Cutting-Edge Tutorial: Complex Reasoning over Natural Language

Wenting Zhao
Cornell University
wzhao@cs.cornell.edu

Mor Geva*
Google Research
pipek@google.com

Bill Yuchen Lin*
Allen Institute for AI
yuchenl@allenai.org

Michihiro Yasunaga*
Stanford University
myasu@cs.stanford.edu

Aman Madaan*
Carnegie Mellon University
amadaan@cs.cmu.edu

Tao Yu*
The University of Hong Kong
tyu@cs.hku.hk

1 Tutorial Overview

Teaching machines to reason over texts has been a long-standing goal of natural language processing (NLP). To this end, researchers have designed a diverse set of complex reasoning tasks that involve compositional reasoning (Geva et al., 2021; Trivedi et al., 2022), knowledge retrieval (Yang et al., 2018; Kwiatkowski et al., 2019), grounding (Budzianowski et al., 2018; Xie et al., 2022; Shi et al., 2021), commonsense reasoning (Talmor et al., 2021a; Lin et al., 2020), etc.

A standard choice for building systems that perform a desired type of reasoning is to fine-tune a pretrained language model (LM) on specific downstream tasks. However, recent research has demonstrated that such a straightforward approach is often brittle. For example, Elazar et al. (2021) and Branco et al. (2021) show that, on question-answering (QA) tasks, similar performance can be achieved with questions removed from the inputs. Min et al. (2019), Chen and Durrett (2019), and Tang et al. (2021) show that models trained on multi-hop QA do not generalize to answer single-hop questions. The reasoning capabilities of these models thus remain at a surface level, i.e., exploiting data patterns. Consequently, augmenting LMs with techniques that make them robust and effective becomes an active research area.

We will start the tutorial by providing an overview of complex reasoning tasks where the standard application of pretrained language models fails (in Sec 2). This tutorial then reviews recent promising directions for tackling these tasks (in Sec 3). Specifically, we focus on the following groups of approaches that explicitly consider problem structures: (1) knowledge-augmented methods, where the knowledge is either incorporated during fine-tuning or pretraining; (2) few-shot prompting methods, which effectively guide the models to follow instructions; (3) neuro-symbolic methods, which produce explicit intermediate representations; and, (4) rationale-based methods, one of the most popular forms of the neuro-symbolic methods, which highlight subsets of input as explanations for individual model predictions. The tutorial materials are online at https://wenting-zhao.github.io/complex-reasoning-tutorial.

2 Problem Introduction

We will start with NLP tasks that require reasoning over multiple pieces of information in a provided context, covering various reasoning skills such as fact composition, mathematical reasoning, inferring semantic structures, and reasoning about entities (Yang et al., 2018; Yu et al., 2018; Budzianowski et al., 2018; Dua et al., 2019; Ho et al., 2020; Dasigi et al., 2019; Cobbe et al., 2021; Trivedi et al., 2022). Then, we will discuss benchmarks that combine multiple sources of information (i.e., modalities), e.g., paragraphs, tables, and images (Chen et al., 2020b; Talmor et al., 2021b; Pasupat and Liang, 2015; Chen et al., 2020a).

We will present open-domain setups where external knowledge should be integrated into the reasoning process (Geva et al., 2021; Onoe et al., 2021; Ferguson et al., 2020; Talmor and Berant, 2018). In addition, we will review tasks that require commonsense reasoning (Talmor et al., 2021a; Rudinger et al., 2020; Sap et al., 2019; Saha et al., 2021).

We will conclude this part by highlighting key practices for dataset creation, that increase data diversity and minimize annotation biases and reasoning shortcuts (Bartolo et al., 2020; Khot et al., 2020; Geva et al., 2019; Parmar et al., 2022).

