Enhancing GraphRAG with Beam Search-Based Path Filtering and Semantic Diversity Score

Anonymous ACL submission

Abstract

Graph-based Retrieval-Augmented Generation (GraphRAG) enhances Large Language Models (LLMs) by integrating structured knowledge graphs, but it faces challenges in suboptimal path selection and redundant entity retrieval. To address these, we propose three key improvements: (1) **LLM-driven Structured Entity Extraction**, which enhances query understanding by extracting structured entities prior to retrieval; (2) **Beam Search-based Path Filtering**, which selects globally coherent reasoning paths over greedy nearest neighbor search; and (3) **Semantic Diversity Score (SDS)**, a novel metric that reduces redundancy by quantifying the diversity of retrieved entity clusters.

We evaluate our approach on multiple-choice QA datasets: MCTest, LexGLUE CaseHold, PubMedQA, and MedQA. Our method improves accuracy by **+1.16%**, **+6.53%**, **+4.9%**, and **+0.31%** compared to the baseline LLaMA 3.1-8b, demonstrating enhanced retrieval informativeness and path coherence. Additionally, experiments on various LLMs, including Qwen2.5-7B, Gemma2-9B, and LLaMA 3.1-8B, show accuracy increases of **+12.34%**, **+22.50%**, and **+1.33%** on MCTest, respectively. While our method improves factual consistency and reasoning quality, further work is needed to adapt SDS to domain-specific tasks such as biomedical question answering.

1 Introduction

Large Language Models (LLMs) have achieved state-of-the-art performance across numerous NLP tasks (Zhao et al., 2024), but they still suffer from limitations such as knowledge staleness and hallucination (Petroni et al., 2019). Retrieval-Augmented Generation (RAG) has emerged as a promising solution, enabling LLMs to retrieve relevant external knowledge to improve factual accuracy (Lewis et al., 2020).

Recently, **GraphRAG** has been proposed as an extension of RAG that integrates structured knowledge graphs (KGs) into retrieval, enhancing LLM reasoning by leveraging structured entity-relationship graphs (Edge et al., 2024; Procko and Ochoa, 2024). GraphRAG constructs an entity-relation Graph from a knowledge graph and retrieves subgraphs relevant to the query (Xu et al.). However, despite its advantages, GraphRAG still faces the following challenges: 1)Traditional word embedding methods while effective for local similarity matching, exhibit limited capacity to capture nuanced contextual semantics or model long-range dependencies in knowledge graphs(Zhou et al., 2020). 2)Existing implementations often prioritize entity-level relevance over graph structural coherence, thereby ignoring hierarchical relationships and topological constraints inherent in KGs(Shi et al.). 3) Dense retrieval mechanisms can cause redundant information to mask key knowledge elements(Karpukhin et al., 2020).

To address these challenges, we propose an enhanced GraphRAG framework with the following key contributions:

- **LLM-driven Structured Entity Extraction**: We improve query understanding by extracting structured entities and keywords using an LLM before retrieval.
- **Beam Search-based Path Filtering**: Instead of greedy nearest neighbor selection, we employ beam search to optimize retrieval paths and improve global coherence.
- **Semantic Diversity Score (SDS)**: We introduce SDS as a novel metric to maximize informativeness while reducing redundant information within retrieved paths.

We evaluate our method based on **Llama3.1-8B**(Grattafiori et al., 2024) on four multiple-choice

QA datasets: **MCTest**, **LexGLUE CaseHold**, **PubMedQA**, and **MedQA**(Chalkidis et al., 2022; Richardson et al., 2013; Jin et al., 2019, 2020). Our approach demonstrates consistent improvements in accuracy, highlighting the effectiveness of structured retrieval and path optimization within GraphRAG. Additionally, we test the performance of **Qwen2-7B**(Qwen et al., 2025) and **Gemma2-9B**(Team et al., 2024) on MCTest, demonstrating that our method provides consistent enhancements across different large models.

