Commonsense Retrieval-Augmented Generation for Large Language Models

CPSC 532V 2023W2 Project Proposal

Juntai Cao[♠] Student #: 50171404 jtcao7@cs.ubc.ca

Yilin Yang[♠] Student #: 24754350 yangyl17@cs.ubc.ca

Yuwei Yin[♠] Student #: 36211928 yuweiyin@cs.ubc.ca

Department of Computer Science, University of British Columbia
2366 Main Mall, Vancouver, BC V6T 1Z4, Canada

1 Introduction

Natural language generation (NLG) has improved considerably owing to the rapid development of large language models (LLMs) (Touvron et al., 2023; OpenAI, 2023; Zhao et al., 2023b). After training on massive corpora and tuning in an instruction-following way (Ouyang et al., 2022; Bai et al., 2022), LLMs can generate fluent and coherent responses in a human-like fashion (OpenAI, 2022). However, the generation process suffers from the hallucination problem (Ji et al., 2023) because LLMs tend to make up plausible answers regardless of whether they understand the question and context.

LLMs frequently encounter challenges in producing satisfactory answers when confronted with tasks that demand commonsense reasoning (Sap et al., 2020), which makes the hallucination problem especially severe. This is rooted in the language modeling training paradigm (Vaswani et al., 2017; Radford et al., 2018), in which LLM models predict the next token based on the previously generated ones. Hence, the models are supposed to produce better output if they are conditioned on more relevant context for solving the question.

Further exploration is needed to determine ***how to effectively incorporate contextual information to enhance problem-solving performance and mitigate hallucination in LLM generation***. Various methods have been proposed in recent years to provide the models with such context. In-context learning (ICL) (Brown et al., 2020; Dong et al., 2023) offers question-answering examples to guide generation, but the method is more helpful in regulating the answer format than augmenting informational content if the examples are not finely sifted. Chain-of-thought (CoT) (Wei et al., 2022) directs LLMs to generate the output step by step (Kojima et al., 2022), yet the rationale

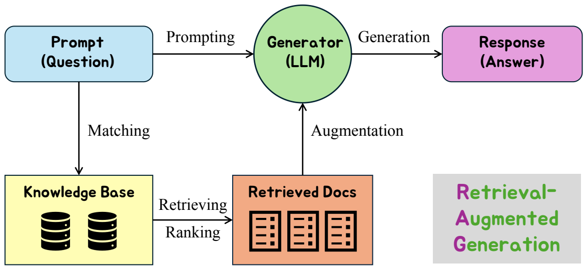

Figure 1: The overview of retrieval-augmented generation (RAG) for large language models (LLMs).

produced in intermediate steps may not consistently contribute to problem-solving and may instead involve hallucination. Retrieval-augmented generation (RAG) (Lewis et al., 2020) directly searches relevant information from external knowledge bases like Wikipedia [1] and then augments the initial prompt for LLMs to improve the generated answers, as demonstrated in Figure 1. Though intuitive and promising, the RAG approach requires extensive exploration because adding extra context to the prompt does not always enhance the performance of LLMs.

In this work, we plan to examine different RAG methods for LLMs on natural language generation tasks focusing on commonsense knowledge. Our project aims to develop an innovative and foundational approach for integrating external knowledge into LLMs. Specifically, we will implement the RAG system in the following steps: (1) **Building KB**: to build the knowledge base from encyclopedic and commonsense knowledge sources; (2) **Indexing**: to develop the document indexing module for handling information in the KB; (3) **Retrieval**: to develop the information extraction module for searching, matching, and ranking the most relevant documents; (4) **Augmentation**: to combine the extracted knowledge with the initial context, question, and options to form the final prompt; (5)

♠ Authors contributed equally and listed alphabetically.

[1] https://pypi.org/project/Wikipedia-API/

Generation: to feed LLMs with the final prompt to generate answers; (6) **Advanced RAG**: to incorporate other advanced RAG methods as auxiliary modules. Each of these components can be implemented with different designs. We will elaborate on it in § 3.

