Commonsense Retrieval-Augmented Generation for Large Language Models

CPSC 532V 2023W2 Project Proposal - 2024.03.06

Juntai Cao, Yilin Yang, *and* Yuwei Yin

{jtcao7, yangyl17, yuweiyin}@cs.ubc.ca
Student #: 50171404, 24754350, 36211928
.
Department of Computer Science, Faculty of Science
University of British Columbia, on unceded Musqueam land
Vancouver, BC V6T 1Z4, Canada

THE UNIVERSITY OF BRITISH COLUMBIA

1. Introduction
2. Implementation Steps
3. Experimental Setup
4. Conclusion
5. References

Background

Natural language generation has improved considerably owing to the rapid development of large language models (LLMs) [1, 2, 3].

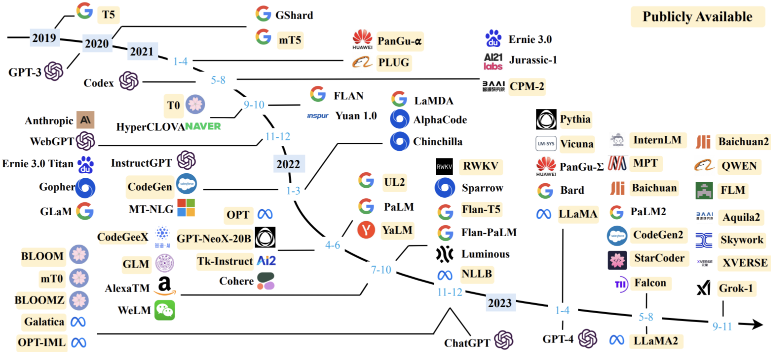

Figure 1: A timeline of existing LLMs (size > 10B) in recent years [3].

Background

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

Figure 2: Claude 3 – a new standard for intelligence [4] (Mar 4, 2024).

Motivation

- **Hallucinations** in LLMs [5]
  producing outputs that are coherent and grammatically correct but *nonsensical*, *unfaithful* [6], or *factually incorrect*.

Motivation

- **Hallucinations** in LLMs [5]
  producing outputs that are coherent and grammatically correct but *nonsensical*, *unfaithful* [6], or *factually incorrect*.
- Especially when confronted with tasks that demand **commonsense reasoning** [7].
  (usually more implicit & need more context)

Motivation

- **Hallucinations** in LLMs [5]
  producing outputs that are coherent and grammatically correct but *nonsensical*, *unfaithful* [6], or *factually incorrect*.
- Especially when confronted with tasks that demand **commonsense reasoning** [7].
  (usually more implicit & need more context)
- LM training paradigm [8, 9]: LLM models **predict the next token** based on the previously generated ones.
  "The president of US is" $\rightarrow$ "Biden" or "Trump" (Which year?)

Motivation

- **Hallucinations** in LLMs [5]
  producing outputs that are coherent and grammatically correct but *nonsensical*, *unfaithful* [6], or *factually incorrect*.
- Especially when confronted with tasks that demand **commonsense reasoning** [7].
  (usually more implicit & need more context)
- LM training paradigm [8, 9]: LLM models **predict the next token** based on the previously generated ones.
  "The president of US is" → "Biden" or "Trump" (Which year?)
- Hence, the models are supposed to produce better output if they are **conditioned on more relevant context**.
  "In 2020, the president of US is" → "Trump" (with context)

Research Problem

- ***How do we mitigate hallucination in LLM generation?***

Research Problem

- ***How do we mitigate hallucination in LLM generation?***
- **Modeling and Inference Methods**:
  > architecture (encoder, attention, decoder),
  > training (planning, reinforcement learning, multi-task learning, controllable generation),
  > post-processing, etc.

Research Problem

- ***How do we mitigate hallucination in LLM generation?***
- **Modeling and Inference Methods**:
  > architecture (encoder, attention, decoder),
  > training (planning, reinforcement learning, multi-task learning, controllable generation),
  > post-processing, etc.
- **Data-Related Methods**:
  > building a faithful dataset,
  > cleaning data automatically,
  > **information augmentation**, etc.