2 Related Work

Retrieval-Augmented Generation (RAG) Retrieval-Augmented Generation (RAG) has emerged as a promising paradigm that enhances large language models (LLMs) with external knowledge retrieval to mitigate issues such as hallucination and factual inconsistency. Since the introduction of RAG by Lewis et al.(Lewis et al., 2020), various works have improved the retrieval and generation components by leveraging dense retrieval(Karpukhin et al., 2020), hybrid search strategies(Izacard and Grave, 2021), and adaptive retrieval selection(Ram et al., 2023). However, most RAG implementations rely on vector-based similarity search, which lacks structured reasoning over retrieved knowledge.

Graph-Based Retrieval in RAG To address the limitations of flat dense retrieval, GraphRAG was introduced as a structured retrieval approach that organizes knowledge into a graph and enables retrieval based on entity relationships. Recent works have explored knowledge graphs (KGs) and graph neural networks (GNNs) for retrieval(Atif et al., 2023; Shi et al., 2024). Graph-based retrieval has demonstrated advantages in multi-hop reasoning and contextual disambiguation, particularly in domains such as scientific literature retrieval(Agarwal et al., 2024) and biomedical knowledge graphs(Soman et al.). However, existing GraphRAG implementations often rely on greedy nearest neighbor search, ranking entities based on cosine similarity with the query, which may lead to suboptimal retrieval paths.

Path Optimization in Knowledge Graphs Path selection and optimization are critical challenges in graph-based information retrieval, particularly in multi-hop reasoning tasks. Early approaches primarily relied on shortest-path algorithms and random walk-based methods for traversing knowledge graphs(Tong et al., 2006). More recent work has explored beam search strategies to enhance multi-hop reasoning in knowledge graphs(Xiong et al., 2017), aiming to ensure that the retrieved paths remain both semantically informative and globally coherent. However, existing methods often rely on heuristic-based ranking mechanisms and lack an explicit optimization criterion for selecting paths that best preserve contextual relevance and knowledge consistency.

3 Approach

In this section, we introduce our approach for improving entity retrieval and path selection in GraphRAG. We propose two key enhancements: (1) **Structured Query-based Entity Extraction**, which refines entity retrieval by explicitly extracting relevant entities from the query, and (2) **Beam Search-based Path Optimization**, which incorporates a novel **Semantic Diversity Score (SDS)** to improve the informativeness and coherence of selected paths.

Figure 1 illustrates the overall workflow of our approach. The process consists of three main steps: first, structured entity extraction is applied to identify relevant entities from the query. Second, a subgraph is constructed based on retrieved entities and their relationships. Finally, beam search is performed to optimize path selection, leveraging SDS to encourage diverse and informative knowledge paths.

3.1 Structured Entity Extraction for Graph-Based Retrieval

Traditional GraphRAG methods retrieve relevant entities by directly comparing the query embedding with entity embeddings using cosine similarity. However, this approach can be suboptimal, as queries often contain multiple relevant entities or concepts that are not well captured by a single embedding.

To address this, we first extract structured entity representations from the query using a large language model (LLM). Given a natural language query, the model outputs a set of potentially relevant entities and key concepts. Formally, given a query q, the model produces:

