GRAFT: Graph Retrieval Augmented Fine Tuning for Multi-Hop Query Summarization

Stanford CS224N Custom Project

Sonya Jin
Department of Computer Science
Stanford University
sonyajin@stanford.edu

Sunny Yu
Department of Computer Science
Stanford University
syu03@stanford.edu

Natalia Kokoromyti
Department of Computer Science
Stanford University
knatalia@stanford.edu

Abstract

Traditional retrieval-augmented generation (RAG) approaches struggle with multi-hop reasoning and global query-focused summarization tasks over large document corpora, which require summarizing broad themes and contexts and a holistic knowledge of documents. We propose GRAFT (Graph Retrieval Augmented Fine-Tuning), a novel approach that combines the strengths of the Retrieval Augmented Fine-Tuning (RAFT) methodology and the GraphRAG technique. GRAFT fine-tunes large language models (LLMs) on a simulated imperfect retrieval setting, training the model to identify relevant documents and ignore distractors in the provided context. The model is then coupled with graphRAG at inference. To investigate the effectiveness of the GRAFT methodology, we constructed a knowledge graph using 74 Wikipedia source documents and extracted communities within this graph. We then summarized these communities, leveraging local and global relationships between documents for retrieval, fine-tuned a Microsoft Phi-2 model using the RAFT approach on a subset of the HotPotQA dataset, and evaluated its performance on a custom set of multi-hop and global questions generated from Wikipedia articles published in 2024. Our experimental results demonstrate that GRAFT outperforms baseline models, including the Baseline RAG model, the RAFT model, and the Baseline GraphRAG model, across various evaluation metrics like BERT, BLEU, ROUGE-1, and Semantic Similarity. In particular, GRAFT achieves the highest scores on global questions, showcasing its effectiveness in query-focused summarization tasks that require understanding broad themes and contexts over large document corpora.

1 Key Information to include

- Mentor: Aditya Agrawal
- External Collaborators (if you have any): N/A
- Sharing project: N/A
- Contributions:
  1. Sonya: Data collection, pre-processing, fine-tuning (RAFT code), knowledge graph construction, GRAFT pipeline design and code, baseline RAG and baseline GraphRAG code, evaluation, report.
  2. Sunny: Data collection, knowledge graph construction, baseline RAG code, evaluation, report.
  3. Natalia: Data collection, knowledge graph construction, report

2 Introduction

Large language models (LLMs) have shown remarkable capabilities in natural language processing tasks, but they often struggle with hallucinations, coherence, and factual consistency, especially in complex, domain-specific scenarios [1, 2]. Retrieval-Augmented Generation (RAG) systems, which combine LLMs with external knowledge bases, have emerged as a promising approach to mitigate these limitations by grounding the model's outputs in factual knowledge and ensuring their consistency with the input context [3]. Despite their potential, traditional RAG systems face challenges in retrieval accuracy, contextual understanding, and response coherence when dealing with multi-hop reasoning and global query-focused summarization tasks over large document corpora [4].

This paper focuses on two types of questions that pose challenges to traditional RAG methods: multi-hop questions, which require reasoning across multiple pieces of information, and global questions, which demand an understanding of an entire dataset. To improve model performance on these question types, we propose GRAFT, a novel approach that combines the RAFT approach [5] with the GraphRAG [6] approach, capturing global and local information within the documents through extracted communities from a knowledge graph.

Our results show that the GRAFT model outperforms all other models, and both RAFT and Baseline GraphRAG outperform the baseline RAG model, which indicates that the combination of fine-tuning and community summary extraction improves answer quality further. Furthermore, compared to baseline RAG and RAFT, the baseline GraphRAG model and GRAFT both show better performance on global questions than non-global questions on most evaluation metrics.

3 Related Work

3.1 RAG and Its Current Limitations

Existing studies have advocated for the need for LLMs to move beyond simply reciting the training data [7] to actively seek out and synthesize information from supplemental knowledge bases to tackle complex, context-rich queries. To this end, RAG marries the open-ended generation capabilities of large language models with targeted information retrieval by first scouring the corpus, surfacing relevant passages, feeding the extracts with the original query, and finally producing an output [8]. Evaluations have shown that RAG models generate more specific, diverse and factual language than seq2seq baselines [3]. RAG has also been applied to fact-checking [9] and has shown promise in document-grounded dialogue [10].

Despite the strengths of current RAG systems, challenges such as maintaining the relevance of the retrieved information and handling complex, domain-specific scenarios remain. In particular, RAG struggles with global questions, such as identifying the main themes throughout an entire text corpora, as this task requires query-focused summarization (QFS) rather than explicit retrieval [11] . The key limitation of RAG systems in answering global questions stems from their top-k retrieval method. By focusing on retrieving only the most relevant passages, RAG fails to capture the broader context and overarching themes necessary to address global queries effectively.

3.2 Advanced RAG Techniques

To address traditional RAG methods' limitation of proficiency in multi-hop reasoning, researchers have proposed several advanced techniques, including Forward-Looking Active REtrieval augmented generation (FLARE) [12], self-memory techniques [13], prompt-guided retrieval augmentation [14], retrieval augmented reinforcement learning [15], Dense Knowledge Retrieval [16], Iter-RetGen [17] (which synthesizes retrieval and generation iteratively), Generation-Augmented Retrieval (GAR) [18], and LTRGR [19], which combines generative retrieval with the classical learning-to-rank paradigm.

While the techniques outlined above can improve question-answering accuracy for open-domain tasks, they fail to tackle multi-hop and global questions. Another common approach focuses on fine-tuning RAG models [20]. One such approach is RAFT (Retrieval-Augmented Fine-Tuning), which enhances the model's ability to differentiate between relevant and irrelevant documents by employing a chain-of-thought fine-tuning process.