To evaluate the results, we will test the performance of different LLMs on a wide variety of commonsense natural language understanding and generation tasks. The evaluation metrics will be accuracy, f1, exact match, and more. Compared to the vanilla LLM generation, we expect that using the RAG approach will enhance the performance. In addition, we will conduct comprehensive analyses and ablation studies on the effectiveness of different RAG components. Our expected contributions are as follows:

- We will systematically implement the RAG system emphasizing commonsense reasoning. The code, data, and results will be released on GitHub [2].
- We will extensively examine the effectiveness of different RAG methods for multiple LLMs on various commonsense tasks. The results will bring findings and ideas for further improvement.
- We will conduct in-depth analyses and ablation studies to assess the efficacy of different RAG components. The analysis is supposed to shed light on future commonsense RAG research.

2 Related Work

2.1 Retrieval-Augmented Generation

There has been a series of retrieval-augmented generation (RAG) research (Gao et al., 2023; Mialon et al., 2023) proposed in recent years. Karpukhin et al. (2020) propose to use dense representations from a dual-encoder framework to replace traditional sparse vector methods, such as TF-IDF or BM25 (Robertson et al., 2009), for open-domain QA tasks. Ma et al. (2023) propose to ask the model to rewrite the original query and then perform web searching to obtain the relevant documents. He et al. (2024) apply RAG for textual graph understanding and question answering using LLMs and graph neural networks (GNNs).

[2] https://github.com/YuweiYin/UBC_CPSC_532V

Multi-modal RAG. Zhao et al. (2023a) examine research involving the augmentation of generative models through the retrieval of multi-modal information, including image, code, structured knowledge, speech, and video. Hu et al. (2023) propose to augment a visual-language model by enabling it to retrieve multiple knowledge entries from diverse sources, thus aiding generation.

Multi-source RAG. Yu (2022) highlights the limitations associated with relying solely on single-source homogeneous knowledge, such as Wikipedia, and offers various solutions for implementing RAG using heterogeneous knowledge sources. Wang et al. (2024) propose a unified multi-source RAG method including three sub-tasks, i.e., knowledge source selection, knowledge retrieval, and response generation, with a self-refinement mechanism for iteratively refining the generated response.

RAG + GAR. Shao et al. (2023) introduce a framework that combines Retrieval-Augmented Generation and Generation-Augmented Retrieval (GAR) (Mao et al., 2021), which utilizes the model output from the previous iteration as the context to enhance the RAG process. Similarly, Feng et al. (2023) propose to iteratively use language models to refine the documents retrieved in the RAG step.

Robust RAG. RAG can harm performance when irrelevant retrieval is used. Recent research proposes methods to improve the robustness of generation (Yoran et al., 2023; Yan et al., 2024). (Berchansky et al., 2023) propose to eliminate non-essential retrieved information at the token level to streamline the answer generation process. In addition, RAG exhibits certain limitations, such as the attribution-fluency trade-off (Aksitov et al., 2023), wherein the quality of output may be influenced by the constraints introduced by the retrieved knowledge.

LLMs as Knowledge Source. While the majority of RAG methods retrieve information from external knowledge bases, recent research suggests utilizing LLMs to generate documents or processing the retrieved ones. Petroni et al. (2019) systematically analyze the factual and commonsense knowledge present in publicly available pretrained language models.

Source	URL / API
CYC (Lenat et al., 1986)	cyc.com
WebChild (Tandon et al., 2014)	mpi-inf.mpg.de
ConceptNet (Speer et al., 2017)	conceptnet.io
NELL (Mitchell et al., 2018)	huggingface.co
Atomic (Hwang et al., 2021)	allenai.org
NCLB (Fung et al., 2024)	github.com
Wikipedia	huggingface.co

Table 1: The knowledge source for building the RAG knowledge base.