$$E_q = \{e_1, e_2, ..., e_n\}, \quad K_q = \{k_1, k_2, ..., k_m\} \tag{1}$$

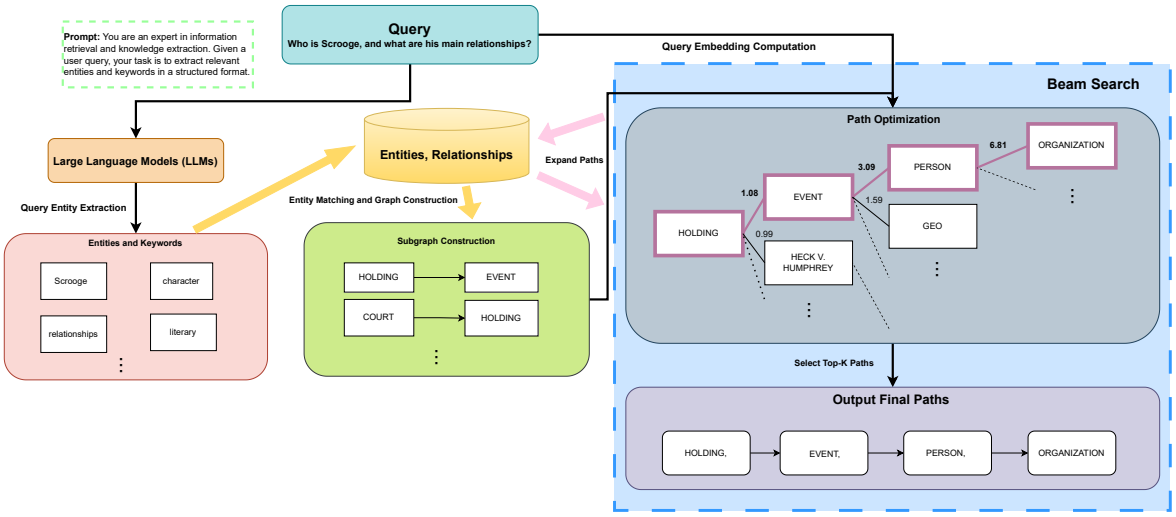

Figure 1: Overview of our proposed GraphRAG-based retrieval method with Beam Search.

where E_q represents extracted entity candidates and K_q denotes relevant keywords. These extracted elements are then used for graph-based retrieval, replacing the traditional method that relies solely on query embeddings.

3.2 Path Filtering with Beam Search and Semantic Diversity Score (SDS)

Once a set of relevant entities is identified, we construct a subgraph by selecting k-hop neighbors around them. Given this subgraph, our goal is to extract coherent and informative reasoning paths that maximize informativeness while avoiding redundant or noisy entities.

3.2.1 Beam Search for Path Exploration

As shown in Figure 2, we employ beam search to explore multiple candidate paths, evaluating them based on entity relevance, relationship strength, and semantic diversity. Beam search maintains a set of top-ranked paths at each expansion step, ensuring global coherence. The paths are scored using a weighted function that incorporates these factors.

3.2.2 Semantic Diversity Score (SDS) for Path Optimization

To explicitly optimize the informativeness of retrieved paths, we introduce **Semantic Diversity Score (SDS)** as a penalty term in the beam search scoring function. SDS measures the diversity of semantic information within a path, encouraging paths that contain more non-redundant information while discouraging paths with excessive semantic overlap.

For a given path $P = \{e_1, e_2, ..., e_k\}$, we compute **SDS** as follows:

1. Compute the **pairwise cosine similarity matrix** of entity embeddings along the path.
2. Perform **clustering** based on a similarity threshold τ, grouping semantically equivalent entities.
3. Compute the Shannon entropy of the cluster distribution:

$$SDS(P) = -\sum_i p_i \log p_i \tag{2}$$

where p_i represents the proportion of entities in cluster i.

Paths with higher SDS contain more diverse and complementary information, while paths with low SDS are penalized.

To ensure the selected paths are not only diverse but also relevant and globally coherent, we incorporate **Relevance Score**, **Relation Score**, and **Path Length Penalty** into the scoring function:

$$\begin{aligned} \text{Score}(P) = &\sum_{e \in P} \text{Relevance}(e) \\ &+ \gamma \sum_{(e_i, e_j) \in P} \text{Relation}(e_i, e_j) \\ &- \lambda \cdot SDS(P) \\ &- \mu \cdot |P| \end{aligned} \tag{3}$$

where $\sum_{e \in P} \text{Relevance}(e)$ measures the individual importance of each entity in the path with

respect to the query. $\sum_{(e_i,e_j)\in P} \text{Relation}(e_i, e_j)$ quantifies the strength of semantic connections between entity pairs along the path, weighted by γ. $SDS(P)$ penalizes redundant paths with high semantic overlap, controlled by λ. $|P|$ represents the length of the path, with μ acting as a penalty factor to discourage unnecessarily long paths.