Another common approach to tackle global questions relies on graph construction [21]. In particular, GraphRAG [11] constructs a knowledge graph from source documents, explicitly modeling the relationships between different pieces of information and uses community summaries and hierarchical indexing to generate comprehensive answers, which effectively captures semantic and contextual relationships.

Moreover, hierarchical methods, such as HiQA (Hierarchical Contextual Augmentation) [22], have also shown promise in addressing the challenges of multi-hop reasoning by integrating cascading metadata and a multi-route retrieval mechanism to enhance precision and relevance in knowledge retrieval.

The three methods present distinct strengths: RAFT allows the model to better identify relevant information in retrieved documents; GraphRAG relies on an explicit knowledge graph representation; HiQA leverages a hierarchical metadata structure and multi-route retrieval to identify relevant information. However, no existing study has explored the confluence of these approaches. Our study fills this gap by combining supervised fine-tuning with knowledge graph construction to examine the performance of such an approach.

4 Approach

4.1 RAFT-inspired Finetuning on Multi-Hop Reasoning Task

We follow the approach outlined in RAFT [5] by fine-tuning the Microsoft Phi-2 model [23] using relevant and distractor documents. Specifically, we fine-tuned Phi-2 with QLoRA [24] on a subset of the HotPotQA dataset [25], modifying the training data to align with the RAFT methodology. Each training example consisted of a question, an answer, and a set of passages: 1-2 oracle passages required to answer the question, and 3-4 irrelevant distractor passages. The preprocessing steps were as follows:

1. Extracting the relevant (oracle) documents and padding with irrelevant documents until the maximum token length of 2048 was reached.

2. Constructing the input prompt by combining the instruction, question, context, and answer.

3. Shuffling the order of the context passages to prevent the model from relying on positional cues (relevant documents tended to be first).

We used Retrieval Augmented Generation (RAG) [3] to enhance the performance of question answering models by leveraging an external knowledge base. The following steps were taken:

- **Embedding and Storage:** Wikipedia articles were chunked into passages of 800 tokens each with 100-token overlap and embedded using the 'all-MiniLM-L6-v2' model from Sentence Transformers [26]. The embeddings were then stored in a FAISS vector database [27].
- **Retrieval:** During inference, for each question, the top 3 relevant passages were retrieved from the vector database based on cosine similarity between the question embedding and the passage embeddings.
- **Generation:** The retrieved passages were provided as additional context to the finetuned Microsoft Phi-2 model at inference.

4.2 GraphRAG

Our GraphRAG approach involves the following steps:

1. **Text Chunking**: The source documents are split into smaller text chunks of a fixed token length (e.g., 800 tokens) with overlapping windows to maintain context continuity.

2. **Entity and Relation Extraction**: A large language model (LLM), notably GPT-4o, is used to extract entity mentions and relation mentions from each text chunk. This is done by providing the LLM with prompts tailored to the domain, which contain few-shot examples of entities and relations to learn from. The LLM outputs tuples containing the entity/relation text, type, and description.

3. **Graph Construction**: The extracted entity and relation mentions are used to construct an undirected, weighted graph. Each unique entity mention becomes a node in the graph. If a relation mention links two entity mentions, an edge is created between their corresponding nodes. The edge weight is assigned based on the frequency of that specific relation mention across all text chunks.

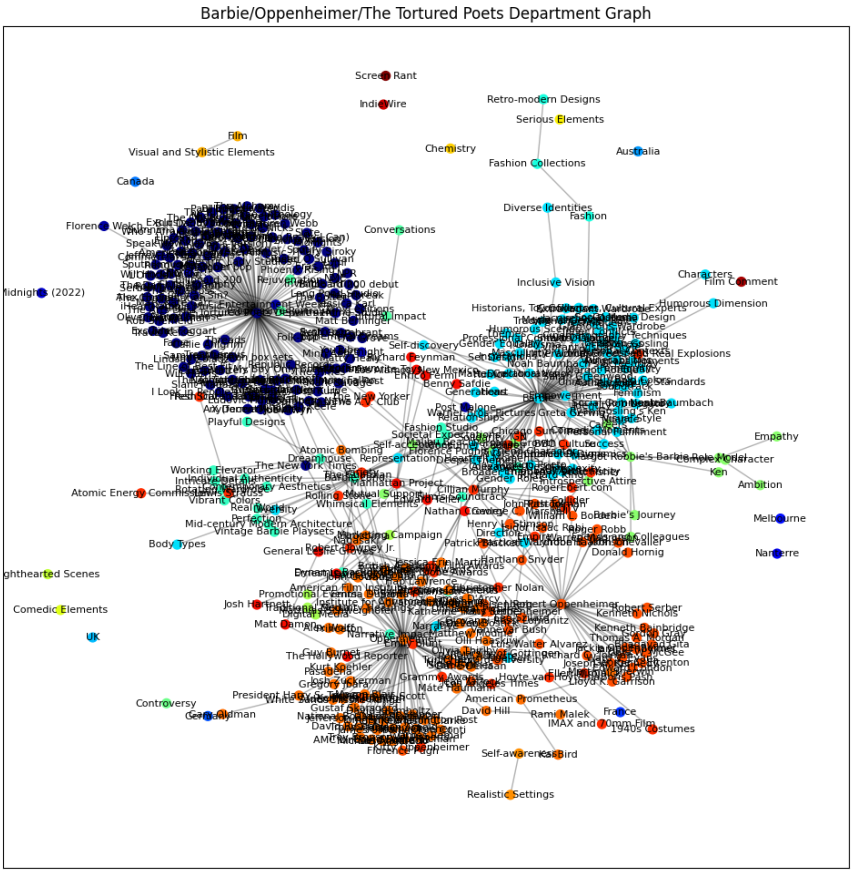

Figure 1: GPT-4o Knowledge Graph, color-coded by community. The graph is constructed based on Wikipedia articles on Oppenheimer, The Barbie Movie (2023), and The Tortured Poets Department. The cluster in dark blue represent nodes constructed from The Tortured Poets Department. The top right cluster is from Barbie, and the bottom cluster is from Oppenheimer.