3 Approaches

(1a) Knowledge-Augmented Fine-Tuning Tackling complex reasoning problems that require commonsense knowledge and entity-centric facts can

*Equal Contribution.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 6: Tutorial Abstracts, pages 11–20
July 9, 2023 ©2023 Association for Computational Linguistics

benefit from access to external knowledge sources. How to incorporate knowledge during fine-tuning has thus been extensively studied. A general method is to retrieving knowledge facts relevant to given situations (e.g., questions) and fusing them with an LM-based neural module. External knowledge can be categorized into three forms: structured (e.g., knowledge graphs like ConceptNet (Speer et al., 2017)), unstructured (e.g., knowledge corpora such as Wikipedia and GenericsKB (Bhakthavatsalam et al., 2020)), and instance-based (i.e., annotated examples).

In this section, we will cover methods for these three forms of knowledge in a variety of reasoning problems. For structured knowledge, KagNet (Lin et al., 2019) is a typical method that focuses on fusing retrieved subgraphs from ConceptNet for fine-tuning LMs to perform commonsense reasoning. Follow-up works include MHGRN (Feng et al., 2020), QA-GNN (Yasunaga et al., 2021), and GreaseLM (Zhang et al., 2022b). For unstructured knowledge, we will introduce methods that encode a large knowledge corpus as neural memory modules to support knowledge retrieval for reasoning. We will start with DPR (Karpukhin et al., 2020), one of the most popular methods that embed Wikipedia as a dense matrix of fact embeddings. Then, we will cover DrKIT (Dhingra et al., 2020), which improves multi-hop reasoning ability by encoding sparse entity mentions. Additionally, we introduce DrFact (Lin et al., 2021), a fact-level extension for DrKIT that focuses on commonsense reasoning. For instance-based knowledge, a promising direction, we will also introduce methods such as RACo (Yu et al., 2022b), ReCross (Lin et al., 2022), and QEDB (Chen et al., 2022b), which aim to exploit annotated examples to enhance reasoning.

(1b) Knowledge-Augmented Pretraining. Pretraining performs self-supervised learning of representations from large-scale data, which holds the potential to help a broader range of downstream tasks. We will review recent efforts to incorporate knowledge and reasoning abilities into LMs during pretraining. We first discuss retrieval-augmented pretraining (Guu et al., 2020; Lewis et al., 2020a; Borgeaud et al., 2021; Yasunaga et al., 2022b), which retrieves relevant documents from an external memory and feeds them to the model as an additional input. This helps not only knowledge-intensive tasks but also some reasoning-intensive tasks because the models learn to process multiple documents for multi-hop reasoning (Yasunaga et al., 2022b). We then discuss works that integrate structured knowledge bases/graphs. For example, some use knowledge graphs to make additional pretraining objectives for LMs (Xiong et al., 2020; Shen et al., 2020; Wang et al., 2021; Liu et al., 2021; Yu et al., 2022a; Ke et al., 2021); others retrieve and feed entity or knowledge graph information as a direct input to the model (Zhang et al., 2019; Rosset et al., 2020; Liu et al., 2020; Sun et al., 2021; Agarwal et al., 2021; Sun et al., 2020; He et al., 2020; Yasunaga et al., 2022a). Recent works show that these retrieved knowledge graphs can provide LMs with scaffolds for performing complex reasoning over entities, such as logical and multi-hop reasoning (Yasunaga et al., 2022a).

(2) Few-Shot Prompting Approaches. The rise of large pretrained LMs, such as GPT-3 (Brown et al., 2020), OPT (Zhang et al., 2022a), and PaLM (Chowdhery et al., 2022), has unlocked the potential of few-shot prompting methods for a wide range of reasoning tasks. However, despite their strengths, these LMs in the few-shot prompting mode have peculiar failure modes, especially when it comes to complex reasoning tasks (Marcus, 2022). Further, the *prompt* has to be designed carefully, and it has been shown that seemingly innocuous changes to the prompt (e.g., order of examples or the format of text) can drastically impact the performance (Le Scao and Rush, 2021; Mishra et al., 2021). In response, several techniques have been developed to make few-shot prompting methods to be less susceptible to the exact prompt choice. This section will cover both a high-level overview of few-shot prompting and introduce specific classes of techniques that can further improve the few-shot prompting methods on complex reasoning tasks.

