PhD Subject
Integrating External Knowledge into Language Models

1 Supervision

- Directeur: Joseph Le Roux
- Co-encadrant: Nadi Tomeh

2 Introduction

Recent advancements in Natural Language Processing (NLP) have been driven by the emergence of Large language models (LLMs) trained on vast amounts of data, which have been able to achieve state-of-the-art performance on a variety of natural language tasks. LLMs' versatility and reasoning capabilities have made them a powerful tool for general-purpose language generation tasks. However, they struggle when processing tasks and scenarios involving data specific to a specialized field such as healthcare or finance, or real-time data that was unavailable when the model was trained. This is primarily due to their inability to access the latest domain-specific data and inadequate domain-specific reasoning. I propose exploring two paradigms of incorporating external knowledge into language models (LMs) to address these limitations - knowledge-aware inference and architecture-level knowledge integration.

3 Knowledge-aware inference

Language models, even LLMs, struggle to learn long-tail knowledge, i.e. knowledge that occurs less frequently such as obscure scientific facts and niche pieces of information, from pre-training data (Kandpal et al., 2023). They are often supplemented this data through additional fine-tuning in order to perform well on domain-specific benchmarks. Retrieval-augmented Generation (RAG) could be used to overcome this lack of access to external knowledge. This typically involves using the input query to search and retrieve relevant indexed documents, and adding them to the LM input prompt as context. Figure 1 provides an example of RAG being used to answer an input query. Prior work has also shown that retrieval improves performance across a variety of natural language processing (NLP) tasks in isolation (Chen et al., 2017; Moghe et al., 2018; Khandelwal et al., 2020). However, the query complexity and the potential necessity to retrieve and reason using multiple pieces of information (as in the case of the example query in Figure 2) make efficient retrieval a challenging task for domain-specific data. In this project, I aim to develop methods of converting the input query into a series of retrieval sub-problems and using reasoned context to aid a language model in resolving tasks requiring domain-specific data.

3.1 Proposed Work

The chief focus of this approach will be in exploring methods to breakdown a given input query into retrieval sub-problems and combining the retrieved contexts. I plan on using a framework similar to step-wise prompting techniques such as Chain-of-thought (CoT) (Wei et al., 2023) and self-ask prompting (Press et al., 2023), where the problem statement is broken down into simpler sub-problems. However, instead of relying on the