2.2 Commonsense RAG

Commonsense knowledge constitutes a fundamental aspect of artificial intelligence (Gunning, 2018; Razniewski et al., 2021) and commonsense reasoning is a significant task in natural language processing (NLP) (Sap et al., 2020). Bosselut et al. (2019) propose the COMET model combining the power of the Transformer model (Vaswani et al., 2017) and commonsense knowledge graphs Atomic (Hwang et al., 2021) and ConceptNet (Speer et al., 2017). Lal et al. (2022) propose to use COMET as the commonsense knowledge source to augment different LLMs for answering why-questions.

Liu et al. (2022) propose to generate knowledge from a language model, and then perform RAG to answer questions. Li et al. (2021) propose a BERT-based filter model to filter low-quality candidates and implement contrastive learning (Chen et al., 2020) in both the encoder and decoder. Yu et al. (2022) propose a unified framework of retrieval-augmented commonsense reasoning, a commonsense corpus with over 20 million documents, and strategies for training a commonsense retriever. Ghosal et al. (2023) propose to train a sequence-to-sequence next-step prediction model by incorporating external commonsense knowledge and employing search techniques to generate intermediate steps for natural language inference (NLI) tasks. Seo et al. (2022) propose to retrieve scene knowledge to enhance compositional generalization and relational knowledge to improve commonsense reasoning. Cui et al. (2024) propose a multi-modal retrieval augmentation framework leveraging both text and images to enhance the commonsense capabilities of language models.

3 Implementation Steps

In this section, we introduce the implementation steps of the project. Our project aims to develop an RAG system to examine and compare different RAG methods of integrating external knowledge into LLMs for solving commonsense NLP tasks. We will elaborate on the key points and hardships in the implementation.

3.1 Step 1: Building Knowledge Base

First, we plan to use the encyclopedic and commonsense knowledge sources, as shown in Table 1. At first, we will collect and use them as separate sources. Then we will consider classifying and merging the common knowledge from them to build an overarching knowledge base (KB) with a unified knowledge structure. The knowledge will be classified into several types, including *cultural*, *social*, *temporal*, *physical*, and *multimodal* commonsense. Additionally, we list commonsense-related datasets that may serve as knowledge sources in Table 5.

3.2 Step 2: Indexing

Then, we will develop the indexing module for searching for information in the commonsense KB. Since the element in the KB can be structured knowledge or raw text, we need multiple processing methods to handle these different types. The processed knowledge nodes should be in the same format, of reasonable size, and easy to perform searching and semantic matching.

3.3 Step 3: Retrieval

The retriever module is responsible for searching, semantic matching, and result ranking. The related documents are retrieved from the KB and sorted based on their relevance to the query (question). Afterward, we will try to add several pre-retrieval (e.g., query routing (Li et al., 2023), rewriting (Ma et al., 2023), and expansion) and post-retrieval (e.g., document re-rank (Zhuang and Zuccon, 2021; Zhuang et al., 2023), summary, and fusion (Raudaschl, 2023)) methods to enhance the results.

3.4 Step 4: Augmentation

After retrieval, we will combine the extracted documents with the original context, question, and options to construct the final prompt as the input of LLMs. The implementation of augmentation can be flexible. We will try zero-shot generation, in-context learning (Brown et al., 2020; Dong et al., 2023), chain-of-thought prompting (Wei et al., 2022; Kojima et al., 2022), and supervised fine-tuning (SFT) with instruction tuning (Wei et al., 2021; Sanh et al., 2021; Longpre et al., 2023; Zhang et al., 2023; Jiang et al., 2024).

3.5 Step 5: Generation

We will implement the generation process of LLM using PyTorch (Paszke et al., 2019) and Hugging Face `transformers` (Wolf et al., 2020) toolkits. For evaluation, we will implement the language model evaluation scripts on selected commonsense benchmarks. In addition, we will implement the fine-tuning of LLM because we plan to explore instruction tuning in the augmentation stage.

3.6 Step 6: Advanced RAG

After implementing the above steps, a basic RAG system is built. We will incorporate various advanced RAG methods (as mentioned in § 2) as auxiliary modules, compare their performance on commonsense benchmarks, analyze the results, and discuss the findings and ideas for future research.

