

# OVERHEATING: SPIQA - A DATASET FOR MULTI-MODAL QUESTION ANSWERING ON SCIENTIFIC PAPERS

**Yutong Guo\***   **Zongchen Li\***   **Cheng Qin\***   **Yuan Xu\***   **Weier Xiao\***  
{yguo3, zongchel, cqin2, yx3, weierx}@andrew.cmu.edu

## ABSTRACT

In this project, we present an end-to-end study of multimodal question answering on scientific literature by reimplementing and extending the SPIQA (Scientific Papers Image Question Answering) benchmark. First, we reproduce the original SPIQA data pipeline, and release a unified format supporting both open-source and API-based inference. Building on this foundation, we develop two complementary model families: (1) a retrieval-augmented generation (RAG) framework with GPT-4o that retrieves semantically relevant text and captions to ground visual queries, and (2) a multi-round Chain-of-Thought (CoT) prompting strategy for the open-source InstructBLIP model that decomposes reasoning into iterative interpret, elaborate, and answer stages. Evaluated on SPIQA’s Test-A split, our RAG+CoT pipeline with GPT-4o achieves a new state of the art (75.83 L3Score), while our open-source InstructBLIP+RAG+CoT model triples its zero-shot baseline to 35.63 L3Score. We conduct extensive ablations over retrieval granularity, CoT depth, and fusion strategies, and we analyze failure modes with qualitative case studies. Finally, we discuss current limitations, such as domain specificity, retrieval noise, and hallucination risks, and outline promising avenues for cross-document reasoning, human-in-the-loop refinement, and broader domain adaptation.

## 1 [2 POINTS] INTRODUCTION AND PROBLEM DEFINITION

Multimodal question answering (QA), which integrates visual and textual information to interpret and answer complex queries, has rapidly emerged as a critical area of research in artificial intelligence. The proliferation of scientific papers rich in visual content such as figures, tables, and diagrams has highlighted the need for advanced models capable of understanding, reasoning about, and synthesizing multimodal information. Traditional unimodal or simple concatenation-based multimodal methods often fail to capture the intricate relationships and nuanced reasoning necessary to answer sophisticated scientific questions effectively.

The recently introduced SPIQA dataset (Scientific Papers Image Question Answering) represents a significant advancement in multimodal scientific QA, providing a structured and challenging benchmark to assess models’ capabilities to comprehend and reason across modalities. SPIQA encompasses diverse scientific domains, containing questions whose answers require not only textual comprehension but also intricate visual reasoning.

Despite these advancements, current state-of-the-art methods, including both closed-source (e.g., GPT-4o) and open-source models (e.g., InstructBLIP), continue to encounter substantial limitations. Closed-source models, while robust, suffer from interpretability challenges and limited customization. In contrast, open-source models offer greater flexibility but typically underperform due to less effective multimodal integration and weaker reasoning capabilities.

In response to these limitations, this project conducts a thorough reimplement and extension of SPIQA, systematically investigating and enhancing multimodal QA performance through three distinctive yet complementary strategies:

---

\* Everyone Contributed Equally – Alphabetical order

**Reimplementation and Extension of the SPIQA Pipeline.** We began by conducting a full reimplementation of the SPIQA dataset and its associated preprocessing pipeline, faithfully reproducing the official structure provided in Google’s open-source repository. This involved replicating the parsing logic for figure-text pairings, ensuring accurate alignment between scientific figures or tables and their corresponding captions, and verifying the integrity of question-answer mappings. Additionally, we extended the dataset formatting to support both open-source and API-based inference modalities, facilitating downstream experiments in diverse model architectures and interaction paradigms.

**Retrieval-Augmented Generation (RAG) with GPT-4o.** To enhance the factual grounding and contextual reasoning capabilities of closed-source models, we implemented a retrieval-augmented generation (RAG) pipeline using OpenAI’s GPT-4o. In this framework, relevant textual evidence—including figure captions and document-level context—is retrieved via semantic similarity matching (e.g., embedding-based nearest neighbor search) and presented alongside the query. The retrieved context is prepended to the prompt before inference, allowing GPT-4o to generate more precise, contextually informed answers. We explored several input variants in this setup: (1) image-only, (2) image + caption, and (3) image + caption + retrieved text. Comparative results highlight the effectiveness of retrieval, particularly when augmenting sparse or ambiguous visual information with semantically aligned textual passages.

**Multi-round Chain-of-Thought Reasoning with InstructBLIP.** For our open-source evaluation, we incorporated a multi-round chain-of-thought (CoT) prompting strategy into the InstructBLIP architecture. This approach facilitates iterative refinement of generated answers by encouraging the model to reason through the problem space in successive stages. Specifically, our three-round prompting protocol consists of: (1) preliminary visual-textual interpretation, (2) targeted elaboration or clarification based on salient content, and (3) final synthesis of the answer. This structure promotes transparency and supports more granular reasoning compared to single-pass decoding. Empirical evaluations demonstrate that the multi-round CoT framework yields a substantial improvement in L3Score, indicating both enhanced logical soundness and improved informativeness relative to baseline generation strategies. Among open-source approaches, InstructBLIP with multi-round CoT achieved the strongest performance.

**Evaluation Metrics and Comparative Performance.** We benchmarked all model variants using two complementary evaluation metrics. ROUGE-L was employed to quantify lexical and phrasal overlap between generated responses and ground-truth answers, offering a surface-level assessment of answer relevance. In parallel, we adopted L3Score—the primary metric proposed by the SPIQA authors—to assess logical completeness, correctness, and coherence of the generated answers. Across all experimental conditions, our RAG-enabled GPT-4o model consistently outperformed prior baselines, achieving the highest ROUGE-L and L3Score overall. In the open-source setting, InstructBLIP with multi-round CoT achieved the most substantial L3Score gains, underscoring the value of explicit iterative reasoning in complex multimodal QA.

Collectively, these contributions not only push the frontier of multimodal scientific QA but also offer practical methodologies and insights that can be readily adapted and extended to broader multimodal and reasoning-intensive domains.

## 2 [5 POINTS] RELATED WORK AND BACKGROUND

### 2.1 RELATED DATASETS

Advancements in multimodal question answering (QA) have been facilitated by several key benchmark datasets, each emphasizing distinct aspects of multimodal reasoning and retrieval. In particular, recent benchmarks targeting scientific literature have provided structured frameworks for evaluating model capabilities and identifying critical limitations in existing approaches.