Research Problem

The focus of this project:

- ***How do we effectively incorporate contextual information to enhance problem-solving performance and mitigate hallucination in LLM generation?***

Retrieval-Augmented Generation

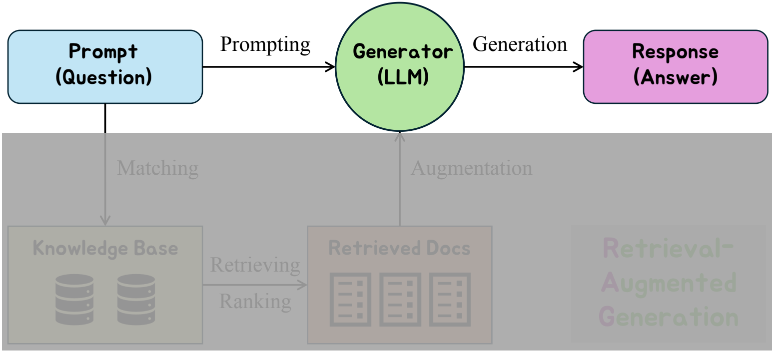

Figure 3: The overview of LLM generation.

Retrieval-Augmented Generation

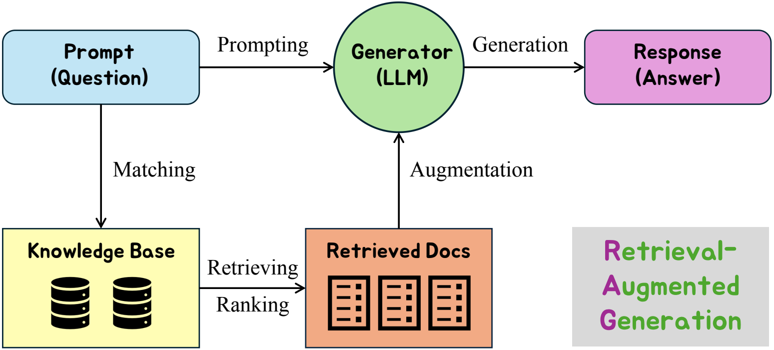

Figure 4: The overview of retrieval-augmented generation (RAG).

Retrieval-Augmented Generation

Key points and hardships of RAG:

- 1. Building knowledge base from multiple sources

Retrieval-Augmented Generation

Key points and hardships of RAG:

- 1. Building knowledge base from multiple sources
- 2. Knowledge indexing for fast searching

Retrieval-Augmented Generation

Key points and hardships of RAG:

- 1. Building knowledge base from multiple sources
- 2. Knowledge indexing for fast searching
- 3. Semantic matching and knowledge retrieval

Retrieval-Augmented Generation

Key points and hardships of RAG:

- 1. Building knowledge base from multiple sources
- 2. Knowledge indexing for fast searching
- 3. Semantic matching and knowledge retrieval
- 4. Augmentation of the original query

Retrieval-Augmented Generation

Key points and hardships of RAG:

- 1. Building knowledge base from multiple sources
- 2. Knowledge indexing for fast searching
- 3. Semantic matching and knowledge retrieval
- 4. Augmentation of the original query
- 5. LLM generation and commonsense evaluation

Retrieval-Augmented Generation

Key points and hardships of RAG:

- 1. Building knowledge base from multiple sources
- 2. Knowledge indexing for fast searching
- 3. Semantic matching and knowledge retrieval
- 4. Augmentation of the original query
- 5. LLM generation and commonsense evaluation
- 6. Incorporating various advanced RAG methods

Retrieval-Augmented Generation

Key points and hardships of RAG:

- 1. Building knowledge base from multiple sources
- 2. Knowledge indexing for fast searching
- 3. Semantic matching and knowledge retrieval
- 4. Augmentation of the original query
- 5. LLM generation and commonsense evaluation
- 6. Incorporating various advanced RAG methods

★ All points will be covered in our implementation.
Skip related work (in our proposal) for the sake of time

Implementation

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.

Steps

1. Building knowledge base
2. Indexing
3. Retrieval
4. Augmentation
5. Generation
6. Advance RAG

Step 1 & 2: Building Knowledge Base and Indexing

- Encyclopedic and commonsense knowledge sources
  1. Atomic [10]
  2. ConceptNet [11]
  3. Wikipedia [1]
  4. NCLB [12]
  5. ...
- From individual KBs with different formats
  $\rightarrow$ a unified, structured KB.

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

Step 3: Retrieval

The **retriever** module is responsible for *searching*, *semantic matching*, and *result ranking*.

