Does Pre-training Induce Systematic Inference? How Masked Language Models Acquire Commonsense Knowledge

Ian Porada[*]
Mila, McGill University
ian.porada@mail.mcgill.ca

Alessandro Sordoni
Microsoft Research Montréal
alsordon@microsoft.com

Jackie Chi Kit Cheung
Mila, McGill University
jcheung@cs.mcgill.ca

Abstract

Transformer models pre-trained with a masked-language-modeling objective (e.g., BERT) encode commonsense knowledge as evidenced by behavioral probes; however, the extent to which this knowledge is acquired by systematic inference over the semantics of the pre-training corpora is an open question. To answer this question, we selectively inject verbalized knowledge into the pre-training minibatches of BERT and evaluate how well the model generalizes to supported inferences after pre-training on the injected knowledge. We find generalization does not improve over the course of pre-training BERT from scratch, suggesting that commonsense knowledge is acquired from surface-level, co-occurrence patterns rather than induced, systematic reasoning.

1 Introduction

Pre-trained Transformers, such as BERT, encode knowledge about the world (Petroni et al., 2019; Zhou et al., 2020); e.g., BERT assigns relatively high probability to "fly" appearing in the context "robins can ___ ." In this work, we investigate whether such knowledge is acquired during pre-training through systematic inference over the semantics of the pre-training corpora; e.g., can models systematically infer "robins can fly" from the premises "birds can fly" and "robins are birds?"

Resolving *how* models acquire commonsense knowledge has important implications. If models learn to make systematic inferences through pre-training, then scaling up pre-training is a promising direction for commonsense knowledge acquisition. If, instead, models only ever generalize based on superficial, surface-level patterns, then the majority of commonsense knowledge, which is only supported implicitly, will never be acquired (Gordon and Van Durme, 2013; Forbes and Choi, 2017).

On the one hand, there is cursory evidence that pre-training might induce the ability to systematically reason about the world. When fine-tuned on supervised training sets, pre-trained models can classify valid inferences better than strong baselines (Clark et al., 2020; Talmor et al., 2020b); and, in zero-shot evaluations, pre-trained models perform relatively well on reasoning tasks that *may* require systematic reasoning, such as number comparison (Talmor et al., 2020a) and Winograd schemas (Sakaguchi et al., 2021).

On the other hand, existing works have argued that pre-training does not generalize by systematic inference over semantics on the basis of theoretical or synthetic results (Bender and Koller, 2020; Merrill et al., 2021; Traylor et al., 2021). Referring to physical commonsense knowledge acquired by BERT, Forbes et al. (2019) conclude that "neural language representations still only learn associations that are explicitly written down."

Our main contribution is a direct evaluation of the training dynamics of BERT's reasoning ability. We inject verbalized knowledge, such as "robins are birds" (where the masked token is the predicate, e.g., "birds"), into the minibatches of BERT throughout pre-training. We then consider how well BERT generalizes to supported inferences; e.g., how does the likelihood of "robins can ___" $\rightarrow$ "fly" change?

We find generalization does not improve over the majority of pre-training which supports the hypothesis that the type of commonsense knowledge studied is not acquired by systematic inference. Rather, our findings suggest this knowledge is acquired from surface-level, co-occurrence patterns.

2 Related Work

Commonsense knowledge acquisition is a long-standing challenge in natural language processing (Charniak, 1973; Hwang et al., 2021; Zhang et al., 2021), and current approaches rely on knowledge

[*] Work conducted while the author was an intern at Microsoft Research Montréal.

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4550 - 4557
July 10-15, 2022 ©2022 Association for Computational Linguistics

acquired by pre-trained Transformer language models (Bosselut et al., 2019; Zhang et al., 2020; West et al., 2021). The commonsense reasoning ability of these language models has been evaluated using behavioral probes (Ettinger, 2020; Misra et al., 2021; He et al., 2021) and downstream, fine-tuned evaluations (Banerjee et al., 2021; Zhou et al., 2021; Tafjord and Clark, 2021). Such works consider the knowledge encoded by a model after pre-training.