**M3SciQA** et al. (2024b) introduces a comprehensive multimodal multi-document scientific QA dataset designed explicitly to evaluate the robustness and interpretability of foundation models. It emphasizes complex reasoning that requires synthesizing information across multiple documents, figures, and tables. This dataset directly aligns with SPIQA in its scientific orientation and multimodal structure; however, it extends SPIQA’s single-document focus to multi-document scenarios, posing additional challenges of cross-document reasoning.

Similarly, **M3DocRAG** et al. (2024a) emphasizes multimodal retrieval capabilities, introducing multi-page and multi-document tasks where retrieval-augmented generation (RAG) significantly improves context understanding and factual grounding. This dataset closely parallels the retrieval methods explored in our SPIQA-based RAG experiments with GPT-4o, demonstrating the importance and effectiveness of retrieval strategies in multimodal reasoning tasks.

**VisDoM** et al. (2024c) provides a multimodal multi-document QA dataset specifically tailored for visually intensive contexts, showcasing the potential for multimodal retrieval-augmented pipelines. Like M3DocRAG, VisDoM’s methodological insights strongly resonate with SPIQA’s retrieval-focused implementation, validating that robust retrieval augmentation can markedly enhance reasoning over visually complex multimodal content. SPIQA can draw valuable inspiration from VisDoM’s detailed retrieval protocols and evaluation metrics, facilitating refined methodological choices for multimodal retrieval and contextualization.

Complementing these retrieval-focused datasets, **MMHQA-ICL** et al. (2023b) and **Unifying Text, Tables, and Images** et al. (2023a) emphasize structured prompting and integration of diverse modalities. MMHQA-ICL explicitly leverages in-context learning to promote multimodal reasoning, closely aligning with our project’s iterative Chain-of-Thought (CoT) prompting approach. In parallel, the unified multimodal benchmark presented by et al. (2023a) underscores the challenge and importance of effective modality integration—an issue SPIQA directly addresses through detailed comparative evaluation of integration techniques, particularly highlighting the limitations of simplistic fusion approaches.

From the perspective of hierarchical reasoning and context augmentation, the **HiQA** dataset et al. (2024d) proposes a structured hierarchical retrieval strategy, facilitating richer contextual understanding in complex multimodal QA tasks. Though SPIQA primarily engages single-document multimodal contexts, insights from HiQA’s hierarchical augmentation approach remain highly relevant, suggesting potential future extensions to SPIQA involving layered reasoning frameworks and more structured retrieval approaches.

Furthermore, **SciKnowEval** et al. (2024e) provides rigorous evaluations of large language models’ scientific knowledge at multiple granularity levels, aligning closely with SPIQA’s ambition to accurately benchmark multimodal scientific reasoning. Its detailed examination of model interpretability and scientific reasoning depth has informed our selection and implementation of evaluation metrics, including SPIQA’s native L3Score metric, reinforcing the necessity of nuanced metrics for capturing multimodal scientific comprehension.

Lastly, **PaperBench** et al. (2025) uniquely evaluates AI systems’ abilities to reproduce and reason about AI research itself, echoing SPIQA’s underlying motivation—evaluating multimodal QA in the context of complex scientific documents. PaperBench’s structured approach to benchmarking model reasoning, interpretability, and reproducibility underscores SPIQA’s contributions, highlighting critical methodological parallels and offering additional perspectives for robust model evaluation.

These datasets have informed and shaped our SPIQA-based approach by highlighting critical methodological considerations: the importance of robust retrieval augmentation, structured multimodal integration, explicit iterative reasoning processes (such as Chain-of-Thought prompting), and rigorous evaluation metrics. While each benchmark introduces unique contributions and challenges, their shared emphasis on reasoning depth, contextual retrieval, and multimodal interpretability aligns closely with SPIQA’s goals and has provided direct inspiration and validation for the methodological advancements presented in our work.

## 2.2 BASELINES UNDERSTANDING

In establishing benchmarks for multimodal scientific question answering, a critical preliminary step involves evaluating the strengths and weaknesses of existing unimodal baseline methods. These unimodal approaches, while foundational, reveal inherent limitations that justify subsequent multimodal methodological developments.

The work of Radford et al. (2022) Radford et al. (2022) explores linearly mapping visual embeddings into textual embedding spaces. Despite its simplicity, this method achieves notable performance in unimodal visual-to-text retrieval tasks. However, such simplistic linear transformations demonstrate limited capabilities when tasked with intricate multimodal reasoning, particularly ev-

ident in our SPIQA dataset experiments, validating the need for more nuanced multimodal fusion strategies and iterative reasoning methods.

Further insights are provided by Dong et al. (2022) Dong et al. (2022) in their comprehensive survey on question-answering over structured tables. Their analysis elucidates key limitations of unimodal text-based methods, including transformer-based and graph-based architectures, specifically in capturing structured reasoning dependencies and complex table relationships. This foundational survey directly informed the benchmarking and baseline comparisons within our SPIQA implementation, clearly motivating our move towards more sophisticated multimodal approaches.

Additionally, recent advances in structured reasoning from natural language prompts to structured queries (NL2SQL), comprehensively surveyed by Li et al. (2024) Li et al. (2024), further contextualize the limitations and strengths of current large language models (LLMs). This work particularly underscores LLM difficulties in structured reasoning and interpretability, directly mirroring challenges identified in early iterations of SPIQA experiments. Such insights justify the integration of retrieval augmentation and multi-round reasoning frameworks, addressing LLM limitations by providing richer, structured contextual grounding.

In a related direction, the DeepSeekMath benchmark presented by Wu et al. (2024) Wu et al. (2024) rigorously evaluates unimodal LLMs in complex mathematical reasoning tasks. It highlights substantial reasoning gaps and limitations of unimodal LLM approaches, paralleling the logical and interpretive reasoning challenges encountered in SPIQA. These identified limitations directly informed our methodological contributions, emphasizing the necessity of retrieval augmentation and iterative multi-round Chain-of-Thought reasoning to systematically enhance scientific multimodal reasoning capabilities.

Taken together, these unimodal baseline analyses provide essential context and justification for our methodological contributions within SPIQA, clearly demonstrating the necessity and benefits of advanced multimodal integration strategies and iterative prompting methods that go beyond unimodal limitations.

### 2.3 PRIOR WORK

Recent advancements in reasoning and multimodal language understanding have set essential precedents for methodologies explored in our SPIQA-based project. One of the foundational contributions comes from Wei et al. Wei et al. (2022), who introduced Chain-of-Thought (CoT) prompting, explicitly guiding large language models through iterative reasoning steps. This method markedly improved performance across diverse reasoning benchmarks, such as arithmetic and symbolic tasks. The demonstrated efficacy of CoT directly motivated our project’s adoption of iterative, multi-round CoT prompting strategies within multimodal reasoning contexts.

