Can Pretrained Language Models (Yet) Reason Deductively?

Zhangdie Yuan◇*, **Songbo Hu**♠*, **Ivan Vulić**♠, **Anna Korhonen**♠, **Zaiqiao Meng**♣♠†

◇Department of Computer Science and Technology, University of Cambridge
♠Language Technology Lab, University of Cambridge
♣School of Computing Science, University of Glasgow
◇♠{zy317,sh2091,iv250,alk23}@cam.ac.uk
♣zaiqiao.meng@glasgow.ac.uk

Abstract

Acquiring factual knowledge with Pretrained Language Models (PLMs) has attracted increasing attention, showing promising performance in many knowledge-intensive tasks. Their good performance has led the community to believe that the models do possess a modicum of reasoning competence rather than merely memorising the knowledge. In this paper, we conduct a comprehensive evaluation of the learnable *deductive* (also known as explicit) *reasoning capability* of PLMs. Through a series of controlled experiments, we posit two main findings. (i) PLMs inadequately generalise learned logic rules and perform inconsistently against simple adversarial surface form edits. (ii) While the deductive reasoning fine-tuning of PLMs does improve their performance on reasoning over unseen knowledge facts, it results in catastrophically forgetting the previously learnt knowledge. Our main results suggest that PLMs cannot yet perform reliable deductive reasoning, demonstrating the importance of controlled examinations and probing of PLMs' deductive reasoning abilities; we reach beyond (misleading) task performance, revealing that PLMs are still far from robust reasoning capabilities, even for simple deductive tasks.

1 Introduction

Pretrained Language Models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have orchestrated tremendous progress in NLP across a large variety of downstream applications. For knowledge-intensive tasks in particular, these large-scale PLMs are surprisingly good at memorising factual knowledge presented in pretraining corpora (Petroni et al., 2019; Jiang et al., 2020b) and infusing knowledge from external sources (Wang et al., 2021a; Zhou et al., 2022, *among others*), demonstrating their effectiveness in learning and capturing knowledge.
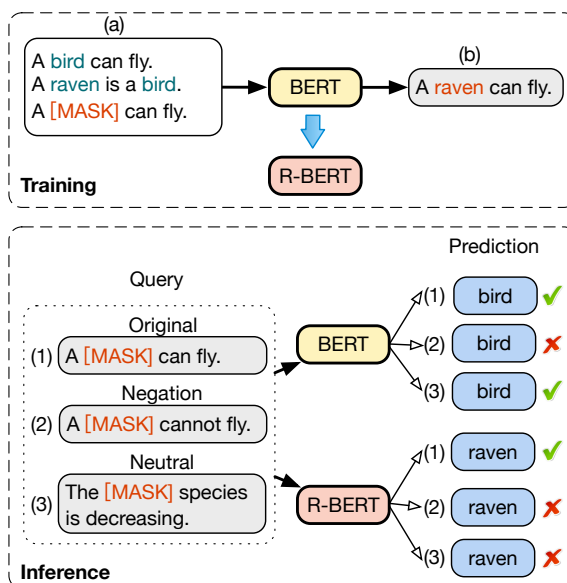

Figure 1: Training and inference for deductive reasoning. Given the explicit premises (a), the input BERT model is trained to get transformed into a reasoner R-BERT model by deductively predicting a previously unseen conclusion (b). This inference process requires R-BERT to understand factual knowledge and interpret rules (e.g. taxonomic relations), intervening directly in the deduction process.

Automatic reasoning, a systematic process of deriving previously unknown conclusions from given formal representations of knowledge (Lenat et al., 1990; Newell and Simon, 1956), has been a long-standing goal of AI research. In the NLP community, a modern view of this problem (Clark et al., 2020), where the formal representations of knowledge are substituted by the natural language statements, has recently received increasing attention,[1] yielding multiple exploratory research directions: mathematical reasoning (Rabe et al., 2021), symbolic reasoning (Yang and Deng, 2021), and com-

---

*Indicates equal contribution.
†Corresponding author.

[1]Following Clark et al. (2020), we also define natural language rules as linguistic expressions of conjunctive implications, $condition[\wedge condition]^* \rightarrow conclusion$, with the semantics of logic programs with negations (Apt et al., 1988).

monsense reasoning (Li et al., 2019). Impressive signs of progress have been reported in teaching PLMs to gain reasoning ability rather than just memorising knowledge facts (Kassner et al., 2020; Talmor et al., 2020), suggesting that PLMs could serve as effective reasoners for identifying analogies and inferring facts not explicitly/directly seen in the data (Kassner et al., 2020; Ushio et al., 2021).

In particular, *deductive reasoning*[2] is one of the most promising directions (Sanyal et al., 2022; Talmor et al., 2020; Li et al., 2019). By definition, deduction yields valid conclusions, which must be true given that their premises are true (Johnson-Laird, 1999). In the NLP community, given all the premises in natural language statements, some large-scale PLMs have shown to be able to deductively draw appropriate conclusions under proper training schemes (Clark et al., 2020; Talmor et al., 2020). Figure 1 shows an example of the training and inference processes of deductive reasoning.

Despite promising applications of PLMs, some recent studies have pointed out that they could only perform a shallow level of reasoning on textual data (Helwe et al., 2021). Indeed, PLMs can be easily affected by mispriming (Misra et al., 2020) and still hardly differentiate between positive and negative statements (i.e., the so-called *negation issue*) (Ettinger, 2020). However, given that some evidence suggests that PLMs can learn factual knowledge beyond mere rote memorisation (Heinzerling and Inui, 2021) and their limitations (Helwe et al., 2021), it is natural to ask, "*Can the current PLMs potentially serve as reliable deductive reasoners over factual knowledge?*" To answer it, as the main contribution of this work, we conduct a comprehensive experimental study on testing the learnable deductive reasoning capability of the PLMs.

In particular, we test various reasoning training approaches on two knowledge reasoning datasets. Our experimental results indicate that such deductive reasoning training of the PLMs (e.g., BERT and RoBERTa) yields strong results on the standard benchmarks, but it inadequately generalises learned logic rules to unseen cases. That is, they perform inconsistently against simple surface form perturbations (e.g., simple synonym substitution, paraphrasing or negation insertion), advocating a careful rethinking of the details behind the seemingly flawless empirical performance of deduc-

tive reasoning using the PLMs. We hope our work will inspire further research on probing and improving the deductive reasoning capabilities of the PLMs. Our code and data are available online at `https://github.com/cambridgeltl/deductive_reasoning_probing`.