When fine-tuned on supervised datasets, pre-trained models can learn to make systematic inferences to some extent (Clark et al., 2020; Tafjord et al., 2021; Gontier et al., 2020; Shaw et al., 2021; Li et al., 2021). By *systematic inferences*, we refer to the ability to learn general rules and apply them in novel settings, as opposed to learning only particular instances of the rule (Fodor and Pylyshyn, 1988; Lake and Baroni, 2018; Bahdanau et al., 2019).

Similar to our experiments, recent work has considered the training dynamics of pre-trained models (Brown et al., 2020; Kaplan et al., 2020). Notably, Liu et al. (2021) evaluate the zero-shot performance of RoBERTa on the oLMpics reasoning tasks throughout pre-training, but find the knowledge studied is never learned. In contrast, we explore *how* learned knowledge is acquired.

Close in spirit to our work, Kassner et al. (2020) pre-train a masked language model on a synthetic dataset to isolate reasoning ability. Wei et al. (2021) also intervene on BERT's pre-training data in a syntactic evaluation and conclude that subject-verb agreement is sometimes inferred from systematic rules for frequent words.

Finally, De Cao et al. (2021) explore how knowledge encoded in BERT is affected by gradient updates when fine-tuning on a downstream classification task. Hase et al. (2021) build on this work and explore how gradient updates on verbalized premises affect models' performance on supported inferences. In contrast, we focus on knowledge obtained by the pre-training objective itself.

3 Method

The purpose of our evaluation is to answer the question: *does BERT systematically infer commonsense knowledge from premises present in the pre-training corpora?*

We focus on one specific type of commonsense knowledge that BERT is known to encode, namely entity properties annotated in ConceptNet (Speer et al., 2017). This knowledge can be represented abstractly as (`subject`, `relation`, `object`) triples. We verify BERT's encoding of knowledge by the ability to predict the `object` conditioned on a verbalization of the knowledge containing only the `subject` and `relation`; e.g., for (`robin`, `capable-of`, `fly`), we evaluate the ability to predict "fly" appearing in the context "robins can ___ ."

Type	Example
Super-statement	*A boat has a* ___ . $\rightarrow$ *hull*
Sub-statement	*A canoe has a* ___ . $\rightarrow$ *hull*
Class Relation	*A canoe is a* ___ . $\rightarrow$ *boat*

Table 1: An example of the three knowledge types as masked-token prediction.

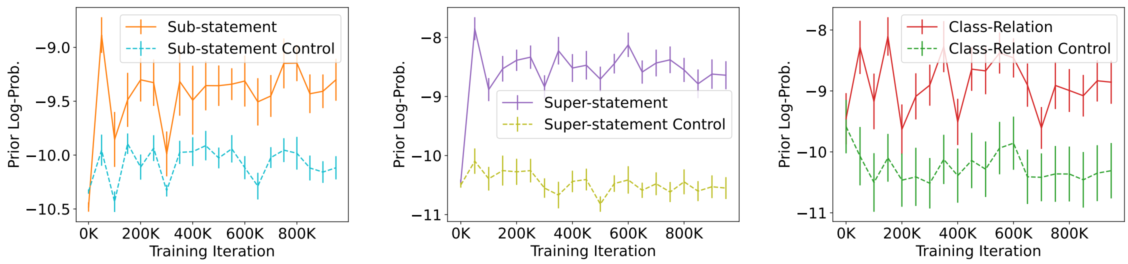
Such knowledge may be supported by simple co-occurrence patterns (such as "robins" and "fly" having high co-occurrence), but we are interested in the extent to which knowledge might also be supported by induced, systematic inference. We focus on the inference of *downward monotonicity* (*A is-a B* $\wedge$ *B has-property C* $\models$ *A has-property C*). We refer to the hypernym property (*B has-property C*) as the super-statement, the hyponym property (*A has-property C*) as the sub-statement, and the hypernymy relation (*A is-a B*) simply as the class relation (Table 1).

We can then evaluate, for example, whether "robins can fly" is influenced by the inference "robins are birds" $\wedge$ "birds can fly" $\models$ "robins can fly." For this evaluation, we inject a supporting premise into a pre-training minibatch (i.e., we replace one of the sentences in the minibatch with the premise) and then evaluate BERT's knowledge of the supported inference after a gradient update on the minibatch containing the premise.