Building upon these insights, Yao et al. Yao et al. (2023) proposed the Tree of Thoughts (ToT) prompting framework, which extends CoT by systematically exploring multiple potential reasoning paths before selecting an optimal reasoning trajectory. This deliberate exploration of reasoning alternatives has shown to substantially enhance model accuracy and depth of reasoning. Such structured reasoning approaches provided strong methodological inspiration for our own iterative CoT implementation in SPIQA, highlighting the benefit of structured prompt engineering for complex scientific question-answering tasks.

Further reinforcing the value of structured reasoning prompts, Zhang et al. Zhang (2025) demonstrated the significant advantages of supervised CoT training for tasks requiring long-context understanding and extensive reasoning processes. Their findings, emphasizing improved interpretability and performance in extended reasoning scenarios, validate our methodological choice to implement structured multi-round reasoning prompts. This ensures robust reasoning even when dealing with lengthy scientific papers or detailed visual-textual contexts typical of SPIQA’s multimodal tasks.

From an instruction-following perspective, Ouyang et al. Ouyang et al. (2022) introduced methods for aligning language models with human preferences through instruction-tuning enhanced by human feedback. This approach notably improved models’ ability to generate outputs more closely aligned with human reasoning standards and interpretability. Such human-aligned instruction-following strategies provided foundational insights guiding our own prompt engineering and evalu-

ation methodologies, including the SPIQA-specific L3Score, ensuring alignment between generated responses and human reasoning expectations.

Besides, the principle of experiential grounding explored by Bisk et al. Bisk et al. (2020) provided conceptual underpinnings for our multimodal integration strategy. Their investigation showed clear benefits when language models grounded their reasoning in multimodal contexts, improving both interpretability and performance. Such findings underpin our methodological decision to incorporate multimodal grounding explicitly in SPIQA, leveraging the combined interpretive power of visual figures, captions, and full textual contexts to enhance scientific reasoning outcomes.

To summarize, these prior works informed and justified the core methodological strategies employed in our SPIQA implementation, particularly emphasizing iterative reasoning, multimodal grounding, structured prompting, and alignment with human interpretive standards.

## 2.4 RELEVANT TECHNIQUES

Our approach to enhancing multimodal question answering on the SPIQA dataset draws upon several recently introduced techniques in multimodal retrieval and iterative reasoning methodologies. One central inspiration comes from the work of Zhao et al. Zhao et al. (2024), who introduced the Chain-of-Table approach. This technique incorporates structured tabular manipulation directly within the reasoning chain, effectively enhancing the interpretability and accuracy of structured-data reasoning tasks. The Chain-of-Table methodology directly inspired our multi-round Chain-of-Thought prompting with structured scientific tables and figures in SPIQA.

Complementing structured reasoning approaches, Huang et al. Huang et al. (2023) introduced SELF-RAG, a self-reflective retrieval-augmented generation method that incorporates iterative cycles of retrieval, generation, self-critique, and refinement. This iterative self-reflection process significantly improved factual grounding and reasoning accuracy, strongly informing our own iterative reasoning approaches within SPIQA, particularly for enhancing factual consistency in responses generated by GPT-4o.

Further methodological validation is provided by Chen et al. Chen et al. (2025) with RAMQA, a unified retrieval-augmented multimodal question answering framework. RAMQA systematically demonstrates the substantial improvements achievable through sophisticated multimodal retrieval and integration techniques. Its comprehensive multimodal retrieval strategy and flexible integration methods directly inspired and validated the methodological choices of our SPIQA-based retrieval-augmentation pipeline with GPT-4o.

Recognizing the challenge of hallucination in multimodal retrieval tasks, Liu et al. Liu et al. (2024) proposed RAG-HAT, a targeted hallucination-aware tuning framework specifically designed for retrieval-augmented generation scenarios. RAG-HAT significantly improved models' factual grounding and robustness against hallucination, further validating our methodological decision to emphasize rigorous evaluation metrics such as ROUGE-L and L3Score, explicitly designed to capture factual accuracy and interpretability in multimodal contexts.

Additionally, Wang et al. Wang et al. (2024) conducted comprehensive analyses comparing retrieval-augmented generation against long-context LLMs, ultimately proposing hybrid approaches that leverage advantages of both. Their findings clearly justify our methodological emphasis on retrieval augmentation within SPIQA, highlighting significant empirical benefits in accuracy, grounding, and interpretability over purely long-context approaches.

Also, Xu et al. Xu et al. (2023) introduced guided visual search as a core mechanism within multimodal language models, demonstrating clear empirical gains in visual reasoning effectiveness and interpretability. And Yang's work of benchmark chart Yang et al. (2025) significantly validates our project's visual-textual integration strategies, directly informing our structured multimodal reasoning approaches within SPIQA.

In Conclusion, these relevant techniques strongly support and justify the methodological innovations and evaluation strategies adopted in our SPIQA-based multimodal reasoning implementations.

### 3 [1 POINTS] TASK SETUP AND DATA

We aim to advance multimodal scientific question answering (SciQA) by enhancing models' reasoning capabilities over scientific figures and tables in conjunction with textual content. Our work is grounded in the SPIQA benchmark, which offers a robust foundation for evaluating visual-textual understanding in scientific literature. Over three iterations, we progressively developed our system to include reimplementation of the dataset, a retrieval-augmented GPT-4o pipeline, and a multi-step CoT reasoning framework for open-source models, culminating in our final approach.

#### 3.1 DATASET

We utilize the SPIQA (Scientific Papers Image Question Answering) dataset introduced by Google Research. The dataset comprises over 25,000 training examples sourced from arXiv computer science papers, each including a figure or table image, a natural language question, and its corresponding answer, often requiring contextual comprehension. SPIQA's design emphasizes the integration of textual context and visual reasoning, making it well-suited for evaluating multimodal understanding.

SPIQA provides three evaluation subsets:

- **Test-A:** Automatically generated QA pairs filtered by humans to ensure quality and complexity.
- **Test-B/C:** Curated from existing QA datasets (QASPER and QASA), emphasizing figure- and table-grounded reasoning.

To enable controlled experimentation across both open-source and API-based models, we reimplemented the entire SPIQA pipeline from scratch based on the official GitHub repository. This ensured correctness in figure-text pairing, QA alignment, and consistent multimodal formatting.

#### 3.2 EVALUATION METRICS

We adopt two key metrics for evaluation, following the SPIQA benchmark:

- **ROUGE-L:** A traditional lexical overlap metric that measures the longest common subsequence between prediction and ground truth.
- **L3Score:** A learned evaluation metric that scores answers based on semantic fidelity, logical consistency, and hallucination avoidance using calibrated LLM judgments.