2 Related Work

Knowledge Probing, Infusing, and Editing with PLMs. PLMs appear to memorise (world) knowledge facts during pretraining, and such captured knowledge is useful for knowledge-intensive tasks (Petroni et al., 2019, 2021). A body of recent research has aimed to (i) understand how much knowledge PLMs store, i.e., *knowledge probing* (Petroni et al., 2019; Meng et al., 2022); (ii) how to inject external knowledge into them, i.e., *knowledge infusing* (Wang et al., 2021b; Meng et al., 2021); and (iii) how to edit the stored knowledge, i.e. *knowledge editing* (De Cao et al., 2021). In particular, De Cao et al. (2021) have shown that it is possible to modify a single knowledge fact without affecting all the other stored knowledge. However, some empirical evidence suggests that existing PLMs generalise poorly to unseen sentences and are easily misled (Kassner and Schütze, 2020).[3] Moreover, this body of research focuses only on investigating how to *recall or expose* the factual and commonsense knowledge that has been encoded in the PLMs, rather than exploring their capabilities of deriving previously unknown knowledge via *deductive reasoning*, as done in this work.

Knowledge Reasoning with PLMs. In recent years, PLMs have also achieved impressive progress in *knowledge reasoning* (Helwe et al., 2021). For example, PLMs can infer a conclusion from a set of knowledge statements and rules (Talmor et al., 2020; Clark et al., 2020), with both the knowledge and the rules being mentioned explicitly and linguistically in the model input. Some generative PLMs, such as T5 (Raffel et al., 2020), are even able to generate natural language proofs that support implications over logical rules expressed in natural language (Tafjord et al., 2021). In particular, some large PLMs, such as LaMDA (Thoppilan et al., 2022), have been shown to be able to conduct multi-step reasoning under the chain

---

of thought prompting (Wei et al., 2022) or proper simple prompting template (Kojima et al., 2022). Although the generated 'reasoning' statements potentially benefit some downstream tasks, there is currently no evidence that the statements are generated via deductive reasoning, rather than obtained via pure memorisation. Generative reasoning models are difficult to evaluate since this requires huge effort of manual assessment (Bostrom et al., 2021).

Although some research has demonstrated that PLMs can learn to effectively perform inference which involves taxonomic and world knowledge, chaining, and counting (Talmor et al., 2020), preliminary experiments on a single test set in more recent research have revealed that fine-tuning PLMs for editing knowledge might negatively affect the previously acquired knowledge (De Cao et al., 2021). Our work performs systematic and controlled examinations of the deductive reasoning capabilities of PLMs and reaches beyond (sometimes misleading) task performance.

3 Deductive Reasoning

What is Deductive Reasoning? Psychologists define reasoning as a process of thought that yields a conclusion from precepts, thoughts, or assertions (Johnson-Laird, 1999). Three main schools describe what people may compute to derive this conclusion: relying on factual knowledge (Anderson, 2014; Newell, 1990), formal rules (Braine, 1998; Braine and O'Brien, 1991), mental models (Johnson-Laird, 1983), or some mixture of them (Falmagne and Gonsalves, 1995). Our experimental study focuses on a 'computational' aspect of reasoning — namely, whether computational PLMs for reasoning inadequately generalise learned logic rules and perform inconsistently against simple adversarial reasoning examples.

We investigate deductive reasoning in the context of NLP and neural PLMs. In particular, the goal of this deductive reasoning task is to train a PLM (e.g. BERT) over some reasoning examples (each with a set of premises and a conclusion) to become a potential reasoner (e.g. R-BERT as illustrated in Figure 1). Then, the trained reasoner can be used to infer deductive conclusions consistently over explicit premises, where the derived conclusions are usually unseen during the PLM pretraining/training. This inference process requires the underlying PLMs to understand factual knowledge and interpret rules intervening in the deduction pro-
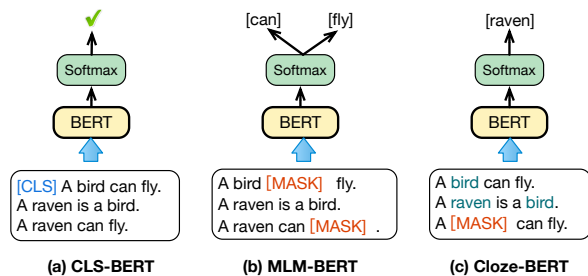

Figure 2: Different reasoning training approaches.

cess. In this paper, we only focus on the encoder-based PLMs (e.g. BERT and RoBERTa) as they can be evaluated under more controllable conditions and scrutinised via automatic evaluation. In particular, we investigate two task formulations of the deductive reasoning training: **1)** *classification-based* and **2)** *prompt-based reasoning*, as follows.

3.1 Classification-based Reasoning

The classification-based approach formulates the deductive reasoning task as a sequence classification task. Let $\mathcal{D} = \{\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \cdots, \mathcal{D}^{(n)}\}$ be a reasoning dataset, where n is the number of examples. Each example $\mathcal{D}^{(i)} \in \mathcal{D}$ contains a set of premises $\mathcal{P}^{(i)} = \{\mathbf{p}_1^{(i)}, \mathbf{p}_2^{(i)} \dots \mathbf{p}_j^{(i)}\}$, a hypothesis $\mathbf{h}^{(i)}$, and a binary label $\mathbf{l}^{(i)} \in \{0, 1\}$. A classification-based reasoner takes the input of $\mathcal{P}^{(i)}$ and $\mathbf{h}^{(i)}$, then outputs a binary label $\mathbf{l}^{(i)}$ indicating the faithfulness of $\mathbf{h}^{(i)}$, given that $\mathcal{P}^{(i)}$ is hypothetically factual.

The goal of the classification-based reasoning training is to build a statistical model parameterised by θ to characterise $P_\theta(\mathbf{l}^{(i)}|\mathbf{h}^{(i)}, \mathcal{P}^{(i)})$. Those PLMs built on the transformer encoder architecture, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), can be used as the backbone of such a classification-based reasoner. Figure 2(a) shows an example of using the BERT model to train a classification-based reasoner (CLS-BERT). In particular, given a training example $\mathcal{D}^{(i)} = \{\mathbf{l}^{(i)}, \mathbf{h}^{(i)}, \mathcal{P}^{(i)}\}$, the BERT model is trained to predict the hypothesis label by encoding $[\mathbf{h}^{(i)}; \mathcal{P}^{(i)}]$ and computing $P_\theta(\mathbf{l}^{(i)}|\mathbf{h}^{(i)}, \mathcal{P}^{(i)})$. To do so, the contextualised representation of the '[CLS]' token is subsequently projected down to two logits and passed through a softmax layer to form a Bernoulli distribution indicating that a hypothesis is *true* or *false*.