4. **Community Detection**: The Leiden algorithm [28] is applied to the constructed graph to identify communities of closely related nodes (entities) based on their edge connections and weights. This results in a hierarchical partition of the graph into communities at different levels.
5. **Community Summarization**: For each community at each level of the hierarchy, the LLM generates a concise summary describing the entities, relations, and other relevant information contained within that community. These summaries act as the index for retrieval during query time. A comparison of a Leiden extracted community and its generated community summary is shown in 10
6. **Query Processing**: Given a user query, the top k = 1 relevant community summary is retrieved by computing the cosine similarity between the query embedding and the pre-computed embeddings of the community summaries. The summary is then chained with the query for input into the LLM. The answer is then generated with this retrieval-augmented generation (RAG) approach.

5 Experiments

5.1 Data

For our experiments, we utilize two datasets: the HotPotQA dataset [25] for model fine-tuning, and a custom dataset consisting of Wikipedia articles for evaluating global question answering.

HotPotQA The HotPotQA dataset [25] is a multi-hop reasoning dataset that requires retrieving and combining information from multiple supporting documents to answer each question. It contains 112,779 questions based on Wikipedia articles, with each question being accompanied by a set of relevant documents. The

questions are designed to test a model's ability to perform multi-hop reasoning and gather information from multiple sources. We utilize this dataset to fine-tune the RAFT model.

Global Question Evaluation Dataset To evaluate our models' performance on global question answering, we curated a custom dataset from Wikipedia articles. Specifically, we selected 74 source documents from the Wikipedia 2024 corpus, which includes articles published after the model's training cutoff date. From these articles, we constructed a set of 67 global questions that require understanding the broad themes and overall context of the source documents. These questions are designed to assess a model's ability to perform global summarization and answer queries that necessitate comprehending the entire document corpus, rather than focusing on specific details or facts.

The global questions in our evaluation dataset (examples shown in 9) are intended to simulate real-world scenarios where users seek high-level insights and overarching themes from a collection of documents. By evaluating our models on this custom dataset, we can assess their effectiveness in performing query-focused summarization and providing comprehensive answers that capture the overall context and main topics of the source documents.

5.2 Evaluation method

For each question, we provided GPT-4 with the source documents and used model-generated responses as the gold standard answers. We evaluated the question answering performance using both human evaluation and four other metrics, including BERT-Score [29], BLEU Score, ROUGE-1 Score [30], and Semantic Similarity [31]. The human evaluation reflects human judgment of the correctness of model responses given the gold standard answer, providing a direct assessment of the model's performance from a user's perspective. BERT-Score leverages the contextual embeddings of the generated response and reference and is calculated by aggregating token-level cosine similarities between the generated response and reference. The BERT-Score is a helpful metric for our purpose because it captures semantic similarity and can handle paraphrasing, which is crucial in evaluating the model's ability to generate semantically equivalent answers even if they differ in exact wording.

The BLEU (Bilingual Evaluation Understudy) Score is a metric for evaluating the quality of machine-translated text by comparing it to reference texts using N-gram matching. While originally designed for machine translation, BLEU has been widely adopted in various natural language generation tasks, including question answering. It provides a measure of the model's ability to generate answers that match the reference in terms of n-gram overlap.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics used to evaluate the quality of text summarization and machine-generated text by comparing it to reference texts. Specifically, ROUGE-1 calculates the overlap of unigrams, which helps assess the model's ability to generate relevant content at the word level. Similar to BERT-Score, Semantic Similarity measures how much two pieces of text are similar in terms of their meaning by calculating the cosine similarity between word embeddings. This metric allows for contextual relevance and flexibility in expression, capturing the model's ability to generate answers that are semantically similar to the reference, even if they use different words or phrasing.

By using a combination of human evaluation and these four automated metrics, we aim to comprehensively assess the model's performance in generating accurate, semantically relevant, and fluent answers, while accounting for various aspects of language generation quality.

5.3 Experimental details

We fine-tuned a Microsoft Phi-2 model using the PEFT (Parameter-Efficient Fine-Tuning) technique, specifically utilizing the QLoRA (Quantized Low-Rank Adaptation) method on a subset of the HotPotQA dataset (20,000 samples). QLoRA combines LoRA with 4-bit quantization to reduce memory usage and enable efficient training on a single GPU.

Model Configuration:

- Base Model: Microsoft Phi-2
- Quantization: 4-bit quantization using BitsAndBytesConfig
- QLoRA Configuration:
  - Rank (r): 32
  - LoRA alpha: 32

  - Target modules: [q_proj, k_proj, v_proj, out_proj, 'dense']
  - LoRA dropout: 0.05

Machine Configuration:

- GPU: NVIDIA T4
- GPU Memory: 16 GB
- CPU: Intel Xeon
- RAM: 64 GB

Fine-tuning Hyperparameters:

- Number of epochs: 1
- Per device train batch size: 1
- Gradient accumulation steps: 8
- Learning rate: 2e-4
- Optimizer: 8-bit Paged AdamW

5.4 Results

Figures 2, 3, and 4 present the evaluation results on global questions. The GRAFT model achieves the highest scores across all metrics: BERT (0.330), BLEU (0.143), ROUGE-1 (0.428), and Semantic Similarity (0.863), outperforming the Baseline RAG model, RAFT, and Baseline GraphRAG. For each metric, the best score is bolded.