First, we will introduce prompt-design techniques like chain-of-thought prompting (Wei et al., 2022b) and least-to-most prompting (Wei et al., 2022c), which encourage an LM to generate reasoning steps as part of the solution, helping with problem decomposition and enhanced reasoning. Next, we will cover techniques that change the prompt dynamically for each input query. The methods covered in this part include selecting the training examples in the prompt (Liu et al., 2022a) and editing the prompt to incorporate feedback received on a similar-input (Madaan et al., 2022a).

Finally, we will cover techniques that leverage code-generation models for complex reasoning tasks. Representative techniques in this part will cover i) the use of code-generation model for structured commonsense reasoning (Madaan et al., 2022b), ii) algorithmic reasoning by expanding detailed instructions in the prompt (Zhou et al., 2022), and iii) generating chain-of-thought styled reasoning chains in Python code to tackle complex symbolic reasoning tasks (Gao et al., 2022).

(3) Neuro-Symbolic Approaches. Although performance on NLP tasks is dominated by neural *end-to-end* systems that directly map inputs to outputs (Devlin et al., 2019; Raffel et al., 2020), these approaches lack interpretability and robustness. *Symbolic* approaches, on the other hand, produce explicit intermediate reasoning trajectories such as logical forms, reasoning paths, or program code, which might then be executed to derive a final output (Zettlemoyer and Collins, 2005; Chen et al., 2019b, *i.a.*). Compared to both end-to-end and chain-of-thought methods (Wei et al., 2022a, *i.a.*), the reasoning processes produced by the symbolic methods are interpretable, and the resulting execution makes them more robust to input changes.

Researchers (Andreas et al., 2016; Liang et al., 2017; Gupta et al., 2019; Khot et al., 2021; Zhu et al., 2022; Cheng et al., 2022; Gao et al., 2022; Schick et al., 2023, *i.a.*) also propose to combine neural modules and symbolic components to leverage advantages of both approaches. More specifically, Neural-Symbolic Machines (Liang et al., 2017) adopt a seq-to-seq model to generate programs and a Lisp interpreter that performs program execution. (Chen et al., 2019b) designs a domain-specific language for question answering over text. BREAK (Wolfson et al., 2020) proposes a meaningful representation, QDMR, that decomposes the question into multiple steps. Thorne et al. (2021) propose a mixed pipeline of logic forms and neural networks, aiming at solving the scale problem and noisy, messy data over a natural language database.

Another stream of works called neural module networks (Andreas et al., 2016; Das et al., 2018; Gupta et al., 2019) propose to generate symbolic programs that are further softly executed by the corresponding neural modules. Khot et al. (2021) propose text module networks to solve complex tasks by decomposing them into simpler ones solvable by existing QA models and a symbolic calculator. However, most prior neural-symbolic methods require the elaborate human design of the symbolic language and the calibration of corresponding neural modules to tackle problems in a specific domain with large training data. Recently, Cheng et al. (2022) propose Binder, a new neural-symbolic system based on GPT-3 Codex (Chen et al., 2021) that supports *flexible* neural module calls that will enable *higher coverage* for the symbolic language, while only requiring *few annotations*. Also, Gao et al. (2022) introduce PAL, a new method based on Codex that generates executable programs as the intermediate reasoning steps and leverages a Python interpreter to derive final answers.

This section will begin by discussing the high-level comparison among the end-to-end, chain-of-thought, symbolic (e.g., semantic parsing), and neural-symbolic approaches. We will then move to provide a high-level overview of different neural-symbolic approaches. In this part, we will mainly focus on neural-symbolic approaches with LMs. Finally, we will cover recent techniques incorporating GPT-3 Codex in neural-symbolic approaches.

(4) Rationale-Based Approaches. Rationale-based approaches extract parts of input to be *reasoning certificates*, offering end users a way to evaluate the trustworthiness of the predictions. Based on reasoning types, rationales of different granularity are identified – they can be tokens, sentences, or documents (DeYoung et al., 2020; Kwiatkowski et al., 2019). NLP systems can benefit from rationales in several ways. Yang et al. (2018) show that providing rationales as additional supervision improves models' capacity to perform multi-hop reasoning. More recently, Chen et al. (2022a) demonstrate the potential of using such methods to build more robust NLP systems.