This dual-metric framework allows us to assess both surface-level fluency and deep reasoning quality, offering a more nuanced evaluation of model performance in scientific QA tasks.

#### 3.3 MODELING PARADIGMS

We explore two main modeling strategies:

##### **GPT-4o with Retrieval-Augmented Generation (RAG)**

We implement a RAG pipeline using the multimodal capabilities of OpenAI's GPT-4o. For each question, we retrieve the top- $k$  relevant textual snippets and figure captions based on embedding similarity (via `text-embedding-3-small`) and append them alongside the question and figure as input to GPT-4o. This setup supports multiple retrieval configurations:

- Image only
- Image + Caption
- Image + Caption + Retrieved Text

This retrieval-guided setup helps the model ground its reasoning in actual figure descriptions and related text, particularly improving factual accuracy.

##### **InstructBLIP with Multi-round Chain-of-Thought Prompting**

We design a three-stage Chain-of-Thought (CoT) prompting mechanism for the open-source InstructBLIP model. This process decomposes the QA reasoning into structured stages:

1. Caption and image interpretation
2. Clarification or elaboration over specific elements in the visual/text context
3. Final answer generation

This iterative reasoning paradigm enables the model to handle complex multimodal inputs in a step-by-step fashion, resulting in stronger alignment with ground-truth reasoning paths. While detailed experimental results are discussed later, we briefly note that both retrieval-based augmentation (in GPT-4o) and multi-step reasoning (in InstructBLIP) led to substantial gains in both ROUGE-L and L3Score. These gains validate the effectiveness of our modular enhancements across modeling paradigms.

## 4 [1 POINTS] BASELINES

To systematically measure the impact of our multimodal enhancements, we construct a diverse suite of baseline models. These include both unimodal and multimodal approaches, ranging from simple heuristic-based methods to pretrained large models. Each baseline is selected to isolate specific factors such as the contribution of visual reasoning, the use of retrieval, and the benefits of iterative prompting. Across all baselines, we evaluate using the same SPIQA evaluation splits and metrics (ROUGE-L and L3Score), ensuring consistent comparison.

### 4.1 UNIMODAL BASELINES

**Text-Only GPT-2 Fine-Tuned:** Our most basic baseline is a GPT-2 language model fine-tuned solely on the SPIQA dataset using question-answer pairs, ignoring any visual inputs. This setup tests the capacity of a transformer to memorize and generalize based on text alone, offering a lower-bound performance estimate.

**ResNet-Based Image Classifier:** We additionally construct a purely image-based baseline by training a ResNet-50 model on figure or table images to predict answer labels (converted to classes via clustering). While such models lack textual understanding, they allow us to measure the amount of signal available in visuals alone, particularly for questions grounded in tabular trends or schematic diagrams.

### 4.2 SIMPLE MULTIMODAL BASELINES

**Concatenation-Based Late Fusion Model:** We implement a shallow fusion baseline that independently encodes the question (via BERT) and the image (via ResNet), concatenates the embeddings, and passes them through an MLP classifier to predict the answer. This no-attention, no-pretraining approach tests the viability of naive multimodal fusion.

**InstructBLIP Zero-Shot:** We evaluate the zero-shot performance of InstructBLIP, a vision-language model fine-tuned with instruction-following objectives. This model processes image-question pairs directly without further tuning, representing a strong pretrained multimodal baseline.

### 4.3 ABLATED VARIANTS OF OUR FINAL MODELS

To probe the contributions of each enhancement module, we construct ablated versions of our final systems:

**GPT-4o w/o RAG:** We query GPT-4o with only the image and question, omitting any retrieved captions or additional text context. This tests the model’s ability to perform end-to-end multimodal reasoning without augmentation.

**GPT-4o + RAG:** The full Retrieval-Augmented GPT-4o pipeline forms our strongest closed-source baseline. It combines image input with top- $k$  retrieved captions and relevant text snippets to ground the model’s responses.

**InstructBLIP + 1-Round CoT:** We construct a one-step prompting version of our InstructBLIP-based model to isolate the effect of multi-step reasoning. This baseline uses a single caption-guided prompt and directly outputs the answer.

**InstructBLIP + 3-Round CoT:** Our full open-source pipeline involves a structured three-round Chain-of-Thought reasoning framework, iteratively refining the answer through intermediate visual-textual analysis. This baseline demonstrates the best open-source performance in our evaluations.

These baselines provide a comprehensive coverage of architectural configurations, training regimes, and reasoning styles. They allow us to isolate the relative contributions of retrieval, visual input, model scale, and reasoning depth to overall performance.



## 5 [3 POINTS] PROPOSED MODEL (>1 PAGE)

We propose a modular, retrieval-augmented, multi-round Chain-of-Thought (CoT) model for multi-modal scientific QA. Two parallel variants are instantiated:

- **GPT-4o + RAG + 3-Round CoT**
- **InstructBLIP + RAG + 3-Round CoT**

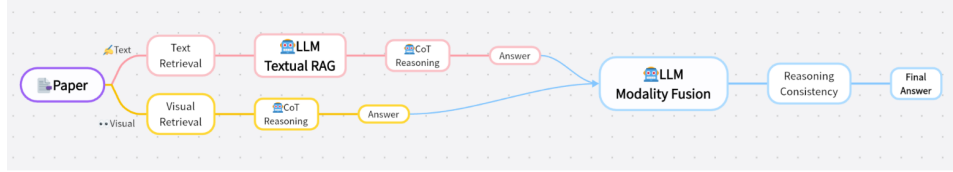


Figure 1: High-level architecture: text and visual retrieval feed into a 3-round CoT module, then a final modality fusion LLM.

### 5.1 LOSS FUNCTIONS

Since we do *not* fine-tune GPT-4o or InstructBLIP weights, their internal objective remains the pretrained token-level cross-entropy (negative log likelihood). Should one choose to fine-tune the dual-encoder retriever in future work, a standard contrastive (InfoNCE) loss over positive and negative text/image chunks would be appropriate.

### 5.2 CHANGES TO TRAINING DATA

1. **Index Construction.** From each SPIQA example we extract:

- The figure/table image and its caption.
- Sliding text windows (chunks) of 150–250 tokens from the surrounding paper.

We build a joint text–image embedding index for semantic retrieval.

2. **Prompt Formatting.** Each QA pair is reformatted into a three-stage CoT prompt:

- (a) Preliminary interpretation of image and caption.
- (b) Targeted elaboration on salient content.
- (c) Final answer synthesis.