3.2 Prompt-based Reasoning

Deductive reasoning can also be approached as a cloze-completion task by formulating a valid con-

clusion as a cloze test. Specifically, given a reasoning example, i.e., $\mathcal{D}^{(i)}$ with its premises $\mathcal{P}^{(i)}$, and a cloze prompt $\mathbf{c}^{(i)}$ (e.g. *"A [MASK] can fly"*), instead of predicting a binary label, this cloze-completion task is to predict the masked token $\mathbf{a}^{(i)}$ (e.g. *raven*) to the cloze question $\mathbf{c}^{(i)}$.

The BERT-based models have been widely used in the prompt-based reasoning tasks (Helwe et al., 2021; Liu et al., 2022), by concatenating the premises and the prompt as input and predicting the masked token based on the bidirectional context. In general, there are two training objectives for the prompt-based reasoning task, i.e., the mask language modelling (MLM) and task-specific (cloze-filling) objectives. For MLM, the given PLMs are trained over the reasoning examples using their original pretraining MLM objective to impose deductive reasoning ability; see Figure 2(b) for an example of the BERT reasoner MLM-BERT. For the cloze-filling objective, the PLMs are trained with a task-specific cloze filling objective. As illustrated in Figure 2(c), Cloze-BERT is trained to predict the masked token in the cloze prompt, by computing the probability $P_\theta(\mathbf{a}^{(i)}|\mathbf{c}^{(i)}, \mathcal{P}^{(i)})$. We note that, unlike the original pretraining MLM objective where 15% tokens of the input are masked randomly, the cloze-filling objective only masks the answer token $\mathbf{a}^{(i)}$ in the cloze prompt $\mathbf{c}^{(i)}$.

This prompt-based reasoning task matches the mask-filling nature of BERT. In this way, we can probe the native reasoning ability of BERT without any further fine-tuning and evaluate the contribution of reasoning training to the PLMs' reasoning ability. Foreshadowing, our experimental results in Section 5 indicate that reasoning training impacts the model both positively and negatively.

4 Experiments and Results

Recent PLMs have shown surprisingly near-perfect performance in deductive reasoning (Zhou et al., 2020). However, we argue that high performance does not mean PLMs have mastered reasoning skills. To validate this, we run controlled experiments to examine whether PLM-based reasoners genuinely understand the natural language context, produce conclusions robustly against lexical and syntactic variance in surface forms, and apply learned rules to unseen cases.

4.1 Datasets

Two datasets are used to examine the PLM-based reasoners, namely, the Leap of Thought (**LoT**) dataset (Talmor et al., 2020) and the WikiData (**WD**) dataset (Vrandecic and Krötzsch, 2014).

LoT was originally proposed for conducting the classification-based reasoning experiments for deductive reasoning (Talmor et al., 2020) and has been used as a standard (and sole) benchmark to probe the deductive reasoning capabilities of PLMs (Tafjord et al., 2021; Helwe et al., 2021). This dataset is automatically generated by prompting knowledge graphs, including ConceptNet (Speer et al., 2017), WordNet (Fellbaum, 1998) and WikiData (Vrandecic and Krötzsch, 2014). LoT contains 30,906 training instances and 1,289 instances for each validation and testing set. Each data point in LoT also contains a set of distractors that are similar but irrelevant to deriving the conclusion.

For the prompt-based reasoning task, we can reformulate the LoT dataset to fit our cloze-completion task. Instead of having a set of premises $\mathcal{P}$, a hypothesis $\mathbf{h}$, and a binary label $\mathbf{l}$, we rewrite the hypothesis in LoT into a cloze $\mathbf{c}$ and the answer $\mathbf{a}$ (e.g. *A raven can fly.* $\rightarrow$ *A [MASK] can fly.*). Note that we only generate those cloze questions on the positive examples. Consequently, the results across these two tasks are not directly comparable.

The **WD** dataset is an *auxiliary* reasoning dataset which we generated and extracted from Wikidata5m (Wang et al., 2021b). Similar to previous work (Petroni et al., 2019; Talmor et al., 2020), we converted a set of knowledge graph triples into linguistic statements using manually designed prompts. The full description of the dataset construction is provided in Appendix C. The final WD dataset contains 4,124 training instances, 413 validation instances, and 314 test instances. WD only contains positive examples: therefore, we only use this dataset for the cloze-completion task.

4.2 Adversarial Probing

Previous work demonstrates that PLMs can achieve near-perfect empirically results in reasoning tasks. For example, RoBERTa-based models record a near-perfect accuracy of 99.7% in the deductive reasoning task on LoT (Talmor et al., 2020). However, another recent study shows that in some natural language inference benchmarks, PLMs are still not robust to the negation examples (Hossain et al., 2020), while humans can handle negations with

Perturbation	Premises	Conclusion	Valid
Original	A bird can fly. A raven is a bird.	A raven can fly.	✓
Paraphrasing	A bird is able to fly. A raven is a bird species.	A raven can fly.	✓
Synonym Substitution	A fowl can fly. A raven is a fowl.	A raven can fly.	✓
Negation	A bird can fly. A raven is a bird.	A raven cannot fly.	✗
Retained knowledge	A bird can live up to 100 years.	–	✓

Table 1: Examples of different perturbations strategies that were used to create the adversarial dataset (see §4.2).

ease. In order to systematically probe PLMs' deductive reasoning capabilities, we design controlled experiments over three different adversarial test settings by generating different *surface form perturbations*, *negation of the original examples*, and creating additional *retained knowledge anchors*. Table 1 shows some different adversarial examples.

Surface Form Perturbations. The theory of mental models postulates that deductive reasoning is based on manipulations of mental models representing situations. In other words, envisaging the situations and making a deduction can be viewed as a semantic process (Johnson-Laird, 1999; Polk and Newell, 1995). On the other hand, previous works demonstrate that, instead of learning interpretable meaning representations and generalising across different surface forms, PLMs tend to learn artefacts in the training data, e.g., higher-order word co-occurrence statistics (Sinha et al., 2021).