Existing methods for extracting rationales often require supervision; they either apply multi-task loss functions (Joshi et al., 2020; Groeneveld et al., 2020), or design specialized network architectures to incorporate inductive biases (Tu et al., 2019; Fang et al., 2020). Because rationale annotations are expensive to collect and not always available, recent effort has been devoted to semi-supervised and unsupervised methods. Chen et al. (2019a) leverage entity taggers to build silver reasoning chains used for rationale supervision. Glockner et al. (2020) and Atanasova et al. (2022) design unsupervised objectives for extracting rationales in multi-hop QA systems. Finally, latent-variable approaches are a natural fit for unsupervised learning (Lei et al., 2016; Zhou et al., 2020; Lewis et al., 2020b). By modeling rationales as a latent variable, it provides a principled way to explicitly impose constraints in the reasoning process.

3.1 Schedule

1. Introduction & Motivations (15 min.)
2. Benchmarks & Evaluation (25 min.)
3. Knowledge-augmented Fine-tuning (25 min.)
4. Knowledge-augmented Pretraining (25 min.)
5. Break (30 minutes)
6. Neuro-Symbolic Approaches (25 min.)
7. Few-shot Prompting Approaches (25 min.)
8. Rationale-Based Approaches (25 min.)
9. Concluding discussion (15 min.)

4 Instructor information

Wenting Zhao is a Ph.D. student in Computer Science at Cornell University. Her research focuses on the intersection of reasoning and NLP. She is especially interested in developing explainable methods for complex reasoning problems.

Mor Geva is a postdoctoral researcher, now at Google Research and previously at the Allen Institute for AI. Her research focuses on debugging the inner workings of black-box NLP models, to increase their transparency, control their operation, and improve their reasoning abilities. She is organizing the next edition of the Workshop on Commonsense Reasoning and Representation.

Bill Yuchen Lin is a postdoctoral researcher at the Allen Institute for AI. He obtained his Ph.D. at USC advised by Prof. Xiang Ren. His research goal is to teach machines to think, talk, and act with commonsense knowledge and commonsense reasoning ability as humans do. He was a co-author of the tutorial on *Knowledge-Augmented Methods for Natural Language Processing* and the *Workshop on Commonsense Representation and Reasoning* at ACL 2022.

Michihiro Yasunaga is a Ph.D. student in Computer Science at Stanford University. His research interest is in developing generalizable models with knowledge, including commonsense, science, and reasoning abilities. He co-organized the Workshop on Structured and Unstructured Knowledge Integration (SUKI) at NAACL 2022.

Aman Madaan is a Ph.D. student at the School of Computer Science, Carnegie Mellon University. He is interested in large language models, feedback-driven generation, and the intersection of code generation and natural language reasoning. He helped organize the 1st and 2nd Workshops on Natural Language Generation, Evaluation, and Metrics (GEM) at ACL 2021 and EMNLP 2022.

Tao Yu is an assistant professor of computer science at The University of Hong Kong. He completed his Ph.D. at Yale University and was a postdoctoral fellow at the University of Washington. He works on executable language understanding, such as semantic parsing and code generation, and large LMs. Tao is the recipient of an Amazon Research Award. He co-organized multiple workshops in Semantic Parsing and Structured and Unstructured Knowledge Integration at EMNLP and NAACL.

5 Other Information

Reading List Rogers et al. (2022); Storks et al. (2019); Liu et al. (2022b); Lyu et al. (2022); Wiegreffe and Marasović (2021); Andreas et al. (2016); Cheng et al. (2022); Gao et al. (2022).

Breadth We estimate that approximately 30% of the tutorial will center around work done by the presenters. This tutorial categorizes promising approaches for complex reasoning tasks into several groups, and each of this group includes a significant amount of other researchers' works.

Diversity considerations The challenges of building robust and generalizable NLP systems exist in every language. The methods covered in this tutorial are language-agnostic and can be extended to non-English context.

For instructors, they all have different affiliations (i.e., Cornell, Google, Stanford, USC, HKU, and CMU). They are three PhD students, two postdoctoral researchers, and one assistant professor; two of the instructors are female.