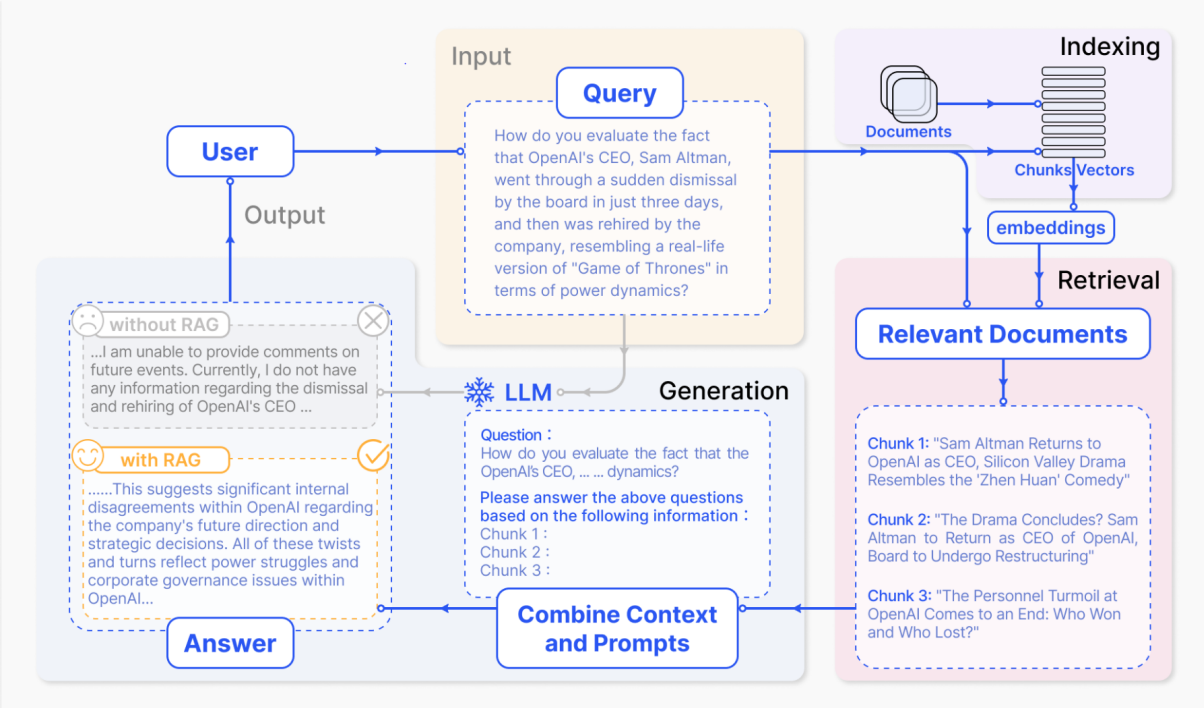

Figure 1: A representative instance of the RAG process applied to question answering. It mainly consists of 3 steps. 1) Indexing. Documents are split into chunks, encoded into vectors, and stored in a vector database. 2) Retrieval. Retrieve the Top k chunks most relevant to the question based on semantic similarity. 3) Generation. Input the original question and the retrieved chunks together into LLM to generate the final answer. Source: Gao et al. (2024).

model responses like the aforementioned approaches, I plan to focus on using the sub-tasks to retrieve relevant external knowledge. Zhou et al. (2023) offer one possible solution to create these sub-tasks by adding a separate problem decomposition stage where a language model is asked to decompose a given problem into sub-problems, and appending the sub-problem and solution pairs to the original problem to produce the reasoned output. I'm also interested in evaluating the utility of LMs for resolving sub-problems with the help of retrieved context(s), and checking if they can be used for rephrasing, re-ranking or making inferences using the retrieved contexts to improve the quality of information eventually used to resolve the input query.

4 Architecture-level knowledge integration

While prompt-level augmentation can be used to provide relevant contexts to an LM, it still needs to understand and process the domain-specific knowledge efficiently. As evidenced by studies evaluating LMs' understanding of domain-specific knowledge (Poerner et al., 2019; Kassner and Schütze, 2020), models do well in reasoning about surface-level entities but fail to capture the rich factual knowledge require to reason efficiently on domain-specific knowledge. Typically, fine-tuning LMs has also been the primary method of incorporating domain-knowledge understanding and up to date knowledge (Tian et al., 2023) into LMs. This resulted in the emergence of multiple domain-specific variants of smaller, BERT-based (Devlin et al., 2019) architectures such as BioBERT (Lee et al., 2019), FinBERT (Araci, 2019) and SciBERT (Beltagy et al., 2019). However, this

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Figure 2: An example query that requires retrieving and inferring the answer from multiple pieces of information such as the name of a band, its former members and news related to former members. Source: (Yang et al., 2018)

approach requires LLMs to be separately fine-tuned for every domain-specific task, and updated every time the underlying knowledge base is updated, both of which require a large amount of compute time and resources. As such, this project aims to develop efficient methods of integrating external knowledge which can be used to overcome these shortcomings.