As both LoT and WD datasets are prompted from knowledge graphs, the lexical and syntactical variance of the dataset is minimal, with imaginable artefacts. To examine if the PLM-based reasoner could consistently perform reasoning against linguistic diversity and variability (in terms of both the token-level and the syntactic-level diversity), we employ two types of surface form perturbations to the data items from the original datasets:

- **Synonym Substitution:** In order to investigate to what extent the PLM-based reasoners would be sensitive to the token-level semantic diversity in terms of deriving their conclusions, we employ synonym substitution (Dhole et al., 2021) to the premises $\mathcal{P}$. Synonym substitution does not modify the syntactic structures and the premises' semantics, preserving all the original input's structural information. In our setting, a word is replaced by a uniform-randomly selected synonym based on WordNet (Fellbaum, 1998) with a probability of 50%.

- **Paraphrasing:** To further investigate the PLM-based reasoners' robustness on sentence-level semantic variability, we paraphrase the premises $\mathcal{P}$ with two paraphrasing systems: (i) PEGASUS, an end-to-end model fine-tuned for paraphrasing (Zhang et al., 2020) (ii) Syntactically Diverse Paraphrasing (SD-Paraphrasing), a two-step framework that incorporates neural syntactic preordering for better diversity (Goyal and Durrett, 2020).

Negated Examples. Understanding negation is often considered as the first test case in natural language understanding tasks (Ettinger, 2020; Khemlani et al., 2012; Schon et al., 2021). To examine whether PLMs can handle negation in the deductive reasoning task, we construct a set of negated samples by negating the hypothesis **h** or the cloze prompt **c** while keeping the premises $\mathcal{P}$ unchanged (Hosseini et al., 2021). For the classification-based reasoning task, the label of the negated hypothesis is then also flipped. For the cloze-completion task, since the answer for the original query will unlikely be the same answer for the negated queries, predicting the original answer would be regarded as a wrong prediction.

Anchors of Retained Knowledge. Prior work has shown that PLMs are prone to forgetting previously learnt knowledge when fine-tuning with new knowledge data (De Cao et al., 2021), the so-called catastrophic forgetting issue (Kirkpatrick et al., 2017; de Masson d'Autume et al., 2019). It is thus natural to also probe whether deductive reasoning training still retains the knowledge already stored in the original PLM. In this work, in order to measure to which extent PLMs retain the knowledge acquired during pretraining, we introduce a set of 'retained knowledge statements' or *anchors* for each dataset. There are two criteria for such *anchors*: (i) they are semantically close to the conclusions in our test data, (ii) they do not meet the conditions of the

premises for a reasoning replacement.[4] Ideally, the reasoning training should not affect the prediction of PLMs when reasoning over these anchors.

We create such a set of anchors for both LoT and WD, and investigate the behaviour of the reasoning models over these anchors based on the prompt-based reasoning task. In particular, these anchors should be real-world textual statements that contain the target word (to meet criterion (i) above), but their newly composed sentences (by the reasoning replacement) are unlikely true statements (to meet criterion (ii) above). To this end, we use the BM25 algorithm (Sparck Jones et al., 2000) to retrieve the top 10 sentences for each 'reasoning' target word from the Wikipedia corpus.[5] Then, we construct the anchor sentences based on these top 10 retrieved sentences by replacing their target words with their hyponym/hypernym words. The final anchors are selected from these top 10 sentences only when the newly created sentences do not likely exist in the entire Wikipedia corpus (i.e. their top-1 similar sentence should have less than a BM25 score of 50). Ideally, this set of retrained knowledge statements is relevant but should not be affected by the reasoning training.

4.3 Evaluation

Following previous work (Talmor et al., 2020), the evaluation metric for classification-based reasoning is accuracy. For prompt-based reasoning, we calculate top K recall (R@K) by measuring what fraction of the correct answers are retrieved in the top K predictions. For the negation examples, we report the top K error rate (E@K) because a retrieved answer **a** and the negated cloze question ¬**c** would compose a fallacy.

In the following, we report our findings and numerical results based on the BERT-based reasoners (in particular `bert-base-uncased`), but we note that other PLMs (such as RoBERTa) of various sizes observe the same performance trends and result in the same findings and conclusions. Appendix B provides results for other PLMs.

---

[4]Taking the premise (a) of Figure 1 as an example, the 'bird' token in the statement 'A bird can fly.' was replaced by its hyponym 'raven' after reasoning training. However, in the anchor statement 'The bird species is decreasing.', the 'bird' token should not be replaced by 'raven'.

[5]ElasticSearch: https://github.com/elastic/elasticsearch.

Model	R@1	R@5	R@10
Dataset: **LoT**			
Pretrained	13.15	59.18	70.96
MLM-BERT	98.36	98.36	98.36
Cloze-BERT	99.73	100	100
Dataset: **WD**			
Pretrained	27.07	64.97	72.93
MLM-BERT	99.04	99.36	99.36
Cloze-BERT	100	100	100

Table 2: Recall of the correct answer in the top K predictions (R@K) from the BERT model before and after deductive prompt-based reasoning fine-tuning. Both MLM-BERT and Cloze-BERT achieve (near-)perfect R@1 scores after fine-tuning.

5 Results and Discussion

We evaluate the impact of reasoning training on the PLMs and investigate their robustness against three well-known issues of PLMs: utilising artefacts from data, incapability of modelling negation, and catastrophic forgetting. We further conduct qualitative analysis to understand the inference errors introduced by deductive reasoning training.

5.1 Deductive Reasoning Training

Finding 1 *All the deductive reasoning training approaches significantly improve PLMs' reasoning capabilities, achieving near-perfect deductive reasoning performance on both the reasoning test sets.*

Table 2 reveals that the prompt-based reasoners achieve near-perfect performance on both datasets, regardless of the reasoning training method. In particular, on LoT dataset the R@1 score of BERT has increased from 13.15% to 98.36% and 99.73% after the MLM reasoning training (i.e. MLM-BERT) and the cloze reasoning training (i.e. Cloze-BERT) respectively, which are in line with previously reported result (Talmor et al., 2020). The near-perfect trends are observed in the classification-based reasoning models, where CLS-BERT also achieves a high accuracy score of 94.72% after reasoning training (Table 5).[6] In sum, while the off-the-shelf BERT model already demonstrates a decent level of empirical performance, conducting reasoning training on the pretrained BERT achieves strong or even near-perfect performance on both datasets.