Metric	Baseline RAG model	RAFT	Baseline GraphRAG	GRAFT
Human Eval	0.310	0.500	0.413	**0.620**
BERT	-0.0196	0.0575	**0.346**	0.330
BLEU	0.0130	0.0344	0.0290	**0.143**
ROUGE-1	0.168	0.239	0.288	**0.428**
Semantic Sim	0.620	0.669	0.753	**0.863**

Figure 2: Evaluation results for global questions

Metric	Baseline RAG model	RAFT	Baseline GraphRAG	GRAFT
Human Eval	0.263	0.393	**0.605**	0.526
BERT	0.00364	0.0634	0.298	**0.312**
BLEU	0.0194	0.0321	0.0245	**0.146**
ROUGE-1	0.188	0.242	0.270	**0.397**
Semantic Sim	0.563	0.634	0.648	**0.791**

Figure 3: Evaluation results for non-global question

Metric	Baseline RAG model	RAFT	Baseline GraphRAG	GRAFT
Human Eval	0.284	0.403	0.522	**0.567**
BERT	-0.00645	0.060	0.319	**0.320**
BLEU	0.0166	0.0331	0.0265	**0.145**
ROUGE-1	0.179	0.240	0.278	**0.410**
Semantic Sim	0.588	0.649	0.694	**0.822**

Figure 4: Joint metrics (whole dataset)

Figure 3 shows the evaluation results on non-global questions. GRAFT maintains its superior performance, achieving the highest scores on BERT (0.312), BLEU (0.146), ROUGE-1 (0.397), and Semantic Similarity (0.791). For the majority of the metrics, GRAFT has a higher score on global questions compared to non-global questions, whereas the opposite is observed for Baseline RAG, meaning that GRAFT achieves especially improved performance for global questions. Considering the joint metrics across the entire dataset in Table 4 4, GRAFT consistently outperforms all other models across BERT (0.320), BLEU (0.145), ROUGE-1 (0.410), and Semantic Similarity (0.822), demonstrating its versatility and robustness. Moreover, for all metrics, both RAFT and Baseline GraphRAG outperform Baseline RAG model, supporting previous findings that fine-tuning and graph construction improve question-answering performance.

6 Analysis

6.1 Qualitative Evaluation

Method	Response	Response Summary
RAFT	The Barbie film series is a collection of American animated films produced by Mattel Creations and distributed by 20th Century Fox Home Entertainment. The film was released on DVD and Blu-ray on October 6, 2011. The film was released on DVD and Blu-ray on October 6, 2011. The film was released on DVD and Blu-ray on October 6, 2011. [...]	Repeatedly states incorrect DVD release information without addressing the question.
Baseline GraphRAG	The chemistry between Barbie and Ken, as portrayed by Margot Robbie and Ryan Gosling, is a key aspect of the film's narrative. Their interactions highlight the themes of mutual respect, partnership, and the search for individual identity within a relationship. The film also features strong supporting characters, such as Florence Pugh's friend character, who provide additional layers to the narrative. [...]	It doesn't specifically address Barbie and Ken's dynamic (mutual respect and partnership are irrelevant knowing the narrative structure of the film). Additionally, the response includes extraneous information that doesn't directly answer the prompt at hand.
GRAFT	The chemistry between the characters in the Barbie film is a key aspect of its storytelling. Barbie, portrayed by Margot Robbie, embarks on a journey of self-discovery that challenges traditional gender roles and highlights the importance of being true to oneself. Ryan Gosling's portrayal of Ken adds depth to the discussion of masculinity and the pressures men face to conform to certain stereotypes. The interplay between Barbie and Ken, along with other characters, emphasizes mutual respect, empathy, and the value of authentic relationships.	Concisely explains how character chemistry and relationships contribute to the film's narrative, focusing on key themes such as gender identity, societal expectations, and authenticity.

Figure 5: Comparison of GRAFT, RAFT, and Baseline GraphRAG responses on "How do the chemistry and relationships in the Barbie film contribute to its narrative?"

To gain a deeper understanding of GRAFT's performance, we compare the responses generated by GRAFT, RAFT, and Baseline GraphRAG for a selected example (Figure 5). RAFT's response repeatedly mentions incorrect DVD release information, failing to address the question about character relationships in the Barbie film. This indicates poor query understanding and irrelevant information retrieval. Baseline GraphRAG provides a more relevant account of Barbie and Ken's chemistry but includes the factual error of mentioning Florence Pugh, who is not in the film. This suggests issues with factual accuracy and extraneous information. GRAFT, however, generates a concise, focused response that directly addresses the question, accurately depicting character dynamics, themes, and narrative elements. GRAFT's response demonstrates superior coherence, relevance, and accuracy compared to RAFT and Baseline GraphRAG.

The context retrieved by RAFT and Baseline models (RAFT and baseline RAG) shown in 6 demonstrates a significant deficiency in relevance and specificity concerning the core aspects of the query, particularly character relationships and overarching social themes. These models tend to retrieve information that is either tangential or repetitive, failing to address the nuances of the query adequately. By examining the retrieved contexts and their impact on answer quality, we see how GRAFT's approach to capturing both global and local relationships significantly enhances its performance. In contrast, GRAFT's retrieved context, generated with its community-based summarization shown in 11, exhibits a marked improvement in both relevance and comprehensiveness. The community summary effectively identifies Barbie and Ken as pivotal entities and delves into their evolving dynamic relationship, emphasizing themes such as mutual respect, partnership, and the quest for individual identity. Moreover, the summary acknowledges the critical role of supporting characters in enriching the narrative and underscores the film's engagement with societal issues, including unrealistic beauty standards and traditional gender roles. The global perspective is evident in its understanding of broad social themes such as gender roles and unrealistic beauty standards. Simultaneously, the local relationships,

particularly the dynamics between Barbie and Ken and their interactions with supporting characters, provide depth and specificity to the response.