By integrating these components, our method prioritizes paths that are semantically rich, well-connected, and concise, leading to improved retrieval quality in GraphRAG.

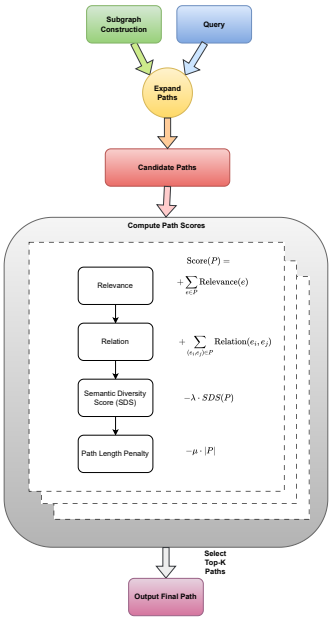

Figure 2: Path Scoring Framework

3.3 End-to-End Query Processing Pipeline

Our full retrieval and reasoning pipeline is as follows:

1. **Entity Extraction**: Use an LLM to extract structured entities and keywords from the query.

2. **Subgraph Construction**: Retrieve relevant entities and construct a local knowledge subgraph.

3. **Beam Search Path Filtering**: Generate candidate reasoning paths and rank them using SDS.

4. **Final Context Selection**: Retrieve the top-ranked paths and use them for response generation.

This approach ensures that the retrieved knowledge is **contextually relevant**, **semantically diverse**, and **globally coherent**, significantly improving the effectiveness of GraphRAG for knowledge-intensive tasks.

4 Experiments

4.1 Experimental Setup

We evaluate our approach on multiple-choice question-answering tasks across four datasets: MCTest, LexGLUE CaseHold, PubMedQA (pqd_label) and MedQA. We utilize LLaMA 3.1-8B as the underlying large language model (LLM) and do not apply any fine-tuning, relying solely on its inference capabilities. Accuracy is used as the primary evaluation metric.

We evaluate the effectiveness of our approach across different Large Language Models (LLMs), including Qwen2.5-7B, Gemma2-9B, and LLaMA 3.1-8B.

The experiments were conducted on a system equipped with two NVIDIA A40 GPUs.

4.2 Results

Table 1 presents the accuracy of different configurations on the three datasets. Our baseline is the LLaMA 3.1-8B model. We then compare various enhancements strategies:

- **GraphRAG**: Standard GraphRAG retrieval.
- **GraphRAG + Beam Search**: Path filtering using beam search.
- **GraphRAG + LLM Entity Extraction**: Query entities extracted using an LLM before retrieval.
- **GraphRAG + LLM Entity Extraction + Beam Search**: A combination of both optimizations.

Additionally, we investigate the performance of our method across different LLMs on the MCTest dataset. As shown in Table 2, we achieve significant accuracy improvements across all models.

4.3 Analysis

The results demonstrate the effectiveness of integrating GraphRAG into retrieval-augmented generation (RAG). Across most datasets, GraphRAG consistently improves accuracy compared to the vanilla LLaMA 3.1-8B model, with notable gains

Table 1: Ablation Study: Accuracy (%) on different datasets with various retrieval and path optimization strategies. The base LLM is **LLaMA 3.1-8B**.

Method	MCTest	LexGLUE CaseHold	PubMedQA	MedQA
LLaMA 3.1-8B	93.17%	55.47%	71.90%	59.31%
GraphRAG (Base)	94.00%	60.67%	76.80%	59.07%
+ Beam Search (Path Filtering)	93.33%	61.81%	76.30%	59.31%
+ LLM Entity Extraction	94.17%	61.86%	76.40%	59.54%
+ Beam Search + SDS (Full Model)	**94.50%**	**61.92%**	**76.80%**	**59.62%**

Table 2: Accuracy (%) comparison on MCTest dataset using different LLM models with and without our method.