1452

Adversarial Probing	LoT			WD		
	Pretrained	MLM-BERT	Cloze-BERT	Pretrained	MLM-BERT	Cloze-BERT
Original	13.15	98.36	99.73	27.07	99.04	100.00
+ Pegasus-Paraphrasing	11.79 (↓1.36)	64.66 (↓33.70)	50.96 (↓48.77)	16.88 (↓10.19)	49.36 (↓49.68)	51.27 (↓48.73)
+ SD-Paraphrasing	5.75 (↓7.40)	9.32 (↓89.04)	0.55 (↓99.18)	16.56 (↓10.51)	27.39 (↓71.65)	31.21 (↓68.79)
+ Syn. Substitution	20.00 (↑6.85)	64.38 (↓33.98)	64.66 (↓35.07)	15.29 (↓11.78)	61.78 (↓37.26)	62.42 (↓37.58)

Table 3: R@1 scores on test sets obtained via applying various surface form perturbations from Section 4.2.

Examples	BERT	MLM-BERT	Cloze-BERT
A holly is not music. A holly is part of a forest. A plant is an actor. A music is not an actor. A holly is a plant. A bluebottle is an organism. A marigold is not an angiosperm. A holly is not an important food source.	musician (0.1213) ✓	holly (0.7168) ✗	holly (0.9999) ✗
A \<MASK\> is not an actor.			
A perry is not a tree. A tree is not capable of burn. Alcohol is capable of burn. A perry is an alcohol. A \<MASK\> is not capable of burn.	tree (0.5244) ✓	person (0.0300) ✗	perry (0.9999) ✗

Table 4: Examples of top 1 predictions from the set of negated examples based on LoT test set. The wrong predictions are in red, and the reasonable predictions are in green. Some other key-related entities are in the same colour. The numbers in the parentheses are the respective probabilities of each prediction after the softmax layer.

Adversarial Probing	LoT
	CLS-BERT
Original	94.72
+ Pegasus-Paraphrasing	83.86 (↓10.86)
+ SD-Paraphrasing	71.14 (↓23.58)
+ Syn. Substitution	84.48 (↓10.24)
+ Negation	9.34 (↓85.49)

Table 5: Accuracy scores on test sets obtained via applying various surface form perturbations from Section 4.2.

5.2 Surface Form Perturbations

Finding 2 *Surface form perturbations drastically decrease PLMs' reasoning performance.*

A natural follow-up question to ask is to what extent the aforementioned near-perfect numbers really reflect the model's reasoning abilities. We thus perform surface form perturbations to add lexical and syntactic variance to the test datasets and probe the model against such variations.

Table 3 demonstrates the performances of the MLM-BERT and Cloze-BERT reasoners, as well as the vanilla pretrained BERT language model, on our controlled test sets generated by different perturbation approaches. We can observe that the scores for both reasoners decrease substantially (>30%) even when applying simple synonym substitution, which only adds lexical variance to the prompt-generated query. In contrast, the pretrained BERT language model is less vulnerable to such an issue. This finding aligns with the hypothesis that PLMs tend to memorise word co-occurrence statistics (Sinha et al., 2021).[7]

We also observe from Table 3 that the R@1 performances of all reasoners decrease strongly when applying either the Pegasus-Paraphrasing or SD-Paraphrasing methods to the premises. In particular, the drops are especially pronounced with the SD-Paraphrasing method, which is designed exactly to generate syntactically very diverse paraphrases. On the other hand, we see from Table 5 that similar performance degradation trends can be observed in the classification-based reasoning task. These results illustrate that current PLMs perform inconsistently against various surface form perturbations, suggesting that future work should look into the creation of more robust reasoners that should be resilient to lexical, syntactic, and semantic variability.

---

[6]Note that we cannot obtain the performance of original BERT since the classification head has not been trained yet.

[7]The comparison of large models (e.g. bert-large) can be found in Appendix B, which indicates the similar trends.
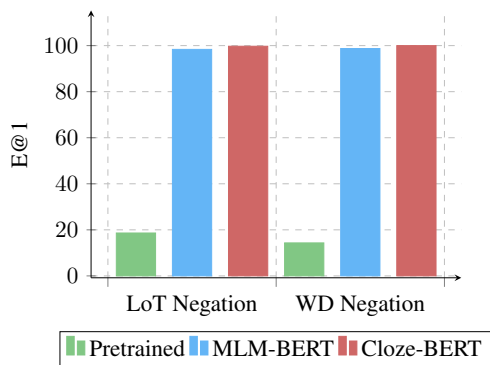
Figure 3: Results (E@1 scores, lower is better) on the test set comprising negated examples; before versus after deductive-reasoning training.
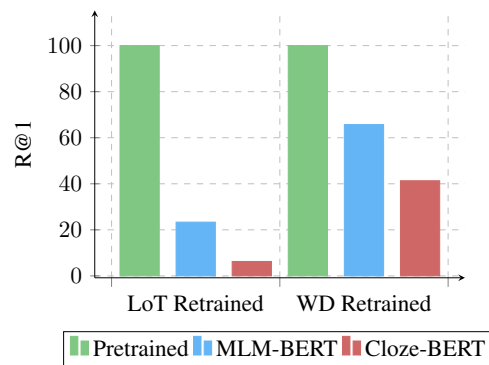

Figure 4: Results (R@1 scores, higher is better) on the test set comprising anchors of retained knowledge; before versus after deductive-reasoning training.

5.3 Negated Examples

Finding 3 *All reasoners cannot distinguish between negated and non-negated examples.*

Figure 3 reveals that the error rate E@1 scores (the lower, the better)[8] for all reasoners on the set of negated examples largely increase after reasoning training. The original off-the-shelf BERT model achieves E@1 of 18.63% on LoT and 14.33% on the WD dataset. However, after reasoning training, the error rate of MLM-BERT significantly increases to 98.36% (LoT) and 98.73% (WD). Cloze-BERT's performance is even worse, with E@1 scores of 99.73% (LoT) and 100% (WD), suggesting a clear case of overfitting to word co-occurrence and other artefacts in the training sets. Moreover, the accuracy score on negated LoT examples is only 9.34%, while a random baseline would score 50%. In sum, these scores indicate that the PLM-based reasoners cannot distinguish between negated and non-negated examples, and their performance on negated examples substantially worsens after reasoning training due to task-specific overfitting.

A quick error analysis, provided in Table 4, further points to the issues with negated examples. The first example shows that the off-the-shelf BERT model makes a reasonable guess, semantically related to an entity mentioned in the premises. This guess is similar to a human guess from the same premises. However, after reasoning training, both MLM-BERT and Cloze-BERT make wrong predictions, and Cloze-BERT is extremely confi-

---

[8]E@1 only denotes a specific type of error: the model makes a wrong prediction due to negation being added, i.e. the negation error. However, many other types of errors that a model can make a wrong prediction, but they are not our focus and thus not measured. In other words, (1 - E@1) is not equivalent to accuracy.

dent in its wrong predictions.