Model	Retrieved Context
RAFT	Mattel began adapting Barbie into various facets of media and entertainment beyond the television advertisement of its dolls and related accessories (which was a prolific marketing strategy in the past). For the first 16 entries in the film series, Barbie is featured as a virtual actress playing the main character, and often being portrayed as a modern girl telling the story to one of her sisters or a younger friend – as a parable to present affairs. Scholars examining how the Barbie films differ from other [...]
GRAFT	Margot Robbie and Ryan Gosling, as Barbie and Ken, exhibit a dynamic relationship that evolves throughout the film. Their interactions highlight the themes of mutual respect, partnership, and the search for individual identity within a relationship. The film also features strong supporting characters, such as Florence Pugh's friend character, who provide additional layers to the narrative. These relationships are not only central to the plot but also serve to explore broader social themes, including diversity, inclusivity, and the importance of supportive friendships. The character dynamics are further enriched by the film's witty dialogue and emotional authenticity, making the relationships believable and engaging. [...]

Figure 6: Retrieved Contexts for RAFT/Baseline and GRAFT

We see that the qualitative evaluation reveals that the GRAFT model significantly outperforms both RAFT and Baseline GraphRAG models in generating relevant, accurate, and coherent responses to complex queries. While RAFT struggled with coherence and relevance, and Baseline GraphRAG failed in factual accuracy despite providing detailed information, GRAFT excelled in delivering a focused and insightful answer.

7 Conclusion

In this work, we introduced GRAFT, a novel approach that unifies Retrieval Augmented Fine-Tuning (RAFT) with GraphRAG to address multi-hop reasoning and global query-focused summarization challenges in large document corpora. GRAFT outperformed strong baselines, including RAG, RAFT, and GraphRAG, across multiple automated evaluation metrics, demonstrating its effectiveness in answering both multi-hop and global questions. However, our work has limitations. The knowledge graph construction process relies on entity and relation extraction from a limited set of documents, potentially leading to a sparse graph that may not capture all relevant information for certain queries. Additionally, the computational resources available during our experiments, specifically the use of a single GPU, constrained the scale and complexity of the models we could train and evaluate. Despite these limitations, GRAFT opens up several avenues for future research:

1. Adaptive graph construction and maintenance: Investigating techniques to dynamically update the knowledge graph based on incoming queries could enable more targeted and efficient retrieval. This could involve developing algorithms to identify and incorporate relevant new information from the document corpus or external sources, ensuring the graph remains up-to-date and comprehensive, storing only a subgraph at a time as communities are extracted and embedded dynamically.
2. Explanation generation using graph reasoning: Enhancing GRAFT with the ability to generate human-readable explanations for its answers by leveraging the graph structure could improve the interpretability and trustworthiness of the system. This could involve using chain-of-thought reasoning in conjunction with citing specific relationships and nodes from the graph to provide a clear, logical justification for the generated answers.
3. Scalability and efficiency improvements: Exploring techniques to optimize the computational efficiency of GRAFT, such as distributed training, model compression, or more efficient graph traversal algorithms, could enable its application to larger-scale datasets and more complex question answering tasks.

By addressing these limitations and pursuing these research directions, GRAFT can be further developed into a powerful, scalable, and interpretable question answering system capable of tackling complex reasoning tasks over large textual corpora.

8 Ethics Statement

8.1 Ethical Challenge 1: Bias and Fairness in Knowledge Representation

One significant ethical challenge in this project is the potential for bias in the knowledge representations embedded within the GRAFT and RAFT models. These biases can arise from the underlying data sources, such as Wikipedia articles and community summaries, which may reflect societal prejudices and historical inequities. For instance, if certain topics or perspectives are underrepresented or misrepresented in the knowledge graph or the retrieved text, the model's outputs might perpetuate these biases, leading to unfair or skewed answers. This issue is particularly pertinent in our project, as the goal is to generate questions and answers that require holistic knowledge, which should ideally be comprehensive and unbiased.

Mitigation Strategy: To mitigate this risk, one strategy is to implement bias detection and correction mechanisms within the model's training and inference pipeline. This could involve using fairness-aware algorithms that detect and adjust for biased representations in the data. Additionally, regular audits of the data sources and the model's outputs by a diverse team of experts can help identify and rectify biases. Policies mandating transparency in the data collection and model training processes, as well as the inclusion of diverse perspectives, can also help mitigate bias and ensure fairer outcomes.

8.2 Ethical Challenge 2: Privacy Concerns with Data Usage

Another ethical concern involves privacy issues related to the data used for training and inference. Given that the project utilizes extensive data from Wikipedia articles and community summaries, there is a possibility that sensitive information about individuals could be inadvertently included in the knowledge graph or retrieved chunks. This raises significant privacy concerns, as the inclusion and potential dissemination of personal data could violate individuals' privacy rights.

Mitigation Strategy: To address this risk, it is crucial to implement rigorous data anonymization and privacy-preserving techniques throughout the data processing pipeline. This includes removing or obfuscating any personally identifiable information (PII) from the data before it is used for training or inference. Additionally, adhering to data privacy regulations, such as GDPR, and conducting regular privacy impact assessments can help ensure that the project complies with legal and ethical standards for data usage. Clear documentation and transparency about the data sources and processing methods can further enhance trust and accountability.

References

[1] Daphne Ippolito, Daniel Khashabi, Sewon Min, Chlo'e Buckingham, and Christopher Callison-Burch. Towards a better benchmark for multi-hop question answering. *arXiv preprint arXiv:2210.07065*, 2022.

[2] Gabriel Stanovsky, Nora A Krizhanovskaya, Haokun Liu, and Samuel R Bowman. Do hallucinated responses matter for the trustworthiness of ai? *arXiv preprint arXiv:2302.07589*, 2023.

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

[4] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Investigating the limitations of transformers with query answering tasks. *arXiv preprint arXiv:2108.04378*, 2021.

[5] Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. Raft: Adapting language model to domain specific rag, 2024.

[6] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2024.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[9] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.

[10] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021.

[11] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[12] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*, 2023.

[13] Xin Cheng, Di Luo, Xiuying Chen, Lemao Liu, Dongyan Zhao, and Rui Yan. Lift yourself up: Retrieval-augmented text generation with self-memory. *Advances in Neural Information Processing Systems*, 36, 2024.