4 Experimental Setup

In this Section, we introduce the plan of experimental settings, which could be limited by the project scope, duration, and available computing resources to some extent.

4.1 Tasks, Datasets, and Evaluation

We will use a series of commonsense NLP tasks and datasets for evaluation, including but not limited to the benchmarks in Table 4. All of them are classification tasks, where the training and validation set have labels, while the test set does not. For the evaluation metrics, we will use **accuracy** for balanced classification, **f1 score** for unbalanced classification, Matthews correlation coefficient (Mcc) (Matthews, 1975; Chicco and Jurman, 2020), and exact match (EM) score (for multiple correct answers scenario) for multiple correct answers scenario.

4.2 Baselines

Baseline LLM Models We plan to use LLMs of different sizes (100M to 1B parameters) and series as the backbone models, such as GPT-2 (Radford et al., 2019), GPT-NEO (Black et al., 2022), OPT (Zhang et al., 2022), and OLMo (Groeneveld et al., 2024; Soldaini et al., 2024).

Baseline Augmentation Methods We plan to test different augmentation methods to incorporate knowledge, such as zero-shot generation (Vaswani et al., 2017; Radford et al., 2018), in-context learning (ICL) (Brown et al., 2020; Dong et al., 2023), chain-of-thought prompting (CoT) (Wei et al., 2022; Kojima et al., 2022), and Supervised fine-tuning (SFT) with Instruction Tuning (Wei et al., 2021; Sanh et al., 2021; Longpre et al., 2023; Zhang et al., 2023; Jiang et al., 2024).

4.3 Expected Results

We expect that the task-solving performance on commonsense NLP benchmarks of baselines will be improved by incorporating RAG methods. In addition, we will implement and experiment with as many RAG methods as possible within the time frame. The discussions of result analysis, ablation study, and error/case study will bring research insights for future work, such as knowledge-aware / explainable generation and effectively mitigating hallucination.

5 Task Division and Timeline

5.1 Task Division

We will work together on research idea discussion and method implementation. The main focus of each member is shown in Table 2.

Task	Juntai	Yilin	Yuwei
Project leading			✔
Run Experiments	✔		✔
Results analysis	✔	✔	
Paper writing		✔	✔
Step 1 Building KB	✔	✔	
Step 2 Indexing	✔	✔	
Step 3 Retrieval	✔	✔	✔
Step 4 Augmentation			✔
Step 5 Gen/Eval/Train			✔
Step 6 Advanced RAG	✔	✔	✔

Table 2: Task division of group members.

6 Timeline

Despite possible schedule changes, we will try to follow the project timeline as shown in Table 3.

Week	Juntai	Yilin	Yuwei
Before Reading Week	Project Planning & Literature Review		
02.19–02.25 02.26–03.03	Method Research	Dataset Research	Step 5 Generation/Evaluation/Training
03.04–03.10	Project Proposal Writing & Presenting		
03.11–03.17 03.18–03.24	Step 1 Building KB Step 1 Building KB & Step 2 Indexing		Step 4 Augmentation
03.25–03.31 04.01–04.07 04.08–04.12	Step 2 Indexing & Step 3 Retrieval Results Analysis & & Step 6 Advanced RAG Results Analysis & Paper Writing		Pipeline & Preliminary Experiments Experiments: Step 1+2+3+4+5 [+6] Paper Writing
Apr 10, Wed Apr 12, Fri	Project Presentation Project Paper Submission		

Table 3: Timeline for the project.