5.4 Anchors of Retained Knowledge

Finding 4 *Previously learnt knowledge is not fully retained after reasoning training, and the trained reasoners (catastrophically) forget it.*

Figure 4 shows that performance on the anchors deteriorates substantially after reasoning training. On LoT, MLM-BERT 'forgets' ∼77% of the previously learnt knowledge, achieving only 23.34% on R@1. Cloze-BERT performs even worse, scoring only 6.27% R@1. The drops are slightly lower but still substantial on WD: MLM-BERT retains 65.7% and Cloze-BERT retains 41.32% of the previously stored knowledge. This result indicates that reasoning training yields the well-known phenomenon of catastrophic forgetting: this effect seems even more pronounced with Cloze-BERT, which relies on a very task-specific objective that might result in overfitting the task data.

Furthermore, Table 6 displays several qualitative examples where the predictions in green refer to the correct predictions based on human judgement. Notice that in the last example, even though the predictions from the two trained reasoners are correct, the probabilities are much lower.

Several strategies might help mitigate catastrophic forgetting. A promising direction is encapsulating lightweight adapter modules (Houlsby et al., 2019; Pfeiffer et al., 2020b; Ansell et al., 2021) within the underlying PLM, where all the 'deductive reasoning capability' will be stored solely in the adapter modules, leaving the original PLM intact (Pfeiffer et al., 2020a). Other similar parameter-efficient and modular methods include prefix tuning (Li and Liang, 2021) or sparse masks (Sung et al., 2021; Ansell et al., 2022). Their main

Examples	BERT	MLM-BERT	Cloze-BERT
a flea is a parasitic <MASK>.	insect (0.2006) ✓	substance (0.1950) ✗	drone (0.0193) ✗
vapour density is a unitless <MASK>.	quantity (0.3494) ✓	quantity (0.2197) ✓	volume (0.0412) ✗
both sexes have a throat <MASK>.	pouch (0.5169) ✓	##lid (0.3606) ✗	hollow (0.0135) ✗
firefox is a web <MASK>.	browser (0.9144) ✓	browser (0.8861) ✓	browser (0.3243) ✓

Table 6: Examples of top 1 predictions from the set of Anchors of Retained Knowledge based on LoT. A wrong prediction is in red, and a reasonable prediction (aligning with human judgement) is in green. The numbers in the parentheses are the respective probabilities of each prediction after the softmax layer.

premise is to separate knowledge extraction and composition, preserving previously learnt knowledge during reasoning training. We leave reasoning training with such methods for future research.

6 Conclusion

In this paper, we probed into the deductive reasoning capabilities of PLMs and conducted comprehensive controlled experiments to examine and compare various deductive reasoning training approaches. Our experimental results showed that current PLM-based deductive reasoners suffer from several issues: **1)** they rely on artefacts from the training data, **2)** they are incapable of modelling negation in deductive reasoning, and **3)** they forget knowledge acquired during pretraining when they get specialised into deductive reasoners. In particular, our experimental study demonstrated that models are vulnerable to multiple adversarial methods, including simple surface form perturbations such as synonym substitution or paraphrasing. While the PLMs trained for deductive reasoning achieve seemingly perfect empirical results in different reasoning datasets, they cannot yet systematically generalise to other deductive reasoning examples. Consequently, our study also calls for further, more rigorous examinations of future PLM-based models' deductive reasoning abilities.

Acknowledgements

Limitations

Despite the thorough experiments on standard and popular PLMs of various sizes, this study explores only encoder-based models. Some generation-based models under other Transformer architectures, such as encoder-decoder (T5) or decoder-only (GPT-3), were also deployed in the reasoning tasks (Bostrom et al., 2021; Wei et al., 2022). We do not probe those groups of models here due to the difficult of evaluation and we leave them for future research.

Further, for prompt-based reasoning, the current reasoners only support and have been evaluated on single-token prediction in a single language (i.e. English). Several prior works have demonstrated that conducting multi-token prediction is considerably more difficult (Jiang et al., 2020a), which would pose an additional challenge to the PLMs in deductive reasoning tasks. One avenue of future work should also extend the scope of analyses to multilingual and multi-token prediction.

In addition, we note that better evaluation resources that could address paraphrases and word senses, especially for mask-filling tasks, are still lacking. This limitation is particularly significant in our setting. For example, in addition to the single-token answer in the evaluation datasets we used, there are some other feasible answers (e.g. synonyms) for the same query, which should also be considered a correct prediction. However, such answers are ignored by the current standard evaluation protocols. As a result, there is a certain level of unavoidable noise in the evaluation process.

Finally, introducing a reasoning dataset is highly challenging and appreciated by the community. Leap-of-thought is to our knowledge the only existing dataset that is suitable for our deductive reasoning evaluation. To solidify our conclusions, we further constructed an auxiliary dataset (WD) following a similar procedure to LoT. Although our data construction method is commonly used to extract reasoning examples, such an automatic procedure, unfortunately, inevitably reflects the quality and errors (e.g. nonsensical statements) from our source (WikiData). To reduce such noisy examples, we have conducted multiple rounds of filtering (see

Appendix for the filtering process) and manually removed as many meaningless relations as we can, given that manually verifying each reasoning example is a highly labour-intensive task.

References

Somak Aditya, Yezhou Yang, and Chitta Baral. 2018. Explicit reasoning over end-to-end neural architectures for visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

John R Anderson. 2014. *Rules of the mind*. Psychology Press.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 1778–1796.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781.

Krzysztof R. Apt, Howard A. Blair, and Adrian Walker. 1988. Towards a theory of declarative knowledge. In Jack Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 89–148. Morgan Kaufmann.

Kaj Bostrom, Xinyu Zhao, Swarat Chaudhuri, and Greg Durrett. 2021. Flexible generation of natural language deductions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6266–6278.

Martin D Braine and David P O'Brien. 1991. A theory of if: A lexical entry, reasoning program, and pragmatic principles. *Psychological review*, page 182.

Martin DS Braine. 1998. Steps toward a mental-predicate logic. In *Mental logic*, pages 281–340. Psychology Press.

John Broome. 2013. *Rationality through reasoning*. John Wiley & Sons.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 3882–3890.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*.

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*, pages 4171–4186.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard H. Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, and et al. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation. *CoRR*.

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*, pages 34–48.

R J Falmagne and J Gonsalves. 1995. Deductive inference. *Annual Review of Psychology*, 46(1):525–559.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 238–252.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main xvolume*, pages 1772–1791.