Prerequisites Following knowledge is assumed:

- Machine Learning: basic probability theory, supervised learning, transformer models
- NLP: Familiarity with pretrained LMs; standard NLP tasks such as question answering, text generation, etc.

Estimated number of participants 150.

Preferable venue ACL.

Targeted audience Researchers and practitioners who seek to develop a background in complex reasoning tasks where standard application of pretrained language models fail. By providing a systematic overview of recent promising approaches for these tasks, this tutorial hopefully reveals new research opportunities to the audience.

References

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. 2021. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Diagnostics-guided explanation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10445–10453.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *ArXiv*, abs/2005.00660.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2021. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*.

Ruben Branco, António Branco, João António Rodrigues, and João Ricardo Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521, Online and Punta Cana, Dominican Republic. 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, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022a. Can rationalization improve robustness? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3792–3805, Seattle, United States. Association for Computational Linguistics.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. 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 4026–4032, Minneapolis, Minnesota. Association for Computational Linguistics.

Jifan Chen, Shih-ting Lin, and Greg Durrett. 2019a. Multi-hop question answering via reasoning chains. *arXiv preprint arXiv:1910.02610*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*.

Wenhu Chen, William W. Cohen, Michiel de Jong, Nitish Gupta, Alessandro Presta, Pat Verga, and John Wieting. 2022b. Qa is the new kr: Question-answer pairs as knowledge bases. *ArXiv*, abs/2207.00630.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020a. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Xinyun Chen, Chen Liang, Adams Wei Yu, Denny Zhou, Dawn Song, and Quoc V Le. 2019b. Neural symbolic reader: Scalable integration of distributed and symbolic representations for reading comprehension. In *International Conference on Learning Representations*.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2022. Binding language models in symbolic languages. *ArXiv*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Abhishek Das, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Neural modular control for embodied question answering. In *Conference on Robot Learning*, pages 53–62. PMLR.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. 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 5925–5932, Hong Kong, China. Association for Computational Linguistics.

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

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Bhuwan Dhingra, Manzil Zaheer, Vidhisha Balachandran, Graham Neubig, Ruslan Salakhutdinov, and William W. Cohen. 2020. Differentiable reasoning over a virtual knowledge base. In *International Conference on Learning Representations*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to square one: Artifact detection, training and commonsense disentanglement in the Winograd schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, Online. Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

James Ferguson, Matt Gardner, Hannaneh Hajishirzi, Tushar Khot, and Pradeep Dasigi. 2020. IIRC: A dataset of incomplete information reading comprehension questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1137–1147, Online. Association for Computational Linguistics.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. 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 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Max Glockner, Ivan Habernal, and Iryna Gurevych. 2020. Why do you think that? exploring faithful sentence-level rationales without supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1080–1095, Online. Association for Computational Linguistics.

Dirk Groeneveld, Tushar Khot, Mausam, and Ashish Sabharwal. 2020. A simple yet strong pipeline for HotpotQA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8839–8845, Online. Association for Computational Linguistics.

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. 2019. Neural module networks for reasoning over text. *arXiv preprint arXiv:1912.04971*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning (ICML)*.

Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020. Integrating graph contextualized knowledge into pre-trained language models. In *Findings of EMNLP*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Pei Ke, Haozhe Ji, Yu Ran, Xin Cui, Liwei Wang, Linfeng Song, Xiaoyan Zhu, and Minlie Huang. 2021. Jointgt: Graph-text joint representation learning for text generation from knowledge graphs. In *Findings of ACL*.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8082–8090.

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2021. Text modular networks: Learning to decompose tasks in the language of existing models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1264–1279.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chen Liang, Jonathan Berant, Quoc Le, Kenneth D. Forbus, and Ni Lao. 2017. Neural symbolic machines: Learning semantic parsers on Freebase with weak supervision. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada. Association for Computational Linguistics.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. 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 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Bill Yuchen Lin, Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Xiang Ren, and William Cohen. 2021. Differentiable open-ended commonsense reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4611–4625, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Kangmin Tan, Chris Miller, Beiwen Tian, and Xiang Ren. 2022. Unsupervised cross-task generalization via retrieval augmentation. *ArXiv*, abs/2204.07937.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022a. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2022b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* Just Accepted.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and P. Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *AAAI Conference on Artificial Intelligence*.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2022. Towards faithful model explanation in nlp: A survey. *arXiv preprint arXiv:2209.11326*.