Dataset	Training	Validation	Test	# Class	Metric
GLUE (Wang et al., 2019b)					
`cola` (The Corpus of Linguistic Acceptability)	8.55k	1.04k	1.06k	2	Mcc
`mnli` (MultiNLI Matched)	393k	9.82k	9.8k	3	Acc
`mnli` (MultiNLI Mismatched)	393k	9.83k	9.85k	3	Acc
`mrpc` (Microsoft Research Paraphrase Corpus)	3.67k	408	1.73k	2	Acc, F1
`qnli` (Question NLI)	105k	5.46k	5.46k	2	Acc
`qqp` (Quora Question Pairs)	364k	40.4k	391k	2	Acc, F1
`rte` (Recognizing Textual Entailment)	2.49k	277	3k	2	Acc
`sst2` (The Stanford Sentiment Treebank)	67.3k	872	1.82k	2	Acc
`wnli` (Winograd NLI)	635	71	146	2	Acc
SuperGLUE (Wang et al., 2019a)					
`boolq` (BoolQ)	9.43k	3.27k	3.25k	2	Acc
`cb` (CommitmentBank)	250	56	250	3	Acc, F1
`copa` (Choice of Plausible Alternatives)	500	100	400	2	Acc
`multirc` (Multi-Sentence Reading Comprehension)	27.2k	4.85k	9.69k	2	Acc
`record` (Reading Comprehension with Commonsense Reasoning)	101k	10k	10k	N	F1, EM
`rte` (Recognizing Textual Entailment)	2.49k	277	3k	2	Acc
`wic` (Words in Context)	5.43k	638	1.4k	2	Acc
`wsc` (The Winograd Schema Challenge)	554	104	146	2	Acc
WSC273 (Levesque et al., 2012)	-	-	273	2	Acc
WinoGrande (Sakaguchi et al., 2021)	9.25k	1.27k	1.77k	2	Acc
ANLI r1 (Nie et al., 2020)	16.9k	1k	1k	3	Acc
ANLI r2	45.5k	1k	1k	3	Acc
ANLI r3	100k	1.2k	1.2k	3	Acc
ARC Easy (Clark et al., 2018)	2.25k	570	2.38k	4	Acc
ARC Challenge	1.12k	299	1.17k	4	Acc
PIQA (Bisk et al., 2020)	16.1k	1.84k	3.08	2	Acc
SWAG (Zellers et al., 2018)	73.5k	20k	20k	4	Acc
HellaSwag (Zellers et al., 2019b)	39.9k	10k	10k	4	Acc
Commonsense QA (Talmor et al., 2019)	9.74k	1.22k	1.14k	5	Acc

Table 4: The statistics of datasets for evaluation. All of them are classification tasks, where the training and validation set have labels, while the test set does not. For the evaluation metrics, “Acc”, “F1”, “EM”, and “MCC” mean accuracy, f1 score, exact match score (for multiple correct answers scenario), and Matthews correlation coefficient (Matthews, 1975; Chicco and Jurman, 2020), respectively.