[14] Zhicheng Guo, Sijie Cheng, Yile Wang, Peng Li, and Yang Liu. Prompt-guided retrieval augmentation for non-knowledge-intensive tasks. *arXiv preprint arXiv:2305.17653*, 2023.

[15] Anirudh Goyal, Abram Friesen, Andrea Banino, Theophane Weber, Nan Rosemary Ke, Adria Puigdomenech Badia, Arthur Guez, Mehdi Mirza, Peter C Humphreys, Ksenia Konyushova, et al. Retrieval-augmented reinforcement learning. In *International Conference on Machine Learning*, pages 7740–7765. PMLR, 2022.

[16] David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. *arXiv preprint arXiv:2102.04643*, 2021.

[17] Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. *arXiv preprint arXiv:2305.15294*, 2023.

[18] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*, 2020.

[19] Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. Learning to rank in generative retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8716–8723, 2024.

[20] Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, and Suranga Nanayakkara. Fine-tune the entire rag architecture (including dpr retriever) for question-answering. *arXiv preprint arXiv:2106.11517*, 2021.

[21] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10313–10322, 2019.

[22] Xinyue Chen, Pengyu Gao, Jiangjiang Song, and Xiaoyang Tan. Hiqa: A hierarchical contextual augmentation rag for massive documents qa. *arXiv preprint arXiv:2402.01767*, 2024.

[23] Microsoft. Microsoft phi-2 language model. `https://huggingface.co/microsoft/phi-2`, 2023.

[24] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.

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

[26] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[27] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2019.

[28] Vincent A. Traag, Ludo Waltman, and Nees Jan van Eck. From louvain to leiden: guaranteeing well-connected communities. *CoRR*, abs/1810.08473, 2018.

[29] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[30] Max Grusky. Rogue scores. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1914–1934, 2023.

[31] Dhivya Chandrasekaran and Vijay Mago. Evolution of semantic similarity—a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–37, 2021.

A Appendix

A.1 Embedding Plots Analysis

To gain insights into the performance of our models on global and non-global questions, we analyzed the embedding space visualizations 8.

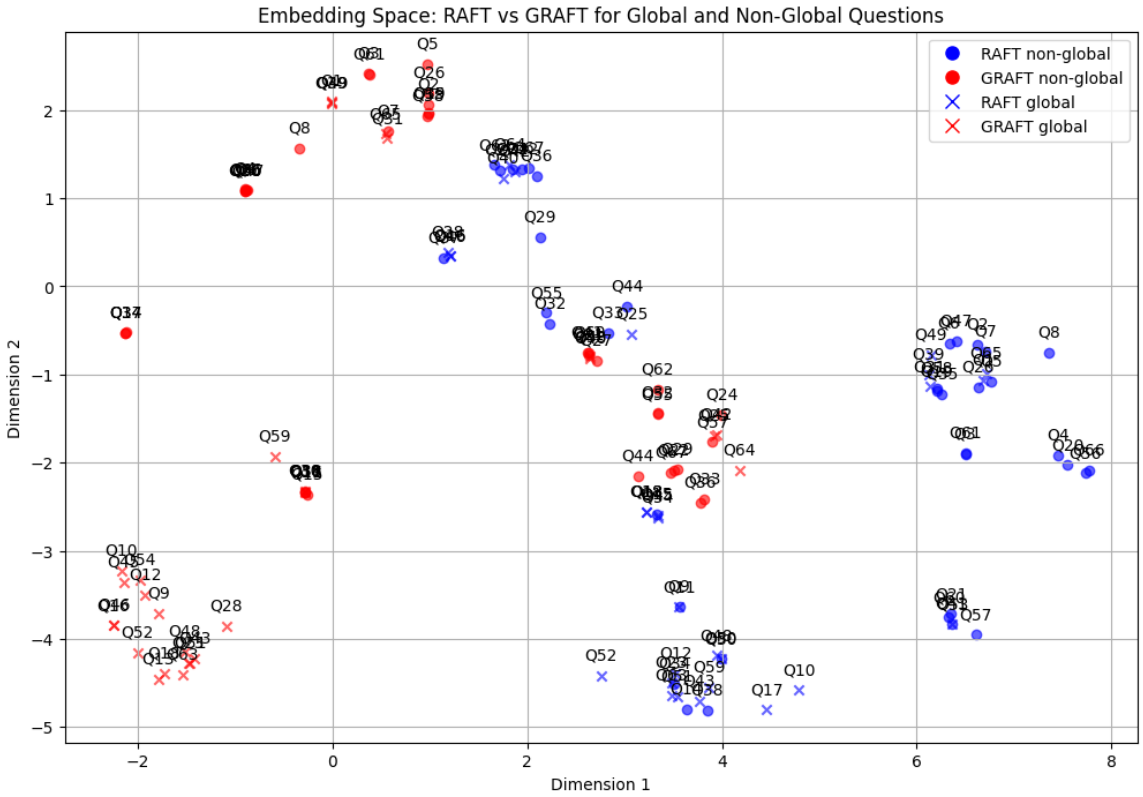

Figure 7: Embedding Plot with RAFT and GRAFT answer embeddings

The embedding plot in Figure 8 provides a visual representation of the embeddings generated by the RAFT and GRAFT models for global and non-global questions. One notable observation from the embedding plot is the presence of a distinct cluster of red crosses (representing GRAFT embeddings for non-global questions) in the lower-left quadrant. This cluster appears to be relatively compact and well-separated from the other embeddings, suggesting that the GRAFT model generates embeddings with similar characteristics for non-global questions. In contrast, the blue crosses (RAFT embeddings for global questions) are more

dispersed and scattered across the embedding space, indicating a higher degree of diversity in the generated embeddings. The compact clustering of GRAFT embeddings for non-global questions could be attributed to the model's ability to effectively capture and represent the relevant information from the source documents, resulting in more accurate and consistent embeddings. However, for global questions, while the GRAFT model still achieves the highest scores on most metrics, the RAFT embeddings (blue crosses) are more dispersed in the embedding space. This dispersion may indicate that the RAFT model struggles to generate consistent and coherent embeddings for global questions, which require a broader understanding of the entire document corpus.