We run this evaluation at intervals throughout the entire pre-training procedure, from random initialization to a fully pre-trained BERT model. If pre-training induces the ability to systematically make the downward monotonicity inference, one would expect that generalization from premise to inference will improve as pre-training progresses.

3.1 Metrics

Let θ_i be the parameterization of BERT at pre-training iteration i, and let $w = \{x, y, z\}$ be a set of knowledge triples where x is a super-statement, y is the corresponding sub-statement, and z is the



(a) Sub-statement (b) Super-statement (c) Class Relation

Figure 1: The prior log-probability of each knowledge type estimated by BERT across pre-training iterations.

corresponding class relation.

Take u to be any logical premise (i.e., $u \in \{x, y, z\}$). Let θ_i^u be θ_i after one gradient update on a minibatch containing u. For a hypothesis $h \in \{x, y, z\}$, we consider:

(1) Prior log-probability: $\log p(h|\theta_i)$

(2) Posterior log-probability: $\log p(h|\theta_i^u)$

(3) PMI: $\log p(h|\theta_i^u) - \log p(h|\theta_i)$

Intuitively, (1) describes the model's prior knowledge of h at step i, and (3) describes how a pre-training update on u affects the knowledge of h. We also consider standard information retrieval metrics such as mean reciprocal rank (MRR).

4 Experiments

4.1 Inference Dataset

We evaluate on the Leap-of-Thought dataset presented by Talmor et al. (2020b). This is a dataset of 30K true or false downward-monotonic inferences which are verbalized using manually written templates. The hypernymy relations are derived from WordNet (Miller, 1995), while the properties are derived from both WordNet and ConceptNet (Speer et al., 2017).

We reformulate this supervised, classification dataset as a zero-shot, cloze-style task. First, we filter the dataset by removing partial examples where one type of knowledge is withheld. Then, we filter out the randomly-generated, negated examples, and those where the object is longer than one word-piece.[1] The filtered dataset consists of 711 examples. Each example is converted into a cloze task by masking the object.

To evaluate relative performance, we also generate a control entity (CE) for each example by randomly sampling a WordNet sibling of the super-statement hypernym as a pseudo-negative (e.g., "A robin is a ____ ." → "fish"). For the super and sub-statements, we take the predicate of the CE under the same relation to be a control (e.g., "Robins can ____ ." → "swim").