Pre-retrieval

- Rewriting [13]
- Query routing [14]

Post-retrieval

- Re-ranking [15, 16]

Step 4: Augmentation

- Zero-shot generation [8, 9]
- In-context Learning (ICL) [17]
- Chain-of-Thought prompting (CoT) [18, 19]
- Supervised fine-tuning (SFT) with Instruction Tuning [20, 21, 22, 23, 24]

Step 5 & 6: Generation/Evaluation/Training and advanced RAG

- We will implement the language model evaluation scripts on selected commonsense benchmarks.
- If time allows, we will try to incorporate several advanced RAG methods like Robust RAG [25, 26], GAR (Generation Augmented Retrieval) [27, 28, 29], etc.

Tasks, Datasets, and Evaluation

- **Tasks**: Multi-choice Commonsense NLP Tasks (classification)
- **Datasets**: GLUE, SuperGLUE, WSC, WinoGrande, ANLI, ARC, PIQA, SWAG, HellaSwag, etc.
- **Evaluation**: Acc, F1, Mcc, EM, etc.
  Acc (Accuracy): for balanced classification
  F1: for unbalanced classification
  Mcc: Matthews correlation coefficient [30][2]
  EM (Exact Match): for multiple correct answers scenario

[2] The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation [31]

Tasks, Datasets, and Evaluation - GLUE

Dataset	Training	Validation	Test	# Class	Metric
GLUE [32]					
`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

Table 1: 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”, “Mcc”, and “EM” mean accuracy, f1 score, Matthews correlation coefficient, and exact match score, respectively.

Tasks, Datasets, and Evaluation - SuperGLUE

Dataset	Training	Validation	Test	# Class	Metric
SuperGLUE [33]					
`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 w/ 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

Table 2: 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", "Mcc", and "EM" mean accuracy, f1 score, Matthews correlation coefficient, and exact match score, respectively.

Tasks, Datasets, and Evaluation - Other Commonsense Benchmark

Dataset	Training	Validation	Test	# Class	Metric
WSC273 [34]	-	-	273	2	Acc
WinoGrande [35]	9.25k	1.27k	1.77k	2	Acc
ANLI `r1` [36]	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` [37]	2.25k	570	2.38k	4	Acc
ARC `Challenge`	1.12k	299	1.17k	4	Acc
PIQA [38]	16.1k	1.84k	3.08	2	Acc
SWAG [39]	73.5k	20k	20k	4	Acc
HellaSwag [40]	39.9k	10k	10k	4	Acc
Commonsense QA [41]	9.74k	1.22k	1.14k	5	Acc

Table 3: 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", "Mcc", and "EM" mean accuracy, f1 score, Matthews correlation coefficient, and exact match score, respectively.

Baselines and Experiments

- **Baseline models**: LLMs of different sizes and series

Baselines and Experiments

- **Baseline models**: LLMs of different sizes and series
- **Baseline methods**:
  > Zero-shot generation;
  > In-context learning (ICL) [17];
  > Chain-of-Thought prompting (CoT) [18, 19];
  > Supervised fine-tuning (SFT) with Instruction Tuning [20, 21, 22, 23, 24]

Baselines and Experiments

- **Baseline models**: LLMs of different sizes and series
- **Baseline methods**:
  > Zero-shot generation;
  > In-context learning (ICL) [17];
  > Chain-of-Thought prompting (CoT) [18, 19];
  > Supervised fine-tuning (SFT) with Instruction Tuning [20, 21, 22, 23, 24]
- **Experiments**: Baseline w/ RAG v.s. Baseline w/o RAG
  Expected Results: Baseline w/ RAG > Baseline w/o RAG

Baselines and Experiments

- **Baseline models**: LLMs of different sizes and series
- **Baseline methods**:
  > Zero-shot generation;
  > In-context learning (ICL) [17];
  > Chain-of-Thought prompting (CoT) [18, 19];
  > Supervised fine-tuning (SFT) with Instruction Tuning [20, 21, 22, 23, 24]
- **Experiments**: Baseline w/ RAG v.s. Baseline w/o RAG
  Expected Results: Baseline w/ RAG > Baseline w/o RAG
- **Analysis**:
  > Task performance (why better or worse);
  > Ablation study (effectiveness of each component);
  > Error analysis & Case study & Discussion → Research insights

Expected Contributions

The focus of this project: ***How do we effectively incorporate contextual information to enhance problem-solving performance and mitigate hallucination in LLM generation?***

- We will systematically implement the RAG system emphasizing commonsense reasoning. The code, data, and results will be released on GitHub [3].