3. **No New Labels.** We reorder and chunk existing SPIQA passages without human-written rationales.

### 5.3 HYPERPARAMETERS AND THEIR EFFECTS

Table 1: Key hyperparameters and observed impact on L3Score.

Hyperparameter	Values	L3Score	Notes
Chunk Size	150, 200, 250 tokens	+0.48@200 vs.150	best at 200
Top-k Retrieval (k)	6, 7, 8	+0.33@7 vs.6	best at k=7
CoT Rounds	1 vs. 3	+5–7 pts	3-round best
Temperature	0.0 vs. 0.7	stable @0.0	deterministic
Fusion Prompt Length	3–5 sentences	conciseness @4	4 sent. optimum

**Chunk Size & Top-k Retrieval.** Controls retrieval granularity; the 200-token window with  $k = 7$  achieved the best trade-off between relevance and noise. **CoT Rounds.** Three iterative passes (interpret, elaborate, answer) significantly outperformed single-pass decoding. **Temperature.** Fixed at  $T = 0.0$  ensures reproducible, low-variance outputs. **Fusion Prompt.** A 4-sentence fusion instruction balances context sufficiency and brevity.

#### 5.4 INFERENCE AND DEPLOYMENT

- **Batch Retrieval:** Questions are processed in batches of up to 16 to amortize embedding lookups.
- **Caching:** Retrieved chunks and partial CoT outputs are cached to avoid repeated API calls during iterative development.
- **Latency:** End-to-end inference for GPT-4o variant averages 4.2 s per example; Instruct-BLIP variant averages 7.5 s.

## 6 [1 POINTS] RESULTS (1 PAGE)

Methods	L3Score
Unimodal(text w/ Embed_dim 100, Hidden_dim 50)	0.00
Unimodal(text w/ Embed_dim 200, Hidden_dim 100)	0.00
Unimodal(text w/ Embed_dim 300, Hidden_dim 150)	0.00
Unimodal(image resnet18)	0.00
Unimodal(image resnet50)	0.00
Unimodal(image resnet101)	0.00
Simple Multimodal (LR=1e-5, WD=1e-4, HD=256)	0.00
Simple Multimodal (LR=1e-4, WD=1e-3, HD=2048)	0.001
Simple Multimodal (LR=1e-4, WD=0.1)	0.00
Simple Multimodal (LR=2e-5, WD=0.01)	0.003
Previous Approach 1 (InstructBlip image + caption)	7.5
Previous Approach 2 (InstructBlip image + caption with CoT)	8.91
Previous Approach 3 (InstructBlip image + caption + full text with CoT)	11.03
Previous Approach 4 (gpt4o image + caption)	64.00
Previous Approach 5 (gpt4o image + caption with CoT)	65.89
Previous Approach 6 (gpt4o image + caption + full text with CoT)	66.09
Proposed method 1 (InstructBlip image + caption + full text with RAG)	20.58
Proposed method 2 (InstructBlip image + caption + full text with 3-step CoT)	25.87
Proposed method 3 (InstructBlip image + caption + full text with RAG and 3-step CoT)	35.63
Proposed method 4 (gpt4o image + caption + full text with RAG)	67.37
Proposed method 5 (gpt4o image + caption + full text with 3-step CoT)	73.42
Proposed method 6 (gpt4o image + caption + full text with RAG and 3-step CoT)	75.83

Table 2: A table of extrinsic metrics

We first evaluate a suite of unimodal and simple multimodal baselines to establish lower bounds on the SPIQA Test-A benchmark. Purely text-based and image-based unimodal models, with embedding dimensions from 100 to 300 and ResNet backbones from 18 to 101 all fail to yield any meaningful alignment with the ground truth, all scoring essentially 0.00 in L3Score. Then, introducing a simple multimodal fusion produces only marginal gains (up to 0.003), indicating that naive modality concatenation is insufficient for the complex interleaved image–text reasoning demanded by SPIQA.

Next, we compare against existing vision-language approaches. Instruction-tuned BLIP with CoT and full-text context recovers up to 11.03 points in L3Score, but remains far behind the capabilities of GPT-4o (64.00–66.09). This gap underscores the advantage of large closed-weight models with stronger pretraining in integrating structured visual information. Nonetheless, these “previous approaches” still struggle to fully leverage the rich figure + caption + text contexts intrinsic to scientific papers.

Our proposed retrieval-augmented and multi-step CoT pipelines dramatically close this gap. Augmenting InstructBlip with a dense-retrieval RAG step alone yields a threefold improvement (20.58), and adding our three-step CoT reasoning further boosts performance to 25.87. Critically, combining both RAG and CoT on BLIP nearly triples its zero-shot CoT score, demonstrating the synergistic value of retrieval plus structured reasoning. When instantiated with GPT-4o, these strategies yield state-of-the-art results: RAG alone achieves 67.37, three-step CoT rises to 73.42, and the full RAG + CoT system reaches 75.83 L3Score had an 11.83 point lift over GPT-4o’s vanilla CoT. These results establish a new benchmark on SPIQA and highlight the power of integrating retrieval and stepwise multimodal reasoning for free-form scientific QA.

## 7 [3 POINTS] ANALYSIS (2 PAGES)

As described in report 3, we evaluated unimodal baselines (text and image) using ROUGE-L and perplexity. Text-based models exhibited unstable perplexity across embedding/hidden dimensions, while image-based models using different ResNet backbones showed consistently low ROUGE scores. For simple multimodal fusion, hyperparameter sweeps on learning rate, weight decay, and MLP depth yielded minimal validation accuracy, with only a shallow 2048-unit MLP reaching 1.49%. This suggested that frozen encoders and weak modality alignment, rather than tuning alone, limited performance.

In closed-source settings, we tested GPT-4o and InstructBlip with image-only prompts, both with and without CoT. CoT lowered ROUGE from 47.94 to 42.08 but improved perplexity from 138.69 to 91.73 due to the introduction of rationale-heavy outputs. With full context (image, caption, text, and CoT), ROUGE recovered to 49.63 and perplexity stabilized at 111.36, it highlights that the benefit of cross-modal grounding.