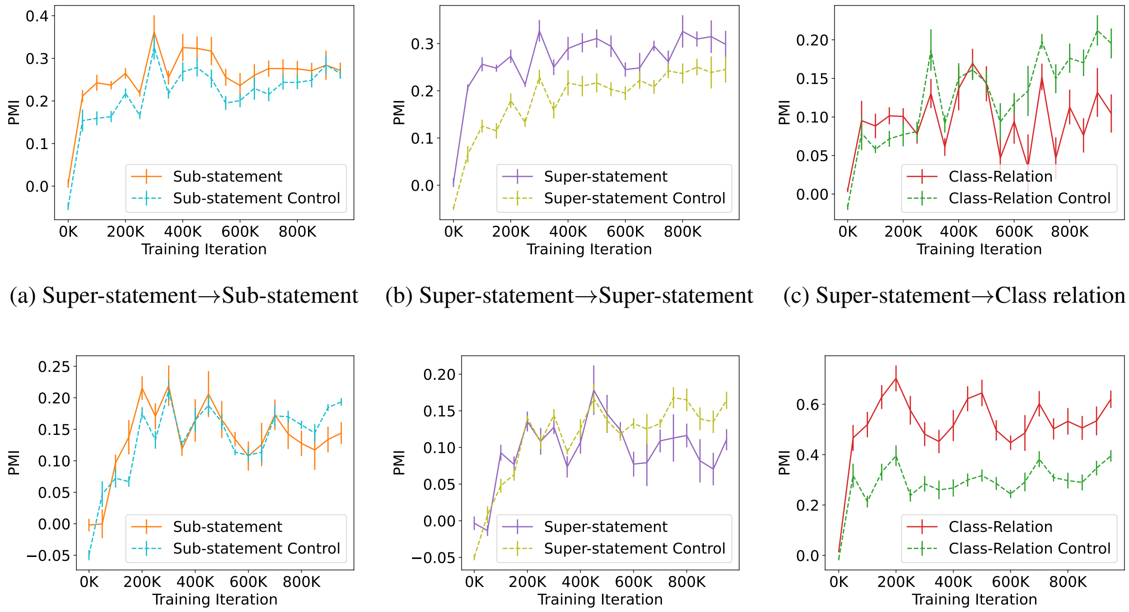
4.2 Model

We consider the training dynamics of a BERT-base model from random initialization to fully pre-trained, replicating details of the original BERT implementation (Devlin et al., 2019).

Specifically, we pre-train the model for 1 million steps on a concatenation of English Wikipedia and the Toronto Book Corpus (Zhu et al., 2015) as released by Huggingface datasets (Lhoest et al., 2021). Training details are given in Appendix A and differ from the original BERT release only in that: 1) we use whole-word masking; 2) we use sentence-order prediction instead of next-sentence prediction as the auxiliary loss (Lan et al., 2020); and, 3) pre-training sentences are extracted using the NLTK Punkt tokenizer (Loper and Bird, 2002) instead of taking random spans of text.

Every 50K pre-training steps, we save a checkpoint of the model's weights and optimizer state. At each checkpoint, we perform the pre-training intervention experiment: we inject 20 random premises into a minibatch and perform one gradient update on this minibatch using the saved optimizer and a constant learning rate of 1e-4 (to control for the effects of the learning rate scheduler). We then evaluate the change in likelihood of h. We perform this evaluation 200 times at each checkpoint so that each of the 711 Leap-of-Thought examples has been evaluated in five separate minibatches.

[1] Evaluating only objects that are a single word-piece follows the procedure of the LAMA evaluation (Petroni et al., 2019) and allows us to evaluate BERT in a zero-shot setting.



(a) Super-statement→Sub-statement (b) Super-statement→Super-statement (c) Super-statement→Class relation

(d) Class relation→Sub-statement (e) Class relation→Super-statement (f) Class relation→Class relation

Figure 2: BERT's generalization from premise to hypothesis across pre-training iterations. Each sub-figure, labelled as $P \rightarrow H$, considers how pre-training on sentences of knowledge type P changes BERT's encoding of supported knowledge of type H. For example, how does a pre-training update on the class relation "robins are ___" → "birds" affect knowledge of the sub-statement "robins can ___" → "fly"?

5 Results

5.1 Model Validation

We first run Talmor et al. (2020b)'s original fine-tuning evaluation on our final BERT checkpoint in order to validate the pre-training procedure. The final implicit reasoning accuracy of our BERT model is 0.89, slightly higher than Talmor et al. (2020b) report for RoBERTa-large. Additional details are presented in Appendix B.

5.2 Pre-training Interventions

Prior prob. Figure 1 shows the prior log-probability of each knowledge type across pre-training. In general, the difference between the correct and control predicates increases during pre-training, suggesting that the knowledge is acquired by BERT. The trend is non-monotonic, however, and interestingly the prior-probability of the correct predicate peaks early in training for all three knowledge types.

Interventions. We evaluate all combinations of knowledge types for premise u and hypothesis h. Some of these inferences are logically sound (e.g., deducing the sub-statement from the super-statement) while others are not (e.g., inducing the super-statement from the sub-statement). We are interested to see when BERT generalizes from u to h as we expect the semantics of the premise to always support the plausibility of the hypothesis relative to the random control.

In Figure 2, we consider PMI for evaluating generalization. When BERT is updated on a pre-training minibatch containing a super-statement, this unsurprisingly increases the probability of the super-statement predicates (Figure 2b) and, as one would expect, there is a similar trend for the class relation (Figure 2f). The control predicates also increase in probability in these cases, but to a lesser extent than the correct predicates.