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

Expected Contributions

The focus of this project: ***How do we effectively incorporate contextual information to enhance problem-solving performance and mitigate hallucination in LLM generation?***

- We will systematically implement the RAG system emphasizing commonsense reasoning. The code, data, and results will be released on GitHub [3].
- We will extensively examine the effectiveness of different RAG methods for multiple LLMs on various commonsense tasks. The results will bring findings & ideas for further improvement.

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

Expected Contributions

The focus of this project: ***How do we effectively incorporate contextual information to enhance problem-solving performance and mitigate hallucination in LLM generation?***

- We will systematically implement the RAG system emphasizing commonsense reasoning. The code, data, and results will be released on GitHub [3].
- We will extensively examine the effectiveness of different RAG methods for multiple LLMs on various commonsense tasks. The results will bring findings & ideas for further improvement.
- We will conduct in-depth analyses and ablation studies to assess the efficacy of different RAG components. The analysis could shed light on future commonsense RAG research.

[3] `https://github.com/YuweiYin/UBC_CPSC_532V`

Task Division

We will work together on research idea discussion, result analysis, and paper writing. The main focus of each member is as follows:

- **Juntai Cao** & **Yilin Yang**:
  Step 1 Building knowledge base;
  Step 2 Indexing;
  Step 3 Retrieval;
  Step 6 Advanced RAG;
  Results analysis
- **Yuwei Yin**:
  Project leading;
  Step 3 Retrieval;
  Step 4 Augmentation;
  Step 5 Generation/Evaluation/Training;
  Step 6 Advanced RAG;
  Paper writing

Timeline

- Before Reading Week: Project planning & Literature review
- 02.19–02.25 (Reading Week) + 02.26–03.03:
  > Done: (Yuwei) Step 5 Generation/Evaluation/Training [4]
- 03.04–03.10:
  > Done: Proposal writing & presenting
  > Doing: collecting knowledge sources
- 03.11–03.17 Todo:
  > (Juntai & Yilin) Step 1 Building knowledge base
  > (Yuwei) Step 4 Augmentation (unify knowledge structure)

[4] The code and preliminary results: `https://github.com/YuweiYin/UBC_CPSC_532V/tree/master/Project#experimental-results`

Timeline

- 03.18–03.24 Todo:
  > (Juntai & Yilin) Step 1 Building KB & Step 2 Indexing
  > (Yuwei) Step 4 Augmentation (Zero-shot, ICL, CoT, SFT)
- 03.25–03.31 Todo:
  > (Juntai & Yilin) Step 2 Indexing & Step 3 Retrieval
  > (Yuwei) Experiment Pipeline: Step 4 + Step 5
- 04.01–04.07 Todo:
  > (Juntai & Yilin) Results analysis & Step 6 Advanced RAG
  > (Yuwei) Experiments (Step 1+2+3+4+5[+6])
- 04.08–04.12 Todo:
  > (Juntai & Yilin) Results analysis & Paper writing
  > (Yuwei) Experiments & Paper writing
  Project Presentation (due Apr 10) & Paper (ddl Apr 12)

Q&A

[1] 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.
Llama 2: Open foundation and fine-tuned chat models.
arXiv preprint arXiv:2307.09288, abs/2307.09288, 2023.

[2] OpenAI.
Gpt-4 technical report.
arXiv preprint arXiv:2303.08774, 2023.

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

[4] Anthropic.
The claude 3 model family: Opus, sonnet, haiku, 4 Mar 2024.

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

[6] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald.
On faithfulness and factuality in abstractive summarization.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics.

[7] Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth.
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, July 2020. Association for Computational Linguistics.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin.
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, 2017.

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

[10] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi.
(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, 2021.

[11] Robyn Speer, Joshua Chin, and Catherine Havasi.
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, 2017.

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

[13] Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan.
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, December 2023. Association for Computational Linguistics.

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

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

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

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

[18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou.
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*, 2022.

[19] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa.
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, 2022.

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

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

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

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

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

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

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

[27] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen.
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, August 2021. Association for Computational Linguistics.

[28] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen.
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, December 2023. Association for Computational Linguistics.

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

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

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

[32] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman.
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, 2019.

[33] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman.
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, 2019.

[34] Hector J. Levesque, Ernest Davis, and Leora Morgenstern.
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, 2012.

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

[36] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela.
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, July 2020. Association for Computational Linguistics.

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

[38] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi.
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, 2020.

[39] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi.
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, October-November 2018. Association for Computational Linguistics.

[40] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi.
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, July 2019. Association for Computational Linguistics.

[41] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant.
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, 2019.

Thanks