Model	Accuracy (Original LLM)	Accuracy (With Our Method)
Qwen2.5-7B	78.33%	90.67%
Gemma2-9B	74.83%%	97.33%
LLaMA 3.1-8B	93.17%	94.50%

in LexGLUE CaseHold and PubMedQA (+5.2% and +4.9%, respectively). However, the improvement on MedQA is more modest (+0.31% for the full model), likely due to its reliance on specialized medical reasoning beyond general knowledge retrieval.

The use of beam search for path filtering further refines retrieval, leading to slight accuracy improvements in LexGLUE CaseHold (+1.14%) and PubMedQA (+0.2%), but shows limited impact on MedQA (+0.31%). Notably, on MCTest, beam search slightly reduced accuracy (-0.67%), likely due to overly aggressive filtering of relevant paths.

The best performance across all datasets is achieved when LLM-based entity extraction is combined with beam search and SDS. This suggests that structured entity extraction is crucial for selecting informative paths, even in domains like MedQA where semantic diversity may conflict with factual specificity.

Table 2 shows our method significantly improves the performance of various LLMs across the MCTest dataset. The accuracy for Qwen2.5-7B improves from 78.33% to 90.67%, representing a +12.34% gain. For Gemma2-9B, the accuracy increases from 74.83% to 97.33%, which is a +22.50% improvement. LLaMA 3.1-8B, on the other hand, improves from 93.17% to 94.50%, a more modest increase of +1.33%.

4.4 Computational Efficiency

Despite accuracy improvements, adding beam search and entity extraction introduces additional computational overhead. In future work, we aim to optimize the efficiency of these steps while maintaining retrieval quality.

4.5 Performance on PubMedQA and MedQA

While our approach demonstrates consistent improvements on MCTest and LexGLUE CaseHold, the performance on PubMedQA exhibits a slight decrease when applying Beam Search-based path filtering. Specifically, GraphRAG achieves 76.80% accuracy, while GraphRAG with Beam Search slightly drops to 76.30%, and GraphRAG with entity extraction + Beam Search achieves 76.40%. On MedQA, the full model achieves 59.62%, only marginally higher than the baseline 59.31%.

We hypothesize that this phenomenon is due to the following factors:

1. **Precision-sensitive nature of medical QA:** Unlike general-domain QA tasks, medical question answering relies on *highly specific factual knowledge*. *Beam Search* may *filter out critical medical concepts*, leading to information loss and reduced performance.

2. **Suboptimal Semantic Diversity Score (SDS) for medical text:** Our path optimization relies on SDS, which encourages diverse paths. However, in medical QA, *the most relevant paths are often highly specific rather than diverse*, making SDS less effective in this domain.

3. **Limitations of LLaMA 3.1-8B on medical reasoning:** The underlying LLM may not

be as specialized in medical knowledge as domain-specific models such as *BioBERT* or *Med-PaLM*, which could further limit performance gains.

Future work may explore *task-specific path filtering strategies*, incorporating *medical knowledge graphs* (e.g., UMLS, SNOMED CT) to improve entity selection. Additionally, adapting *SDS for factual consistency rather than diversity* may further enhance performance on medical QA tasks.

5 Conclusion

In this work, we propose an improved retrieval-augmented generation (RAG) framework by enhancing entity retrieval and path filtering in GraphRAG. Our approach first utilizes a large language model (LLM) to extract potential entities from the query before performing entity retrieval, ensuring more accurate and relevant entity selection. Additionally, we introduce a beam search-based path filtering strategy, incorporating a Semantic Diversity Score (SDS) to balance path informativeness and redundancy.