Less intuitively, however, the PMI of the correct sub-statement predicate is the same as for the control predicate during the final iterations of pre-training (Figure 2a). What's more, the PMI of the class-relation control predicate is higher than the correct predicate during the entire second half of pre-training (Figure 2c). We also see that the control predicate has a higher PMI than the correct predicate when training on the class relation and evaluating on another knowledge type (Figures 2d and 2e).

If knowledge was acquired by induced down-

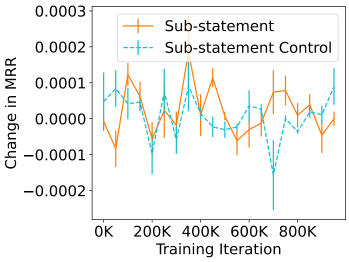


Figure 3: The difference in MRR of predicates before and after updating BERT at each pre-training checkpoint. In this case, we consider MRR of correct and control sub-statement predicates after updating on the corresponding super-statements.

ward monotonicity over semantics, we would expect generalization from class relation to sub-statement to improve over time. The opposite trend suggests knowledge is not being acquired from this semantic inference.

The higher PMI of the control predicate could be in part explained by their lower initial probability, so we also consider changes in MRR (Figure 3). In considering MRR, the difference between predicting the correct and control predicate seems indiscernible across pre-training checkpoints.

6 Conclusion

We show that the ability of BERT to acquire commonsense knowledge from premises and learned inferences does not improve across pre-training, suggesting that the studied knowledge is not acquired from induced semantic inferences.

These results suggest that an explicit reasoning mechanism may be necessary to acquire certain commonsense knowledge.

6.1 Limitations and Future Work

In this work, we only consider one inference type (downward monotonicity) where knowledge is evaluated in one particular way (predicting the predicate) and interventions consist of a single pre-training update. Future work could explore the affects of these experimental design decisions by expanding evaluations to diverse datasets of commonsense inferences and by pre-training for additional steps.

Acknowledgements

We would like to thank the entire team at Microsoft Research Montréal for support with this project. This work was supported by funding from Microsoft Research, and the last author is supported by the Canada CIFAR AI Chair program.

References

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. 2019. Systematic generalization: What is required and can it be learned? In *International Conference on Learning Representations*.

Pratyay Banerjee, Swaroop Mishra, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2021. Commonsense reasoning with implicit knowledge in natural language. In *3rd Conference on Automated Knowledge Base Construction*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Eurene Charniak. 1973. Jack and janet in search of a theory of knowledge. In *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, IJCAI'73, page 337–343, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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, IJCAI-20*, pages 3882–3890. International Joint Conferences on Artificial Intelligence Organization. Main track.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71.

Maxwell Forbes and Yejin Choi. 2017. Verb physics: Relative physical knowledge of actions and objects. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 266–276, Vancouver, Canada. Association for Computational Linguistics.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Nicolas Gontier, Koustuv Sinha, Siva Reddy, and Chris Pal. 2020. Measuring systematic generalization in neural proof generation with transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 22231–22242. Curran Associates, Inc.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, page 25–30, New York, NY, USA. Association for Computing Machinery.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2021. Do language models have beliefs? methods for detecting, updating, and visualizing model beliefs.

Weinan He, Canming Huang, Yongmei Liu, and Xiaodan Zhu. 2021. WinoLogic: A zero-shot logic-based diagnostic dataset for Winograd Schema Challenge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3779–3789, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):6384–6392.

Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *ArXiv*, abs/2001.08361.

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

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *ICML*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1813–1827, Online. Association for Computational Linguistics.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does RoBERTa know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Kanishka Misra, Allyson Ettinger, and Julia Rayz. 2021. Do language models learn typicality judgments from text? In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*.

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

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938, Online. Association for Computational Linguistics.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism.

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*, AAAI'17, page 4444–4451. AAAI Press.

Oyvind Tafjord and Peter Clark. 2021. General-purpose question-answering with macaw.

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, Online. Association for Computational Linguistics.

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

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. 2020b. Leap-of-thought: Teaching pre-trained models to systematically reason over implicit knowledge. In *Advances in Neural Information Processing Systems*, volume 33, pages 20227–20237. Curran Associates, Inc.

Aaron Traylor, Roman Feiman, and Ellie Pavlick. 2021. AND does not mean OR: Using formal languages to study language models' representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 158–167, Online. Association for Computational Linguistics.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models.