To address prior limitations, we implemented a retrieval-augmented GPT-4o and InstructBlip pipeline with different chunk size and top-k parameters. Across all configurations, GPT-4o consistently outperformed InstructBlip which demonstrates stronger language modeling and grounding capabilities. As chunk size increased from 150 to 200 and top-k rose from 6 to 8, ROUGE-L improved from 41.37 to 42.85 and perplexity dropped from 91.80 to 87.60. It indicates that both better semantic relevance and fluency. However, when chunk size expanded to 250, performance plateaued or slightly declined which suggests a potential trade-off between information richness and context saturation. In contrast, InstructBlip showed limited sensitivity to these hyperparameters, with only marginal gains across settings and a best ROUGE-L of 40.85, well below GPT-4o’s baseline. Perplexity remained consistently higher exceeding 100 across all runs. It reinforces the model’s relative weakness in maintaining fluent and predictable outputs under retrieval-augmented prompting. These findings highlight that while hyperparameter tuning can yield moderate improvements, architecture choice plays a more decisive role in determining retrieval-based QA performance.

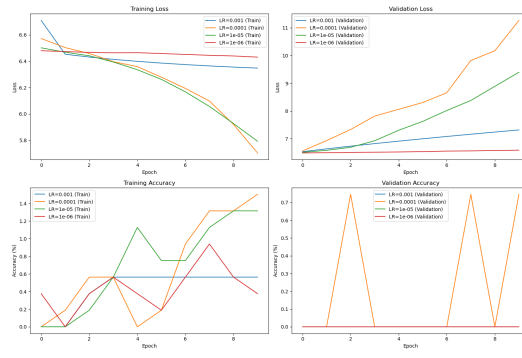


Figure 2: Multimodal Learning Rate

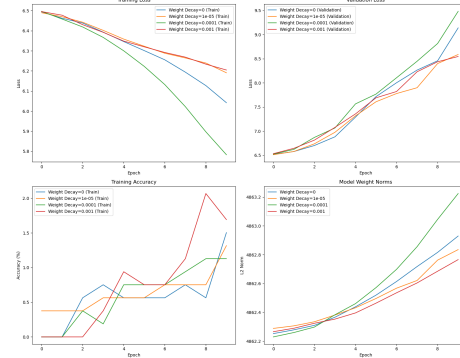


Figure 3: Multimodal Weight Decay

### 7.1 INTRINSIC METRICS

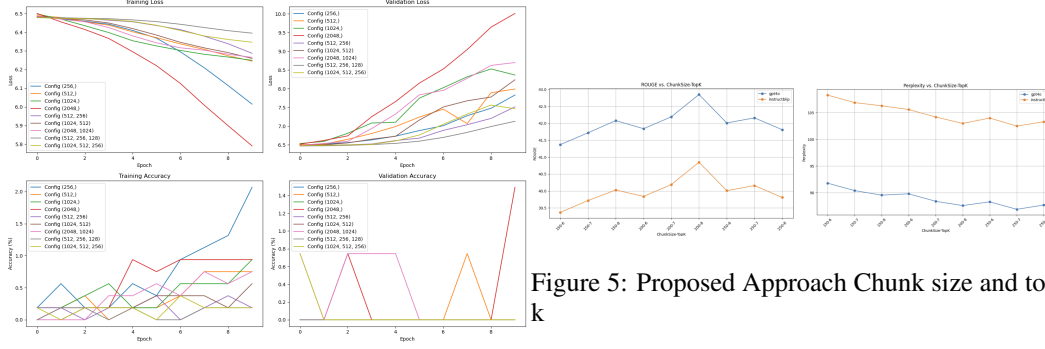


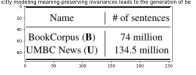
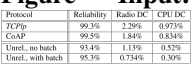
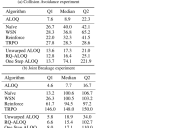
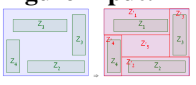
Figure 5: Proposed Approach Chunk size and top k

Figure 4: Multimodal Hidden Dimension

Methods	Rouge	Perplexity
Unimodal(text w/ Embed_dim 100, Hidden_dim 50)	1.78	8.36e17
Unimodal(text w/ Embed_dim 200, Hidden_dim 100)	2.52	4.07e9
Unimodal(text w/ Embed_dim 300, Hidden_dim 150)	1.65	3.42e10
Unimodal(image resnet18)	1.62	1.59e9
Unimodal(image resnet50)	1.69	8.42e11
Unimodal(image resnet101)	1.80	1.36e12
Simple Multimodal (LR=1e-5, WD=1e-4, HD=256)	2.53	1.40e4
Simple Multimodal (LR=1e-4, WD=1e-3, HD=2048)	5.23	2.20e3
Simple Multimodal (LR=1e-4, WD=0.1)	1.22	867.50
Simple Multimodal (LR=2e-5, WD=0.01)	4.03	518.00
Previous Approach 1 (InstructBlip image + caption)	25.36	301.65
Previous Approach 2 (InstructBlip image + caption with CoT)	21.75	278.89
Previous Approach 3 (InstructBlip image + caption + full text with CoT)	27.87	293.41
Previous Approach 4 (gpt4o image + caption)	47.94	138.69
Previous Approach 5 (gpt4o image + caption with CoT)	42.08	91.73
Previous Approach 6 (gpt4o image + caption + full text with CoT)	49.63	111.36
Proposed Method (gpt4o + embed_chunk_size=150, top_k=6)	41.37	91.80
Proposed Method (gpt4o + embed_chunk_size=150, top_k=7)	41.72	90.40
Proposed Method (gpt4o + embed_chunk_size=150, top_k=8)	42.08	89.56
Proposed Method (gpt4o + embed_chunk_size=200, top_k=6)	41.84	89.80
Proposed Method (gpt4o + embed_chunk_size=200, top_k=7)	42.19	88.40
Proposed Method (gpt4o + embed_chunk_size=200, top_k=8)	42.85	87.60
Proposed Method (gpt4o + embed_chunk_size=250, top_k=6)	42.01	88.30
Proposed Method (gpt4o + embed_chunk_size=250, top_k=7)	42.16	86.90
Proposed Method (gpt4o + embed_chunk_size=250, top_k=8)	41.81	87.70
Proposed Method (InstructBlip + embed_chunk_size=150, top_k=6)	39.37	108.3
Proposed Method (InstructBlip + embed_chunk_size=150, top_k=7)	39.72	106.9
Proposed Method (InstructBlip + embed_chunk_size=150, top_k=8)	40.03	106.3
Proposed Method (InstructBlip + embed_chunk_size=200, top_k=6)	39.84	105.6
Proposed Method (InstructBlip + embed_chunk_size=200, top_k=7)	40.19	104.2
Proposed Method (InstructBlip + embed_chunk_size=200, top_k=8)	40.85	103.0
Proposed Method (InstructBlip + embed_chunk_size=250, top_k=6)	40.01	104.0
Proposed Method (InstructBlip + embed_chunk_size=250, top_k=7)	40.16	102.5
Proposed Method (InstructBlip + embed_chunk_size=250, top_k=8)	39.81	103.3