4.1 Proposed Work

I plan on using adapter layers (Houlsby et al., 2019) to augment intermediate representations of an LM with additional knowledge to improve domain-knowledge reasoning. Past research on adapters has focused primarily on developing adapter layers as parameter-efficient fine-tuning methods without the involvement of external knowledge and have demonstrated significant improvements comparable to end-to-end fine-tuning methods. More recent works (Wang et al., 2020; Wang et al., 2024) have also demonstrated that adapters can be an effective way to add external knowledge understanding to pre-trained LMs. However, the extent and method of adding knowledge using adapters is yet to be fully investigated. As a part of this project, I will be focusing on modifying existing adapter mechanisms such as AdapterFusion (Pfeiffer et al., 2021) and UniPELT (Mao et al., 2022) to incorporate existing external knowledge. This existing external knowledge comes in two forms - long-tail knowledge from highly specific domains of data (such as geology, sailing and soccer) that the model struggles to learn from training data even if it were present, and the latest pieces of information that the language models did not have access to during training (such as recent news articles and the latest academic papers).

While past methods of incorporating knowledge have relied on fully fine-tuning (Sun et al., 2019) or re-training models (Wu et al., 2023), I'm interested in following recent works (Meng et al., 2021; Emelin et al., 2022) and injecting knowledge without modifying the pre-trained weights of the underlying LM. In subsequent stages, I will be studying the

effectiveness of methods incorporating these blocks into an LM's architecture. Approaches such as Wang et al. (2020) add adapter blocks alongside an LM's existing architecture, and only change the task-specific head to add the generated encodings, whereas approaches such as AdapterFusion add adapter blocks and modify the intermediate outputs of the LM themselves. Li and Liang (2021) describe another possible way of adding adapters with encoded knowledge as a prefix to the input encodings.

5 Evaluation

I'm interested in using a collection of benchmarks spanning multiple tasks and domains to evaluate the robustness and effectiveness of the proposed approaches. An overview of the types of tasks along with examples of existing benchmarks is given below -

5.1 Question-Answering

Question-Answering tasks such as SQuAD (Rajpurkar et al., 2016), MS MACRO (Bajaj et al., 2018) and PopQA (Mallen et al., 2023) require single-hop reasoning, where as more complex question-answering datasets such as HotpotQA (Yang et al., 2018) and 2WikiMultiHopQA (Ho et al., 2020) examine an LM's multi-hop reasoning ability, and often require the retrieval of multiple pieces of information. In addition to the aforementioned tasks that primarily rely on general and factual knowledge, tasks such as QASPER (Dasigi et al., 2021), InsuranceQA (Feng et al., 2015), CaseHOLD (Zheng et al., 2021), MedQUAD (Ben Abacha and Demner-Fushman, 2019) and RadQA (Soni et al., 2022) contain references to unseen or long-tail knowledge, and serve as important benchmarks in evaluating the knowledge access and reasoning capabilities granted by the developed approaches to LMs.

5.2 Information Extraction

Information Extraction tasks often require the candidate model to be able to understand and extract specific parts of the given corpora, making them suitable benchmarks for evaluating the developed approaches.

Named entity recognition datasets such as the NCBI disease (Doğan et al., 2014), SEC filings (Salinas Alvarado et al., 2015) and SciERC (Luan et al., 2018) evaluate the ability of a model to identify named entities in biomedical data, financial data and computer science literature respectively, and serve as direct points of comparison against end-to-end fine-tuned models such as FinBERT, BioBERT and SciBERT.

Event argument and relation extraction is useful for numerous downstream tasks that involve forming insights from unstructured documents, and often require LMs to have knowledge about obscure and infrequently occuring concepts. Event argument extraction tasks such as WikiEvent (Li et al., 2021) and RAMS (Ebner et al., 2020) have been commonly used to evaluate retrieval-based approaches in the past and serve as important benchmarks. ZsRE proposes a relation extraction task using wikipedia entries and relevant sentences for extracting each relation, allowing us to examine the efficacy of the approach independent of the retrieval system. Biomedical relation extraction tasks such as GAD (Bravo et al., 2015) and CHEMPROT (Islamaj Doğan et al., 2019) require extensive understanding of protein interactions, and relations between genes and diseases.