Experimental results on multiple-choice QA datasets demonstrate the effectiveness of our method. Our approach consistently improves retrieval quality, leading to better downstream task performance. Notably, we achieve substantial gains in accuracy across datasets without any fine-tuning of the LLM. Our method demonstrates consistent improvement across multiple LLMs. The results highlight the versatility and effectiveness of our approach in enhancing the performance of a wide range of models, suggesting that it can be generalized to various language models in future work.

Despite these improvements, our method has certain limitations. The reliance on LLMs for entity extraction introduces additional computational overhead, and SDS-based path filtering may not generalize equally well across different domains. Future work will explore more efficient entity extraction techniques, adaptive path selection mechanisms, and broader evaluation on open-domain question answering tasks.

Limitations

While our approach improves entity retrieval and path selection in GraphRAG, it also introduces several limitations.

First, the reliance on a large language model for query-based entity extraction adds additional computational overhead. Although this step enhances retrieval quality, it may not be feasible for real-time applications or resource-constrained environments. Future work should explore more efficient entity extraction mechanisms, such as lightweight neural classifiers or retrieval-based heuristics.

Second, our Semantic Diversity Score (SDS)-based path filtering prioritizes paths with diverse entity semantics, but it does not explicitly model query-specific relevance beyond cosine similarity. This may result in suboptimal path selection when the most relevant entities are semantically similar. Adaptive weighting strategies that incorporate task-specific signals could further refine our approach.

Finally, our experiments focus only on multiple-choice QA datasets. While the results demonstrate consistent improvements, further evaluation is needed on broader NLP tasks, such as open-domain question answering, document-grounded generation, and structured reasoning benchmarks.

Ethics Statement

Our work focuses on improving entity retrieval and path selection in GraphRAG for knowledge-based reasoning. The datasets used in our experiments are publicly available and widely used in the NLP community. No personally identifiable information (PII) or sensitive data is involved in our study.

However, as our approach relies on large language models (LLMs) for entity extraction, potential biases in LLM-generated outputs could propagate into the retrieval process. While our method does not introduce new biases, future work should explore mitigation strategies to ensure fairness and robustness in knowledge retrieval across different domains.

References

Shubham Agarwal, Issam H. Laradji, Laurent Charlin, and Christopher Pal. 2024. Litllm: A toolkit for scientific literature review. *Preprint*, arXiv:2402.01788.

Farah Atif, Ola El Khatib, and Djellel Difallah. 2023. Beamqa: Multi-hop knowledge graph question answering with sequence-to-sequence prediction and beam search. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 781–790, New York, NY, USA. Association for Computing Machinery.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark

dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *Preprint*, arXiv:2404.16130.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe

Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. 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 2567–2577, Hong Kong, China. Association for Computational Linguistics.

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.

Patrick 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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. 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.

Tyler Thomas Procko and Omar Ochoa. 2024. Graph retrieval-augmented generation for large language models: A survey. In *2024 Conference on AI, Science, Engineering, and Technology (AIxSET)*, pages 166–169.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Yucheng Shi, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, and Ninghao Liu. Retrieval-enhanced Knowledge Editing in Language Models for Multi-Hop Question Answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, pages 2056–2066. Association for Computing Machinery.

Yucheng Shi, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, and Ninghao Liu. 2024. Retrieval-enhanced knowledge editing in language models for multi-hop question answering. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2056–2066, New York, NY, USA. Association for Computing Machinery.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, Sharat Israni, Charlotte A Nelson, Sui Huang, and Sergio E Baranzini. Biomedical knowledge graph-optimized prompt generation for large language models. 40(9):btae560.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Hanghang Tong, Christos Faloutsos, and Jia-yu Pan. 2006. Fast random walk with restart and its applications. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 613–622.

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. DeepPath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 564–573, Copenhagen, Denmark. Association for Computational Linguistics.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering. *Preprint*, arXiv:2404.17723.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang,

Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024. A Survey of Large Language Models. *Preprint*, arXiv:2303.18223.

Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81.