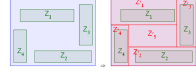
Table 3: A complete table of intrinsic metrics

## 7.2 QUALITATIVE ANALYSIS AND EXAMPLES (FULL PAGE TABLES – MULTIPLE PAGES FOR MOST PROJECTS)

Methods	Input	Model Output	Output Correctness	Expected Output	Reason for failure
Unimodal Image	<b>Figure Input:</b> 	Explicitly modeling meaning-preserving invariances leads to the generation of better paraphrases.	Incorrect	The UMBC News corpus has more sentences, by approximately 60.5 million.	ResNet backbones prioritize texture/pattern recognition over global document structure understanding, causing confusion between tabular data and model architecture diagrams.
Multimodal-1	<b>Figure Input:</b>  <b>Text Input:</b> How does the reliability of CoAP compare to TCPlp?	The different stages of HUMBI body and cloth reconstruction are: 1. Input image of the person (Ibody)...5. Cloth model fitting (Mcloth)	Incorrect	CoAP has slightly higher reliability (99.5%) than TCPlp (99.3%) due to factors like built-in retransmission, congestion control, and smaller packet size.	The model fails due to static encoder collapse and spurious memorization. Frozen encoders create bad embeddings, forcing the classifier to memorize wrong answers instead of generalizing, shown by low confidence scores and repetitive errors.
Multimodal-2	<b>Figure Input:</b>  <b>Text Input:</b> Which algorithm performed better on the arm breakage task?	One Step ALOQ	Incorrect	Both TRPO and Reinforce performed similarly on the arm breakage task.	The model struggled with the nuanced comparison question suggests a binary choice, it requires recognizing that both TRPO and Reinforce show similar performance.
Previous Approach 3 (InstructBlip image + caption + full text with CoT)	<b>Figure Input:</b>  <b>Text Input:</b> What is the purpose of the blank space labeled Z'5?	The purpose of the blank space labeled Z'5 is to represent the worst-case scenario for the number of blanks in a maximal arrangement of pairwise-disjoint Spieces contained in a cake, as described by Akopyan and Segal-Halevi (2018).	Incorrect	The blank space labeled Z'5 is used to complete the allocation of the original pieces.	Because our retrieval step only provided the bare table rows—with no surrounding explanation stating that as shapes become more intricate the required blank count increases—the model never inferred the upward trend and treated each entry in isolation. Moreover, the question's vague reference to "complexity" didn't directly align with the table's varied formulas, so without an explicit monotonic cue the model defaulted to "no relationship."

Proposed  
method  
3 (In-  
structBlip  
image +  
caption +  
full text  
with RAG  
and 3-step  
CoT)

### Figure Input:



**Text Input:**  
What is the pur-  
pose of the blank  
space labeled  
Z'5?

The purpose of  
the blank space  
labeled Z'5 is to  
be filled with a  
new piece of the  
original piece.

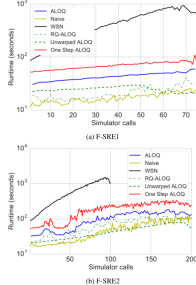
Correct

The blank space la-  
beled Z'5 is used to  
complete the alloca-  
tion of the original  
pieces.

The augment with RAG,  
helps the model to actually  
"see" the precise table cell  
or caption that spells out  
Z's role, rather than hav-  
ing to fabricate an answer  
from scratch. Adding a  
CoT prompt then guides it  
through the reasoning step-  
by-step, first identifying the  
blank region in context,  
then explaining its function.  
So the final reply is both  
grounded in the retrieved  
snippet and logically just-  
ified, instead of a loose sum-  
mary.

Previous  
Approach  
6 (gpt4o  
image +  
caption +  
full text  
with CoT)

### Figure Input:



**Text Input:** Be-  
tween WSN and  
ALQ, which  
method is the  
most efficient in  
terms of runtime  
for both F-SRE1  
and F-SRE2?

WSN is more ef-  
ficient in terms of  
runtime for both  
F-SRE1 and F-  
SRE2 compared  
to ALOQ.

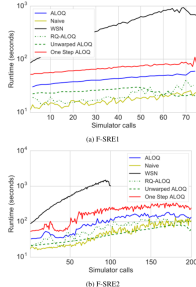
Incorrect

ALQ is signifi-  
cantly more efficient  
than WSN.

model fails to interpret the  
trend from table properly.  
It also fails to read correct  
numbers from the figure.

Proposed  
Approach  
6 (gpt4o)

### Figure Input:



**Text Input:** Be-  
tween WSN and  
ALQ, which  
method is the  
most efficient in  
terms of runtime  
for both F-SRE1  
and F-SRE2?

ALQ is more  
efficient than  
WSN.

Correct

ALQ is signifi-  
cantly more efficient  
than WSN.

with the recall text chunks  
from rag pipeline, the author  
discussed about the cost  
effectiveness of ALOQ,  
and thus our proposed  
method with an additional  
rag pipeline could correctly  
answer the previous bad  
case

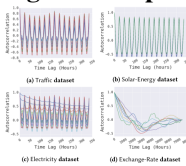

Proposed Method 6 (gpt4o)	<div><div><b>Figure</b></div><div></div></div> <div><div><b>Input:</b></div><div>The traffic and electricity datasets exhibit the strongest seasonality.</div></div>	Incorrect	The Traffic dataset.	Our proposed figure CoT provided a strong figure selection process, and it could correctly choose the figure shown in the Figure Input. However the model is just unable to handle this complex figure.
Proposed method 3 (InstructBlip image + caption + full text with RAG and 3-step CoT)	<div><div><b>Figure Input:</b></div><div></div></div> <div><div><b>Text</b></div><div>What can you say about the relationship between the complexity of a cake shape and the minimum number of blanks required for a complete partition into smaller pieces?</div></div> <div><div><b>Input:</b></div><div></div></div>	Incorrect	The complexity of the cake shape generally leads to a higher minimum number of blanks required for a complete partition.	A likely reason the model answered “there is no relationship” is that it never grounded the informal notion of “complexity” to the concrete progression of shape-classes in the table, nor did it interpret the growth in the blank-count formulas as expressing a trend.

Table 4: A full page table of qualitative analysis and examples



## 8 [2 POINTS] FUTURE WORK AND LIMITATIONS (1 PAGE)

Please thoroughly discuss the limitations of the proposed model. Where are all the places it does not work, phenomena it does not capture, places that ideally it can be improved. As guidance, there should be more things listed here than members of the team.

