

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*.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. *arXiv preprint arXiv:2102.00894*.

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

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*, pages 1145–1149.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019a. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

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 preprint arXiv:1907.11692*.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2020. Exploring bert's sensitivity to lexical cues using tests from semantic priming. *arXiv preprint arXiv:2010.03010*.

Lalchand Pandia and Allyson Ettinger. 2021. Sorting through the noise: Testing robustness of information processing in pre-trained language models. *arXiv preprint arXiv:2109.12393*.

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

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. ACL.

Christopher Potts. 2012. Conventional implicature and expressive content. In Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors, *Semantics: An International Handbook of Natural Language Meaning*, volume 3, pages 2516–2536. Mouton de Gruyter, Berlin.

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

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

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. *Advances in Neural Information Processing Systems*, 32.

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*, 8:842–866.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. *arXiv preprint arXiv:2302.00093*.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? *arXiv preprint arXiv:2109.07154*.

Swabha Swayamdipta, Matthew Peters, Brendan Roof, Chris Dyer, and Noah A Smith. 2019. Shallow syntax in deep water. *arXiv preprint arXiv:1908.11047*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

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*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, et al. 2019. Investigating bert's knowledge of language: five analysis methods with npis. *arXiv preprint arXiv:1909.02597*.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: Deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Kg-bert: Bert for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.

Hongbin Ye, Ningyu Zhang, Shumin Deng, Xiang Chen, Hui Chen, Feiyu Xiong, Xi Chen, and Huajun Chen. 2022. Ontology-enhanced prompt-tuning for few-shot learning. In *Proceedings of the ACM Web Conference 2022*, pages 778–787.

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)*.