5.3 Dialogue generation

With the recent push for dialogue-based models and dialogue-based search agents, augmenting dialogue generation with external knowledge has become an important task with

considerable real-world applications, and can benefit greatly from the external knowledge approaches proposed. As such, I'm interested in evaluating the proposed approaches on benchmarks such as the wizard of Wikipedia (Dinan et al., 2019) and Topical-Chat (Gopalakrishnan et al., 2023) datasets to evaluate the feasibility of using the proposed approaches in a dialogue generation setting.

5.4 Other Tasks

I'd also like to examine the effectiveness of the proposed approaches on tasks such as fact verification using the FEVER and PubHealth benchmarks, which often require LMs to perform multi-step reasoning on unseen domains using credible evidence. Additionally, the Massive Multitask Language Understanding benchmark evaluates LMs on a total of 57 academic domains including specialized areas like law and ethics, making it a comprehensive benchmark for evaluating the proposed approaches.

6 Conclusion

In conclusion, the proposed research aims to investigate and develop approaches to incorporate external knowledge into language models, with the ultimate goal of making it easier to utilize state-of-the-art language models in various domain-specific tasks. Your support for this research endeavor is highly appreciated and I look forward to the contributions it can make towards the academic community.

References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1), October.

Àlex Bravo, Janet Piñero, Núria Queralt-Rosinach, Michael Rautschka, and Laura I Furlong. 2015. Extraction of relations between genes and diseases from text and large-scale data analysis: Implications for translational research. *BMC Bioinformatics*, 01.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada, July. Association for Computational Linguistics.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47:1–10.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online, July. Association for Computational Linguistics.

Denis Emelin, Daniele Bonadiman, Sawsan Alqahtani, Yi Zhang, and Saab Mansour. 2022. Injecting domain knowledge in language models for task-oriented dialogue systems.

Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2023. Topical-chat: Towards knowledge-grounded open-domain conversations.

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 Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

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.

Rezarta Islamaj Doğan, Sun Kim, Andrew Chatr-aryamontri, Chih-Hsuan Wei, Donald C Comeau, Rui Antunes, Sérgio Matos, Qingyu Chen, Aparna Elangovan, Nagesh C Panyam, Karin Verspoor, Hongfang Liu, Yanshan Wang, Zhuang Liu, Berna Altınel, Zehra Melce Hüsünbeyi, Arzucan Özgür, Aris Fergadis, Chen-Kai Wang, Hong-Jie Dai, Tung Tran, Ramakanth Kavuluru, Ling Luo, Albert Steppi, Jinfeng Zhang, Jinchan Qu, and Zhiyong Lu. 2019. Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine. *Database*, 2019:bay147, 01.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR, 23–29 Jul.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 4582–4597, Online, August. Association for Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories.

Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Wen tau Yih, and Madian Khabsa. 2022. Unipelt: A unified framework for parameter-efficient language model tuning.

Zaiqiao Meng, Fangyu Liu, Thomas Clark, Ehsan Shareghi, and Nigel Collier. 2021. Mixture-of-partitions: Infusing large biomedical knowledge graphs into BERT. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4672–4681, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Nikita Moghe, Siddhartha Arora, Suman Banerjee, and Mitesh M. Khapra. 2018. Towards exploiting background knowledge for building conversation systems. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium, October-November. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. Bert is not a knowledge base (yet): Factual knowledge vs. name-based reasoning in unsupervised qa, 11.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In Ben Hachey and Kellie Webster, editors, *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia, December.

Sarvesh Soni, Meghana Gudala, Atieh Pajouhi, and Kirk Roberts. 2022. RadQA: A question answering dataset to improve comprehension of radiology reports. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6250–6259, Marseille, France, June. European Language Resources Association.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters.

Fali Wang, Runxue Bao, Suhang Wang, Wenchao Yu, Yanchi Liu, Wei Cheng, and Haifeng Chen. 2024. Infuserki: Enhancing large language models with knowledge graphs via infuser-guided knowledge integration.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models.