Dataset	Knowledge Type	Website	API
GLUE (Wang et al., 2019b)	NLU & Commonsense	gluebenchmark.com	huggingface.co
SuperGLUE (Wang et al., 2019a)	NLU & Commonsense	super.gluebenchmark.com	huggingface.co
SNLI (Bowman et al., 2015)	NLI	nlp.stanford.edu	huggingface.co
Adversarial NLI (Nie et al., 2020)	NLI	github.com	huggingface.co
OpenBookQA (Mihaylov et al., 2018)	Subject & Commonsense	allenai.org	huggingface.co
ARC (Clark et al., 2018)	Science QA	allenai.org	huggingface.co
CommonGen (Lin et al., 2020)	Daily-life Commonsense	inklab.usc.edu	huggingface.co
Cosmos QA (Huang et al., 2019)	Commonsense Reading Comprehension	github.io	huggingface.co
MultiRC (Khashabi et al., 2018) (in SuperGLUE)	Commonsense Reading Comprehension	cogcomp.seas.upenn.edu	huggingface.co
ReCORD (Zhang et al., 2018) (in SuperGLUE)	Commonsense Reading Comprehension	github.io	huggingface.co
Social IQA (Sap et al., 2019)	Social Commonsense	allenai.org	huggingface.co
COPA (Roemmele et al., 2011) (in GLUE)	Social Commonsense	ict.usc.edu	huggingface.co
WSC (Levesque et al., 2012) (in GLUE)	Social Commonsense	cs.nyu.edu	huggingface.co
RocStories (Mostafazadeh et al., 2016)	Social Commonsense	cs.rochester.edu	huggingface.co
SODA (Kim et al., 2023)	Social Commonsense	github.com	huggingface.co
PIQA (Bisk et al., 2020)	Physical Commonsense	allenai.org	huggingface.co
SWAG (Zellers et al., 2018)	Physical Commonsense	rowanzellers.com	huggingface.co
WinoGrande (Sakaguchi et al., 2021)	Social+Physical Commonsense	allenai.org	huggingface.co
Commonsense QA (Talmor et al., 2019)	Social+Physical Commonsense	tau-nlp.sites.tau.ac.il	huggingface.co
Abductive NLI (Bhagavatula et al., 2020)	Social+Physical Commonsense	allenai.org	-
HellaSwag (Zellers et al., 2019b)	Physical+Temporal Commonsense	rowanzellers.com	huggingface.co
MCTaco (Zhou et al., 2019)	Temporal Commonsense	allenai.org	huggingface.co
TimeDial (Qin et al., 2021)	Temporal Commonsense	github.com	huggingface.co
VQA (Antol et al., 2015)	Multimodal Commonsense	visualqa.org	visualqa.org
VCR (Zellers et al., 2019a)	Multimodal Commonsense	visualcommonsense.com	github.com
NLVR (Suhr et al., 2017, 2019)	Multimodal Commonsense	nlp.cornell.edu	github.com
WebQA (Chang et al., 2022)	Multimodal Commonsense	github.io	github.com
GSR (Pratt et al., 2020)	Multimodal Commonsense	allenai.org	github.com
VGSI (Yang et al., 2021)	Multimodal Commonsense	github.com	-
ALFRED (Shridhar et al., 2020)	Multimodal Commonsense	askforalfred.com	github.com
MaRVL (Liu et al., 2021)	Cultural Commonsense	github.io	github.com
GD-VCR (Yin et al., 2021)	Cultural Commonsense	github.com	-
GIVL (Yin et al., 2023)	Cultural Commonsense	github.com	-

Table 5: Potential knowledge sources for the project.

References

Renat Aksitov, Chung-Ching Chang, David Reitter, Siamak Shakeri, and Yunhsuan Sung. 2023. Characterizing attribution and fluency tradeoffs for retrieval-augmented large language models. *arXiv preprint arXiv:2302.05578*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, abs/2212.08073.

Moshe Berchansky, Peter Izsak, Avi Caciularu, Ido Dagan, and Moshe Wasserblat. 2023. Optimizing retrieval-augmented reader models via token elimination. *arXiv preprint arXiv:2310.13682*.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, abs/2204.06745.

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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, volume 33, pages 1877–1901, Virtual Event. NeurIPS.

Yingshan Chang, Guihong Cao, Mridu Narang, Jianfeng Gao, Hisami Suzuki, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal QA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16474–16483. IEEE.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21:1–13.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. More: Multi-modal retrieval augmented generative commonsense reasoning. *arXiv preprint arXiv:2402.13625*, abs/2402.13625.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, abs/2301.00234.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2023. Retrieval-generation synergy augmented large language models. *arXiv preprint arXiv:2310.05149*, abs/2310.05149.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. No culture left behind: Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, abs/2312.10997.

Deepanway Ghosal, Somak Aditya, and Monojit Choudhury. 2023. Prover: Generating intermediate steps for NLI with commonsense knowledge retrieval and next-step prediction. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 872–884, Nusa Dua, Bali. Association for Computational Linguistics.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, abs/2402.00838.

David Gunning. 2018. Machine common sense concept paper. *arXiv preprint arXiv:1810.07528*.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *arXiv preprint arXiv:2402.07630*, abs/2402.07630.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23369–23379.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual 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 2391–2401, Hong Kong, China. 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. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Victoria Lin, Wen tau Yih, and Srinivasan Iyer. 2024. Instruction-tuned language models are better knowledge learners. *arXiv preprint arXiv:2402.12847*.