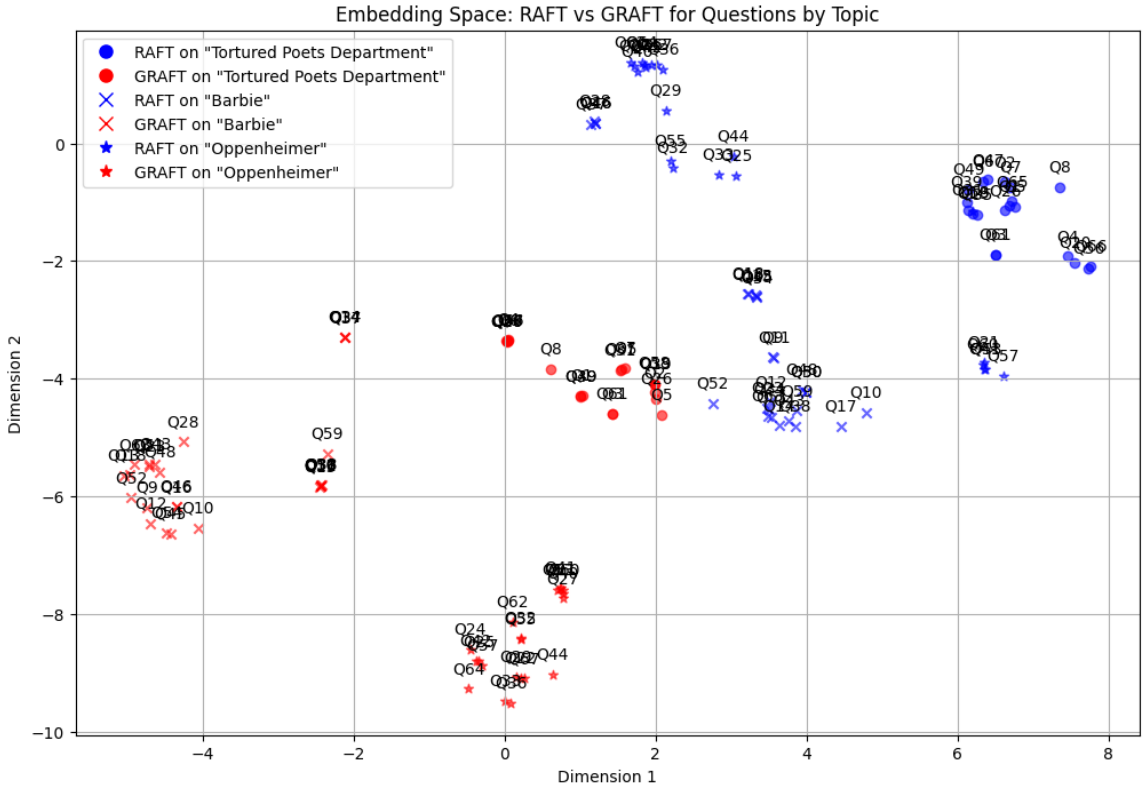

Figure 8: Embedding Plot with RAFT and GRAFT answer embeddings

The embedding plot provides a visual representation of the embeddings generated by the RAFT and GRAFT models for questions across the three different topics: "Tortured Poets Department," "Barbie," and "Oppenheimer." We still observe a distinct cluster of red crosses (representing GRAFT embeddings for "Barbie") in the lower-left quadrant, suggesting GRAFT generates embeddings with similar characteristics for "Barbie." In contrast, there is a higher variation in the representation of the blue crosses (RAFT embeddings for "Barbie"). Therefore, GRAFT is able to outperform RAFT when it comes to the accuracy and consistency of the embeddings. For "Tortured Poets Department," the RAFT embeddings (blue circles) are tightly clustered in the top right, showing consistency, whereas the GRAFT embeddings (red circles) display a similar pattern but with slight variation. This suggests both models are effective for this topic but with some nuanced differences. Regarding "Oppenheimer," the embeddings are more spread out across the embedding space. Both RAFT (blue stars) and GRAFT (red stars) exhibit some clustering but with notable dispersion, especially for RAFT. This dispersion may indicate that the RAFT model struggles to generate consistent and coherent embeddings for "Oppenheimer," which may require a broader understanding of the topic. In summary, while GRAFT demonstrates a tendency to produce more compact and consistent embeddings, RAFT shows a higher degree of dispersion.

A.2 Examples of Global Questions

Topic	Global Questions
Oppenheimer	1. What are the main narrative elements in Oppenheimer? 2. What were the main production techniques used in Oppenheimer?
Barbie	1. How does the Barbie film address controversial themes such as beauty standards and gender roles? 2. How does the Barbie film balance lighthearted and comedic elements with serious and poignant moments?
The Tortured Poets Department	1. What impact did the promotional locations have on The Tortured Poets Department? 2. What are the significant themes in the album The Tortured Poets Department?