Aman Madaan, Niket Tandon, Peter Clark, and Yiming Yang. 2022a. Memory-assisted prompt editing to improve GPT-3 after deployment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2833–2861, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022b. Language models of code are few-shot commonsense learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1384–1403, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gary Marcus. 2022. Experiments Testing GPT-3's Ability at Commonsense Reasoning: Results. Accessed: 2022-08-15.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2021. Reframing Instructional Prompts to GPTk's Language. *arXiv preprint arXiv:2109.07830*.

Yasumasa Onoe, Michael Zhang, Eunsol Choi, and Greg Durrett. 2021. Creak: A dataset for commonsense reasoning over entity knowledge. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. 2022. Don't blame the annotator: Bias already starts in the annotation instructions. *arXiv preprint arXiv:2205.00415*.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

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. *JLMR*, 21.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2022. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.* Just Accepted.

Corby Rosset, Chenyan Xiong, Minh Phan, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Knowledge-aware language model pretraining. *arXiv preprint arXiv:2007.00655*.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *ArXiv*, abs/2302.04761.

Tao Shen, Yi Mao, Pengcheng He, Guodong Long, Adam Trischler, and Weizhu Chen. 2020. Exploiting structured knowledge in text via graph-guided representation learning. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Qi Shi, Yu Zhang, Qingyu Yin, and Ting Liu. 2021. Logic-level evidence retrieval and graph-based verification network for table-based fact verification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *ArXiv*, abs/1612.03975.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *CoRR*, abs/1904.01172.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuan-Jing Huang, and Zheng Zhang. 2020. Colake: Contextualized language and knowledge embedding. In *International Conference on Computational Linguistics (COLING)*.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021a. CommonsenseQA 2.0: Exposing the limits of AI through gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021b. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*.

Yixuan Tang, Hwee Tou Ng, and Anthony Tung. 2021. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3244–3249, Online. Association for Computational Linguistics.

James Thorne, Majid Yazdani, Marzieh Saeidi, Fabrizio Silvestri, Sebastian Riedel, and Alon Halevy. 2021. Database reasoning over text. *arXiv preprint arXiv:2106.01074*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. 2019. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2704–2713, Florence, Italy. Association for Computational Linguistics.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics (TACL)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022a. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022c. Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. In *Proceedings of NeurIPS*.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. 2020. Break it down: A question understanding benchmark. *Transactions of the Association for Computational Linguistics*.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *EMNLP*.

Wenhan Xiong, Jingfei Du, William Yang Wang, and Veselin Stoyanov. 2020. Pretrained encyclopedia: Weakly supervised knowledge-pretrained language model. In *International Conference on Learning Representations (ICLR)*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. 2022a. Deep bidirectional language-knowledge graph pretraining. In *Neural Information Processing Systems (NeurIPS)*.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022b. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2022a. Jaket: Joint pre-training of knowledge graph and language understanding. In *AAAI Conference on Artificial Intelligence*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

W. Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022b. Retrieval augmentation for commonsense reasoning: A unified approach. *ArXiv*, abs/2210.12887.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *UAI*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022a. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022b. GreaseLM: Graph REASoning enhanced language models. In *International Conference on Learning Representations*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Association for Computational Linguistics (ACL)*.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning. *arXiv preprint arXiv:2211.09066*.

Wangchunshu Zhou, Jinyi Hu, Hanlin Zhang, Xiaodan Liang, Maosong Sun, Chenyan Xiong, and Jian Tang. 2020. Towards interpretable natural language understanding with explanations as latent variables. *Advances in Neural Information Processing Systems*, 33:6803–6814.

Zhaocheng Zhu, Mikhail Galkin, Zuobai Zhang, and Jian Tang. 2022. Neural-symbolic models for logical queries on knowledge graphs. *arXiv preprint arXiv:2205.10128*.