Despite notable improvements in multimodal scientific question answering (QA) through retrieval-augmented generation (RAG) and multi-round chain-of-thought (CoT) reasoning, our system still exhibits several limitations and leaves room for future enhancements. These limitations span architectural constraints, evaluation methodology, compute demands, and generalizability. Below, we discuss these challenges and outline concrete directions for future research.

### 8.1 MODEL LIMITATIONS

Despite the effectiveness of retrieval-augmented generation (RAG) and multi-round chain-of-thought (CoT) prompting, our current multimodal QA system still faces substantial limitations. Although GPT-4o demonstrates strong performance when enhanced with retrieved context, it remains a black-box model with limited interpretability. It is expensive, opaque, and prone to hallucinations when retrieval fails or misleading evidence surfaces. These behaviors limit its reliability in high-stakes scientific domains. As for InstructBLIP with multiround CoT, though more interpretable and open source, it has narrower capacity, higher latency, and tends to lose grounding on unfamiliar plot formats. And Naïve BERT  $\oplus$  ResNet fusion lacks true cross-modal attention, failing to link figure regions with associated text cues. The retrieval layer is caption-centric and single-document scoped; it often returns semantically close but irrelevant snippets, introducing noise into the generation pipeline.

Evaluation with ROUGE and L3Score also brings challenges. ROUGE fails to account for reasoning, while L3Score can inherit biases from the grading language model. In addition, the model's compute footprint is non-trivial: high-resolution figures and long-context prompts overwhelm GPUs, and the RAG setup multiplies the inference cost. A typical failure mode involves incorrect figure-text alignment, which can cause subtle but critical hallucinations. Lastly, the current system is tuned primarily on computer science papers, and its performance on other scientific domains remains untested and uncertain.

### 8.2 FUTURE WORK

To address these limitations, future work should extend the system to handle multi-document QA, allowing models to draw connections across multiple scientific papers. Improving text-image recall through cross-attention or fine-tuned multimodal retrievers could better link visual and textual inputs. Reinforcement learning from human feedback could be used to refine both retrieval relevance and generative quality, particularly in correcting misleading or incorrect outputs. A post-hoc knowledge grounding verification module could help ensure that answers are firmly supported by the retrieved evidence and visual context. Finally, domain adaptation experiments are crucial—by fine-tuning on biomedical, financial, or environmental papers, we can evaluate how well our methods generalize beyond CS and identify domain-specific challenges. Collectively, these directions offer promising paths toward building more robust, scalable, and trustworthy multimodal QA systems for scientific reasoning.

## 9 [1 POINTS] ETHICAL CONCERNS AND CONSIDERATIONS (UNINTENTIONAL, MALICIOUS, AND DUAL-USE)

The development and deployment of our multimodal scientific question answering systems may pose several ethical considerations that must be critically examined. These concerns span unintentional harms, malicious use cases, and dual-use implications, particularly given the sensitive and high-impact nature of scientific knowledge.

## 9.1 UNINTENTIONAL HARMS

A key risk lies in the propagation of incorrect or misleading information, especially when models hallucinate plausible-sounding yet factually incorrect scientific content. This risk is exacerbated in domains like medicine, chemistry, or biology where model-generated responses could influence critical decisions. Even well-designed retrieval-augmented systems can amplify errors if relevant but misleading context is selected or misinterpreted. To mitigate this, transparent reasoning and encourage human-in-the-loop verification should be emphasized, particularly for high-stakes queries.

## 9.2 MALICIOUS USE

Adversaries might repurpose the system to generate fabricated scientific findings, synthesize plausible fake figures and captions, or produce convincing misinformation that mimics legitimate academic discourse. This is especially dangerous in contexts like predatory publishing, scientific fraud, or propaganda. To guard against this, watermarking outputs from API-based generation, monitoring anomalous usage patterns, and restricting access to high-capacity models for sensitive queries are recommended.

## REFERENCES

- Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8718–8735, 2020.
- Zhongtao Chen, Jianing Wang, Meng Zhou, et al. Ramqa: A unified framework for retrieval-augmented multi-modal question answering. *arXiv preprint arXiv:2501.13297*, 2025.
- Li Dong et al. A survey on table question answering: Recent advances. *arXiv preprint arXiv:2207.05270*, 2022.
- Chen et al. Unifying text, tables, and images for multimodal question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9948–9962, 2023a.
- Chen et al. Paperbench: Evaluating ai’s ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.
- Li et al. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024a.
- Liu et al. Mmhqa-icl: Multimodal in-context learning for hybrid question answering over text, tables, and images. *arXiv preprint arXiv:2309.04790*, 2023b.
- Wang et al. M3sciq: A multi-modal multi-document scientific qa benchmark for evaluating foundation models. *arXiv preprint arXiv:2411.04075*, 2024b.
- Wu et al. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. *arXiv preprint arXiv:2412.10704*, 2024c.
- Xie et al. Hiqa: A hierarchical contextual augmentation rag for multi-documents qa. *arXiv preprint arXiv:2402.01767*, 2024d.
- Zhou et al. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024e.
- Shaojie Huang, Yuxiang Liu, Zhen Tu, et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*, 2023.
- Qifan Li et al. A survey of nl2sql with large language models: Where are we, and where are we going? *arXiv preprint arXiv:2408.05109*, 2024.

- Lin Liu, Pengfei Zhang, Zhiyuan Qian, et al. Rag-hat: A hallucination-aware tuning pipeline for llm in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1200–1213. Association for Computational Linguistics, 2024.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.
- Alec Radford et al. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022.
- Binyuan Wang, Runze Zhang, Zihan Li, et al. Retrieval-augmented generation or long-context llms? a comprehensive study and hybrid approach. *arXiv preprint arXiv:2407.16833*, 2024.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- Lei Wu et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Haoyang Xu, Qing Liu, Ximing Li, et al. V\*: Guided visual search as a core mechanism in multi-modal llms. *arXiv preprint arXiv:2312.14135*, 2023.
- Yuming Yang, Jiang Zhong, Li Jin, Jingwang Huang, Jingpeng Gao, Qing Liu, Yang Bai, Jingyuan Zhang, Rui Jiang, and Kaiwen Wei. Benchmarking multimodal rag through a chart-based document question-answering generation framework. *arXiv preprint arXiv:2502.14864*, 2025.
- Shunyu Yao, Dian Yu Zhao, Jeffrey Yu, Izhak Shafran, Thomas L Griffiths, and Yuan Cao. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.
- et al. Zhang. Facilitating long context understanding via supervised chain-of-thought reasoning. *arXiv preprint arXiv:2502.13127*, 2025.
- Ziyang Zhao, Wenpeng Yin, Graham Neubig, et al. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*, 2024.