Figure 9: Example of Global Questions for Different Topics

Extracted Community 2	Interpreted Summary
{Heartbreak, Billboard, Randy Merrill, South China Morning Post, Mary Siroky, Billboard 200 debut, Commercial success, The Albatross, I Look in People's Windows, Streaming records, Sputnikmusic, Travis Kelce, American Songwriter, Amanda Petrusich, Robin, Invisible String, AnyDecentMusic?, Eras Tour, Financial Times, Rob Sheffield, I Can Fix Him (No Really I Can), Jeffrey Davies, Olivia Horn, NME, Who's Afraid of Little Old Me?, The Grove, Stereogum, X (formerly Twitter), iHeartRadio, Joe Alwyn, Ellie Roberts, Billboard 200, Slate, Paul Bridgewater, Critical reception, Helen Brown, Mark Savage, The Tortured Poets Department, Josh Kurp, But Daddy I Love Him, Peter, NPR, Bryce Dessner, The Alchemy, Sumnima Kandangwa, Vinyl editions, Vox, Clash, Nathan Hubbard, Exclaim!, Nora Princiotti, The An- thology, Chamber pop, Instagram, Phoenix Rising, The Observer, Album release, The Daily Telegraph, My Boy Only Breaks His Favorite Toys, ..., Threads, Lindsay Zoladz, Ephemeral Dreams }	Community 2 focuses on the commercial success and critical reception of Taylor Swift's album, "The Tortured Poets Department." The album achieved significant commercial milestones, debuting atop the Billboard 200 and setting new streaming records on Spotify and Apple Music. Physical formats, including exclusive vinyl editions and limited-edition box sets, contributed to its high sales figures, underscoring its widespread popularity. Media coverage from outlets like Billboard, NME, and The New York Times highlighted the album's critical reception, with mixed reviews that evolved positively over time. The involvement of renowned producers such as Jack Antonoff and Aaron Dessner ensured a high-quality production, enhancing its appeal. Promotional efforts included appearances on major platforms like "The Tonight Show Starring Jimmy Fallon" and "Saturday Night Live," as well as strategic social media campaigns on Instagram and X (formerly Twitter). The album's inclusion in the setlist of Swift's Eras Tour further amplified its reach and cultural impact. Overall, this community encapsulates the album's remarkable blend of commercial success and critical engagement, solidifying its place in contemporary music culture.

Figure 10: Leiden Extracted Community to Generated Summary

A.3 Community Summary from Extracted Community in Analysis Example

Extracted Community	Interpreted Summary
{Barbie, Margot Robbie, Ken, Ryan Gosling, Florence Pugh's Friend Character, Relationships, Mutual Respect, Partnership, Individual Identity, Diversity, Inclusivity, Supportive Friendships, Witty Dialogue, Emotional Authenticity}	The chemistry between the characters in the Barbie film is a key aspect of its storytelling. Margot Robbie and Ryan Gosling, as Barbie and Ken, exhibit a dynamic relationship that evolves throughout the film. Their interactions highlight the themes of mutual respect, partnership, and the search for individual identity within a relationship. The film also features strong supporting characters, such as Florence Pugh's friend character, who provide additional layers to the narrative. These relationships are not only central to the plot but also serve to explore broader social themes, including diversity, inclusivity, and the importance of supportive friendships. The character dynamics are further enriched by the film's witty dialogue and emotional authenticity, making the relationships believable and engaging.

Figure 11: Community Summary from Leiden Extracted Community for "How do the chemistry and relationships in the Barbie film contribute to its narrative?"

A.4 RAFT and GRAFT performance comparison

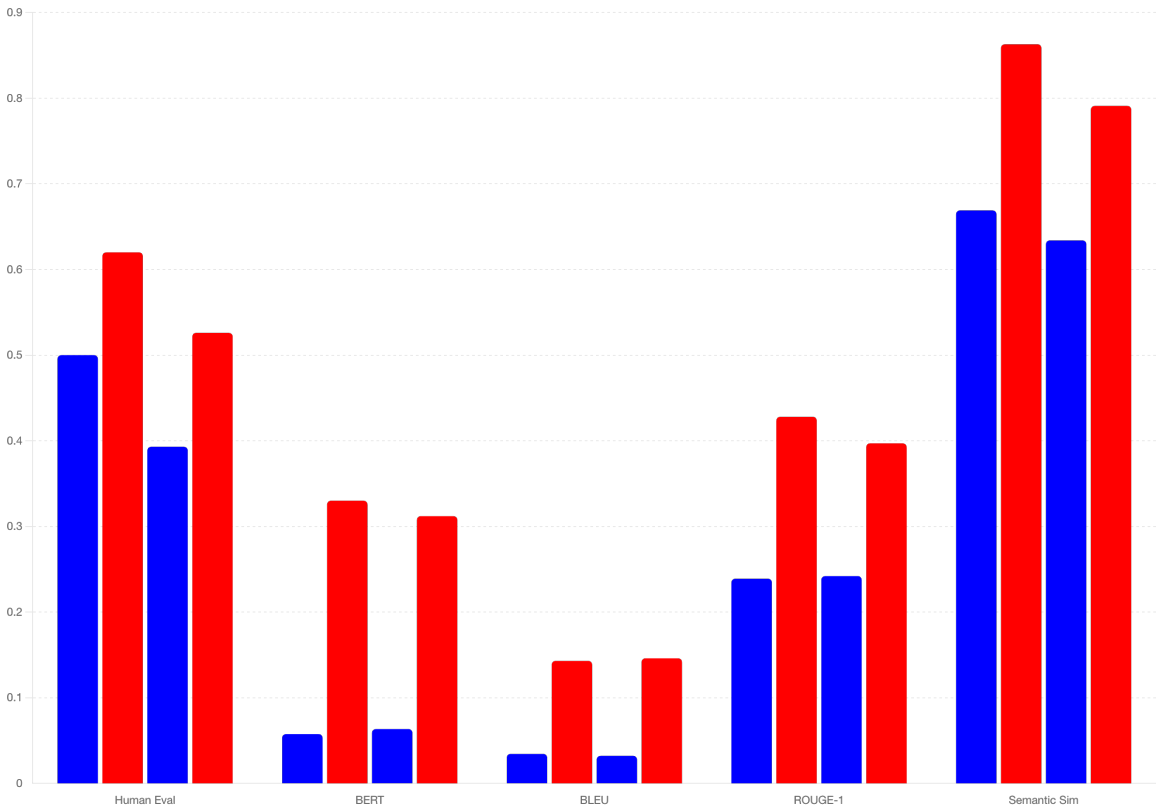

Figure 12: GRAFT outperforms RAFT on all metrics