Chadi Helwe, Chloé Clavel, and Fabian M. Suchanek. 2021. Reasoning with transformer-based models: Deep learning, but shallow reasoning. In *3rd Conference on Automated Knowledge Base Construction*.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 9106–9118.

Arian Hosseini, Siva Reddy, Dzmitry Bahdanau, R Devon Hjelm, Alessandro Sordoni, and Aaron Courville. 2021. Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1301–1312.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 2790–2799.

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*, pages 5943–5959.

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*, pages 423–438.

P. N. Johnson-Laird. 1999. Deductive reasoning. *Annual Review of Psychology*, 50(1):109–135. PMID: 15012459.

Philip N. Johnson-Laird. 1983. *Mental models : towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press. Excerpts available on Google Books.

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.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818.

Sangeet Khemlani, Isabel Orenes, and P. N. Johnson-Laird. 2012. Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, 24(5):541–559.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *CoRR*.

Douglas B. Lenat, Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd. 1990. CYC: toward programs with common sense. *Commun. ACM*, 33(8):30–49.

Shiyang Li, Jianshu Chen, and Dian Yu. 2019. Teaching pretrained models with commonsense reasoning: A preliminary kb-based approach. *arXiv preprint arXiv:1909.09743*.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 4582–4597.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3169.

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

Zaiqiao Meng, Fangyu Liu, Thomas Clark, Ehsan Shareghi, and Nigel Collier. 2021. Mixture-of-partitions: Infusing large biomedical knowledge graphs into BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4672–4681.

Zaiqiao Meng, Fangyu Liu, Ehsan Shareghi, Yixuan Su, Charlotte Collins, and Nigel Collier. 2022. Rewire-then-probe: A contrastive recipe for probing biomedical knowledge of pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 4798–4810.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2020. Exploring BERT's sensitivity to lexical cues using tests from semantic priming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4625–4635.

A Newell. 1990. Unified theories of cognition cambridge university press. *Cambridge, MA*.

A. Newell and H. Simon. 1956. The logic theory machine–a complex information processing system. *IRE Transactions on Information Theory*, pages 61–79.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian

Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 2523–2544.

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*, pages 2463–2473.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 46–54.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7654–7673.

Thad A Polk and Allen Newell. 1995. Deduction as verbal reasoning. *Psychological Review*, page 533.

Markus Norman Rabe, Dennis Lee, Kshitij Bansal, and Christian Szegedy. 2021. Mathematical reasoning via self-supervised skip-tree training. In *International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992.

Soumya Sanyal, Harman Singh, and Xiang Ren. 2022. FaiRR: Faithful and robust deductive reasoning over natural language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 1075–1093.

Claudia Schon, Sophie Siebert, and Frieder Stolzenburg. 2021. Negation in cognitive reasoning. In *KI 2021: Advances in Artificial Intelligence - 44th German Conference on AI*, Lecture Notes in Computer Science, pages 217–232.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913.

K. Sparck Jones, S. Walker, and S.E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 1. *Information Processing & Management*, 36(6):779–808.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Yi-Lin Sung, Varun Nair, and Colin Raffel. 2021. Training neural networks with fixed sparse masks. In *Advances in Neural Information Processing Systems*, pages 24193–24205.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. In *Advances in Neural Information Processing Systems*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*.

Asahi Ushio, Luis Espinosa Anke, Steven Schockaert, and Jose Camacho-Collados. 2021. BERT is to NLP what AlexNet is to CV: Can pre-trained language models identify analogies? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3609–3624.

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

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021a. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kaiyu Yang and Jia Deng. 2021. Learning symbolic rules for reasoning in quasi-natural language. *arXiv preprint arXiv:2111.12038*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11328–11339.

Ming Zhou, Nan Duan, Shujie Liu, and Heung-Yeung Shum. 2020. Progress in neural nlp: Modeling, learning, and reasoning. *Engineering*, 6(3):275–290.

Wenxuan Zhou, Fangyu Liu, Ivan Vulić, Nigel Collier, and Muhao Chen. 2022. Prix-LM: Pretraining for multilingual knowledge base construction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5412–5424.

A Experimental Details

Table 7 lists our model hyperparameters. Among these models, MLM-BERT and Cloze-BERT were implemented using the HuggingFace transformers package (Wolf et al., 2020). We implement CLS-BERT via the SBERT repository (Reimers and Gurevych, 2019), which is built on top of the HuggingFace repository (Wolf et al., 2020). Unless mentioned otherwise, all the hyperparameters are set to the default values provided in the Hugging-Face and SBERT repositories.

Hyper-parameter	Value
MLM-BERT and Cloze-BERT	
batch_size	2
max_sequence_length	512
training_epoch	3
CLS-BERT	
batch_size	64
max_sequence_length	128
training_epoch	20

Table 7: Model hyper-parameters.

B Evaluation Results for PLMs

Table 8 supplements the main paper by providing additional results with classification-based reasoners, where the reasoners start from different PLMs: distil-bert, bert-base, bert-large, roberta-base, and roberta-large. These results corroborate the main findings presented in the main paper; see Section 5. Larger models do perform slightly better than their smaller variants on average. However, the results also demonstrate that different models, regardless of their size, suffer from exactly the same issues, discussed in the main paper.

Table 9 and Table 10 demonstrate performance over adversarial test sets for prompt-based reasoners (1. MLM-based, 2. Cloze-based). The findings from these tables align with the findings from the main results presented in Section 5.

C WD Dataset Construction Pipeline

We construct the **WD** dataset following the pipeline shown in Figure 5, and outlined in what follows.

Source Data. We choose the Wikidata5m dataset (Wang et al., 2021b) as the knowledge source for **WD**. Wikidata5m is a million-scale knowledge graph dataset created upon Wikidata (Vrandecic and Krötzsch, 2014). This dataset comprises 20 million triples, describing relevant and important knowledge statements about real-world entities.
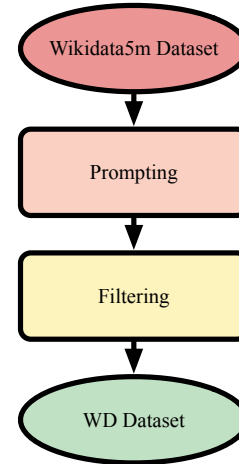

Figure 5: Data construction pipeline for the **WD** dataset. This **WD** dataset is extracted and derived from Wiki-Data5m (Wang et al., 2021b). Following the previous pipeline (Petroni et al., 2019), we convert a set of knowledge graph triples into linguistic statements using manually designed prompts.