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.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yash Kumar Lal, Niket Tandon, Tanvi Aggarwal, Horace Liu, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2022. Using commonsense knowledge to answer why-questions. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Douglas B. Lenat, Mayank Prakash, and Mary Shepherd. 1986. CYC: using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Mag.*, 6(4):65–85.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10-14, 2012*. AAAI Press.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Haonan Li, Yeyun Gong, Jian Jiao, Ruofei Zhang, Timothy Baldwin, and Nan Duan. 2021. Kfcnet: Knowledge filtering and contrastive learning for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2918–2928, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoqian Li, Ercong Nie, and Sheng Liang. 2023. From classification to generation: Insights into crosslingual retrieval augmented icl. *arXiv preprint arXiv:2311.06595*.

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, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315, Singapore. Association for Computational Linguistics.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. 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 4089–4100, Online. Association for Computational Linguistics.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Tom M. Mitchell, William W. Cohen, Estevam R. Hruschka Jr., Partha P. Talukdar, Bo Yang, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matt Gardner, Bryan Kisiel, Jayant Krishnamurthy, Ni Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil A. Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard C. Wang, Derry Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling. 2018. Never-ending learning. *Commun. ACM*, 61(5):103–115.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James F. Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*, abs/1604.01696.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

OpenAI. 2022. Chatgpt. *OpenAI Research*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Sarah M. Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. 2020. Grounded situation recognition. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, volume 12349 of *Lecture Notes in Computer Science*, pages 314–332. Springer.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021. TIMEDIAL: Temporal commonsense reasoning in dialog. 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 7066–7076, Online. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. *OpenAI blog*.

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

Adrian H. Raudaschl. 2023. Forget rag, the future is rag-fusion. Towards Data Science.

Simon Razniewski, Niket Tandon, and Aparna S. Varde. 2021. Information to wisdom: Commonsense knowledge extraction and compilation. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, pages 1143–1146. ACM.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI.

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.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

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.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

Jaehyung Seo, Dongsuk Oh, Sugyeong Eo, Chanjun Park, Kisu Yang, Hyeonseok Moon, Kinam Park, and Heuiseok Lim. 2022. PU-GEN: enhancing generative commonsense reasoning for language models with human-centered knowledge. *Knowledge-Based Systems*, 256:109861.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274, Singapore. Association for Computational Linguistics.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10737–10746. Computer Vision Foundation / IEEE.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, abs/2402.00159.

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, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Niket Tandon, Gerard de Melo, Fabian M. Suchanek, and Gerhard Weikum. 2014. Webchild: harvesting and organizing commonsense knowledge from the web. In *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*, pages 523–532. ACM.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan, and Kam-Fai Wong. 2024. Unims-rag: A unified multi-source retrieval-augmented generation for personalized dialogue systems. *arXiv preprint arXiv:2401.13256*, abs/2401.13256.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, abs/2401.15884.

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. Visual goal-step inference using wikiHow. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2167–2179, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Da Yin, Feng Gao, Govind Thattai, Michael Johnston, and Kai-Wei Chang. 2023. GIVL: improving geographical inclusivity of vision-language models with pre-training methods. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 10951–10961. IEEE.

Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2115–2129, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*, abs/2310.01558.

Wenhao Yu. 2022. Retrieval-augmented generation across heterogeneous knowledge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 52–58, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4364–4377, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019a. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*, abs/1810.12885.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, abs/2308.10792.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, abs/2205.01068.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023a. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4736–4756, Singapore. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023b. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. 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 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Shengyao Zhuang, Bing Liu, Bevan Koopman, and Guido Zuccon. 2023. Open-source large language models are strong zero-shot query likelihood models for document ranking. *arXiv preprint arXiv:2310.13243*.

Shengyao Zhuang and Guido Zuccon. 2021. Tilde: Term independent likelihood model for passage reranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1483–1492.