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.

Hongming Zhang, Daniel Khashabi, Yangqiu Song, and Dan Roth. 2020. Transomcs: From linguistic graphs to commonsense knowledge. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4004–4010. International Joint Conferences on Artificial Intelligence Organization. Main track.

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2021. Aser: Towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *arXiv preprint arXiv:2104.02137*.

Pei Zhou, Rahul Khanna, Seyeon Lee, Bill Yuchen Lin, Daniel Ho, Jay Pujara, and Xiang Ren. 2021. RICA: Evaluating robust inference capabilities based on commonsense axioms. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7579, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9733–9740.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

A BERT Hyperparameters

We train the BERT-base architecture (12 layers, 12 attention heads, hidden size of 768) following the original pre-training hyperparameters: a batch size of 256, sequence length of 128, and train for 1 million steps. We use the Adam optimizer and linearly warmup the learning rate to 1e-4 over the first 10,000 steps of pre-training, and then linearly decay the learning rate.

Our code builds on the Huggingface Transformers (Wolf et al., 2020) and MegatronLM (Shoeybi et al., 2019) implementations of BERT. The pre-training corpus is uncased and pre-processed using the MegatronLM pre-processing. Training takes four days on eight V100 GPUs.

Our conclusions are based on the training dynamics of BERT-base, and future work might investigate if scaling model size allows for more systematic inferences.

B Leap-of-Thought Fine-tuning Results

The original Leap-of-Thought evaluation consists of fine-tuning BERT to classify if a sub-statement is true given supporting premises. In the explicit reasoning evaluation, all supporting premises are given at test time (e.g., the model must determine if "robins can fly" is true given the context "robins are birds and birds can fly."). In the implicit reasoning evaluation, the class relation is withheld (e.g., the model must determine if "robins can fly" given only the context that "birds can fly." This inference relies on the implicit knowledge that robins are birds). We fine-tune for four epochs following Talmor et al. and otherwise use default hyperparameters.

Our main purpose in running this evaluation is to validate our pre-training procedure; however, we also evaluate all intermediate BERT checkpoints in order to understand how the performance changes across pre-training. Interestingly, we find performance increases log-linearly with pre-training iterations in the implicit reasoning test, but performance of the explicit reasoning evaluation peaks at just 15% of pre-training (Figure 4). Numerical results are presented in Table 2.

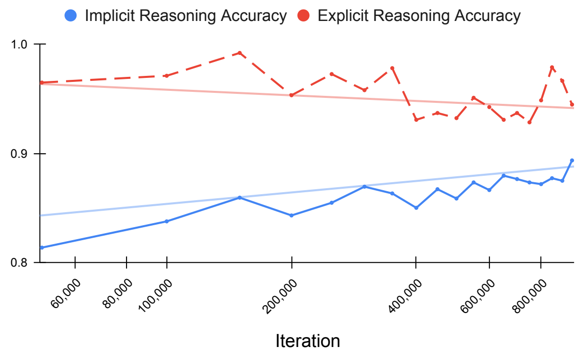


Figure 4: Accuracy on Talmor et al. (2020b)'s original Leap-of-Thought evaluation across pre-training iterations (from 50K to 1M).

Iteration	Implicit	Explicit
0	0.507	0.493
5000	0.507	0.493
10000	0.490	0.490
15000	0.571	0.621
20000	0.625	0.636
30000	0.710	0.763
40000	0.798	0.900
50000	0.814	0.965
100000	0.838	0.971
150000	0.860	0.992
200000	0.843	0.953
250000	0.855	0.973
300000	0.870	0.958
350000	0.863	0.978
400000	0.850	0.931
450000	0.867	0.937
500000	0.859	0.933
550000	0.874	0.951
600000	0.867	0.943
650000	0.880	0.931
700000	0.877	0.937
750000	0.874	0.929
800000	0.872	0.949
850000	0.877	0.979
900000	0.875	0.967
950000	0.894	0.945

Table 2: Fine-tuning accuracy on the original Leap-of-Thought evaluation across pre-training checkpoints.