Prompting. We manually select a set of relations based on their frequency and design their corresponding prompts shown in Table 11. Given a taxonomic relation, such as is_a with relation ID P31, we sample a relevant taxonomic-knowledge graph triple: $\langle raven, is\_a, bird \rangle$. We then retrieve other relevant triples about the subject and the object, for example, $\langle bird, is\_capable\_of, fly \rangle$ and $\langle raven, is\_capable\_of, fly \rangle$. Next, we use our predefined prompts to convert the knowledge graph triples to textual knowledge statements: *A bird can fly*, *A bird is a raven* and *A raven can fly*. Furthermore, these prompted statements assemble an inference instance in the **WD** dataset: if *A bird can fly* and *A bird is a raven*, then *A raven can fly*. In our cloze-completion task setting, we mask the object of the taxonomic triple (e.g. $raven$) in the conclusion statement to form a cloze question (*A [MASK] can fly*). Therefore, $raven$ is the correct answer to this question.

Filtering. We filter those constructed inference instances with the following properties: (i) We only choose examples with answers being a single masked token, and these answers should be included in the BERT vocabulary. (ii) For all in-

Adversarial Probing	LoT				
	distilbert-based-uncased	bert-base-uncased	bert-large-uncased	roberta-base	roberta-large
Original	90.22	94.72	98.76	92.71	99.38
+ Pegasus-Paraphrasing	77.27 (↓12.95)	83.86 (↓10.86)	84.40 (↓14.35)	80.84 (↓11.87)	87.20 (↓12.18)
+ SD-Paraphrasing	66.56 (↓23.66)	71.14 (↓23.58)	73.39 (↓25.39)	75.95 (↓16.76)	78.97 (↓20.40)
+ Syn. Substitution	81.46 (↓8.76)	84.48 (↓10.24)	93.33 (↓5.43)	80.76 (↓11.95)	94.26 (↓5.12)
+ Negation	18.86 (↓71.37)	9.34 (↓85.49)	2.58 (↓96.28)	19.40 (↓73.31)	35.85 (↓63.53)

Table 8: Accuracy on different LoT adversarial test sets for classification-based reasoners with various back-boned PLMs.

Adversarial Probing	LoT			WD		
	MLM-DISTILBERT	MLM-BERT	MLM-BERT-LARGE	MLM-DISTILBERT	MLM-BERT	MLM-BERT-LARGE
Original	99.45	98.36	99.45	99.36	99.04	98.09
+ Pegasus-Paraphrasing	76.71 (↓22.74)	64.66 (↓33.7)	60.27 (↓39.18)	56.05 (↓43.31)	49.36 (↓49.68)	52.23 (↓45.86)
+ SD-Paraphrasing	24.66 (↓74.79)	9.32 (↓89.04)	4.93 (↓94.52)	28.66 (↓70.7)	27.39 (↓71.65)	33.12 (↓64.97)
+ Syn. Substitution	64.11 (↓35.34)	64.38 (↓33.98)	64.93 (↓34.52)	62.1 (↓37.26)	61.78 (↓37.26)	60.19 (↓37.9)
+ Negation (E@1 ↓)	97.53	98.36	95.89	99.36	98.73	98.41
Retained Knowledge	25.86	23.34	26.67	58.33	65.7	56.61

Table 9: Top 1 recall (R@1) on adversarial test sets for various prompt-based reasoners with MLM training. The numbers for the negated examples indicate the top 1 error rates (the lower, the better).

Adversarial Probing	LoT			WD		
	Cloze-DISTILBERT	Cloze-BERT	Cloze-BERT-LARGE	Cloze-DISTILBERT	Cloze-BERT	Cloze-BERT-LARGE
Original	100	99.73	98.63	100	100	100
+ Pegasus-Paraphrasing	57.81 (↓42.19)	50.96 (↓48.77)	47.4 (↓51.23)	57.01 (↓42.99)	51.27 (↓48.73)	56.37 (↓43.63)
+ SD-Paraphrasing	2.74 (↓97.26)	0.55 (↓99.18)	0.82 (↓97.81)	37.26 (↓62.74)	31.21 (↓68.79)	46.82 (↓53.18)
+ Syn. Substitution	64.66 (↓35.34)	64.66 (↓35.07)	63.84 (↓34.79)	62.74 (↓37.26)	62.42 (↓37.58)	62.74 (↓37.26)
+ Negation (E@1 ↓)	100	99.73	98.63	100	100	99.68
Retained Knowledge	11.79	6.27	1.45	35.61	41.32	16.95

Table 10: Top 1 recall (R@1) on adversarial test sets for various prompt-based reasoners with Cloze-filling training. The numbers for the negated examples indicate the top 1 error rates (the lower, the better).

ference instances, the maximum number of occurrences of a single answer is 50 to balance the dataset and avoid excessive repetition.

Final WD Dataset. The final WD dataset contains 4,851 instances, which are randomly split into 4,124/413/314 instances for training/validation/testing while keeping that the answers of the testing set should not appear in the training/validation sets. This is to ensure that trained reasoners need to draw conclusions via conducting deductive reasoning rather than via memorisation.

Relation ID	Prompt
P31	[X] is a [Y] .
P136	[X] is a genre of [Y] .
P179	[X] is part of the [Y] series .
P279	[X] is a subclass of [Y] .
P527	[X] consists of [Y] .
P1269	[X] is a topic of [Y] .
P17	[X] is hosted in [Y] .
P39	[X] holds a [Y] position .
P101	[X] is a subject of [Y] .
P106	[X] is a [Y] by profession .
P140	The religion of [X] is [Y] .
P144	[X] is based on [Y] .
P180	[X] is a painting of [Y] .
P276	[X] is located at [Y] .
P306	[X] runs on [Y] operating system .
P355	[X] owns [Y] .
P360	[X] is a list of [Y] .
P400	[Y] is a platform of [X] .
P404	The game mode of [X] is [Y] .
P462	The color of [X] is [Y] .
P463	[X] is a member of [Y] .
P737	[X] is influenced by [Y] .
P749	[Y] owns [X] .
P1303	[X] plays [Y] .
P1343	[X] is written about in [Y] .

Table 11: Manually written prompts for generating the WD dataset. Given a triple $\langle [X], R\_ID, [Y] \rangle$, the textual knowledge statement (e.g. premises) is written based on the above prompts. R_ID is the unique relation ID in the Wikidata5m dataset. Gray entries (first six rows) denote taxonomic relations.