STUDY COACH: R1. DATASET PROPOSAL AND ANALYSIS

Dipti Aswath[*] **Buddhika Kahawitage**[*] **Jiashan Cui**[*] **Khubi Srinivasan**[*]
{daswath, bmakehelwala, jiashanc, ksubrama}@andrew.cmu.edu

1 [4 POINTS] PROBLEM DEFINITION AND DATASET CHOICE

Paper: Pramanick et al., “SPIQA: A Dataset for Multimodal Question Answering on Scientific Papers” (NeurIPS 2024 Datasets and Benchmarks Track)

Link: https://huggingface.co/datasets/google/spiqa

Problem Motivation: Senior ML interviews increasingly expect candidates to demonstrate paper literacy: reading a paper, understanding experimental results, and explaining key figures. Users struggle with interpreting architecture diagrams, training curves, ablation tables, and result visualizations. Existing tools currently support text-based interview prep but lack the ability to help users engage with the visual components of research.

Dataset: SPIQA (Scientific Paper Image Question Answering) is a large-scale QA dataset designed to interpret complex figures and tables within scientific research papers across computer science domains. The full dataset contains 25,859 papers, 152,487 figures, 117,707 tables, and 270,194 question-answer pairs from 19 CS conferences (2018-2023).

Our Focus: We use the Test-A split (666 QA pairs across 118 papers), which was manually filtered and curated to ensure high quality. We augment this into **SPIQA+** by adding synthetic user answers and structured model evaluations, transforming the task from answer generation to answer evaluation.

Prior Work Gap: The SPIQA paper established that paper context improves figure QA performance (Gemini 1.5 Pro achieves 8+ point L3Score improvements with full context). However, SPIQA only addresses: *Can models generate correct answers?* It does NOT address: *Can models evaluate whether a human’s explanation is correct?* This answer evaluation capability is essential for educational applications like interview coaching.

1.1 [0.5 POINTS] WHAT PHENOMENA OR TASK DOES THIS DATASET HELP ADDRESS?

This dataset helps address **incongruence detection** between user (textual) explanations and visual content in scientific figures. Specifically, given a figure and a user’s explanation of what it shows, can a multimodal model:

1. Detect when the explanation doesn’t match the figure content (verdict: correct/partially correct/incorrect)
2. Categorize the type of error (factual error vs. wrong conclusion)
3. Provide useful coaching feedback (free-text) that localizes the mismatch

This task directly supports the “Study Coach” application where users practice explaining research figures and receive AI-powered feedback on their interpretations.

1.2 [0.5 POINTS] WHAT ABOUT THIS TASK IS FUNDAMENTALLY MULTIMODAL?

The task requires **cross-modal alignment and verification** that cannot be accomplished with either modality alone:

[*] Everyone Contributed Equally

1. **Visual grounding is essential:** To evaluate "BERT achieves 89% accuracy" as incorrect, the model must read the actual bar height (84.6%) from the figure - this information exists only in the visual modality.
2. **Language understanding is essential:** The user's explanation is textual, and the model must parse claims like "outperforms" or "plateaus" to know what visual evidence to seek.
3. **Cross-modal reasoning required:** Wrong conclusion errors (e.g., "loss decreases throughout training" when it plateaus) require the model to map verbal interpretations onto visual patterns and assess whether the interpretation is justified - neither extracting numbers nor parsing text alone suffices.

This is fundamentally different from VQA (generate text from image) or image captioning—we're evaluating whether *someone else's* text-image alignment is correct.

1.3 Hypotheses

[1 points] **Hypothesis 1 (Detection)** Multimodal models can classify user explanations as correct/partially correct/incorrect above random baseline ($\geq$50% on three-class), but will fall short of human performance.

[1 points] **Hypothesis 2 (Context)** Relevant paper context will improve detection accuracy for reasoning-dependent errors (wrong conclusions, omissions) but not for visually-obvious errors (wrong numbers).

[1 points] **Hypothesis 3 (Localization)** Models will detect incongruence at higher rates than they can accurately localize and explain it.

Hypothesis 4 (Error Type) Factual errors will be detected more reliably than interpretation errors (wrong conclusions, omissions).

Hypothesis 5 (Figure Type) Detection accuracy will vary significantly across figure types, with tables being easiest (explicit, localized information) and architecture diagrams hardest (distributed spatial relationships).

2 [6 POINTS] DATASET ANALYSIS

2.1 [1 POINTS] DATASET PROPERTIES

We use the Scientific Paper Image Question Answering (SPIQA) dataset, publicly available through Hugging Face (`https://huggingface.co/datasets/google/spiqa`).

Full Dataset Statistics:

Metric	Count
Papers	25,859 (from 19 CS conferences, 2018-2023)
Figures	152,487
Tables	117,707
QA pairs	270,194 total
Avg question length	12.98 words
Avg answer length	14.56 words

Dataset Splits:

Split	Papers	QA Source	Notes
Train	∼25,459	LLM-generated	For fine-tuning (stretch goal)
Validation	200	LLM-generated	Dev set
Test-A	118	LLM + manual filter	666 QAs – our primary dataset
Test-B	varies	Human-written (QASA)	Optional validation
Test-C	varies	Human-written (QASPER)	Optional validation

We focus on Test-A because: (1) it is the highest-quality evaluation split with manual filtering, (2) our core study uses prompting-based evaluation (no fine-tuning needed), and (3) 666 QA pairs is sufficient for our augmented SPIQA+ dataset of ∼200-400 evaluation examples.

2.2 [0.5 POINTS] COMPUTE REQUIREMENTS

Token Requirement for Dataset Generation

We augment SPIQA data set with wrong student answers along with corresponding model explanations as to why a given student answer is incorrect. We plan to use in-context learning on the test-A dataset with a seed of wrong answer examples which have been manually curated. This synthetically generated dataset will contain about 400 new examples. We will be using GPT-4o or a similar closed LLM API for this data-set generation. We make the following assumptions for the tokens required for this task.

- Each image cost a fixed amount of tokens (used 85 as per OpenAI documentation[1])
- Synthetically generated model explanation as to why the answer is wrong is more verbose as it needs to refer to both the context and the given answer. So we estimate it to be twice the size of ground truth answer explanation.
- Three exemplars used in each prompt.

Under these assumptions, we will consume the following number of tokens for generating ∼400 synthetic wrong answer, explanation examples.

With typically cited GPT-4o token pricing[2], cost estimate will be as follows:

```
Cost = 1888192 * 0.0000025 + 1317330 * 0.000008 + 85830 * 0.0000025 =
$15.47
```

[1] `https://platform.openai.com/docs/guides/images-vision`

[2] `https://platform.openai.com/docs/pricing`

Component	Tokens
Input Text	1,888,192
Input Image	1,317,330
Output Text	85,830

Token Requirement for Model Testing

For testing the SPIQA+ dataset we plan to use GPT-4o and Gemini 1.5 Pro. Assuming that we run through all 666 examples of test-A, the following table gives the estimate for the required number of tokens (question posed along with image and captions).

Component	Tokens
Input Text	669,399
Input Image	712,300
Output Text	114,629

Cost estimate for GPT-4o will be as follows:

```
Cost = 669399 * 0.0000025 + 712300 * 0.000008 + 114629 * 0.0000025 =
$7.66
```

We expect a cost figure in this range for Gemini 1.5 Pro as well.

Hardware Requirements

We also plan to run inference using LLaVA 1.5 7B open source model. A 7B parameter model typically requires an EC2 instance with at least 16GB–24GB of VRAM for efficient inference, making a single `g5.xlarge` (NVIDIA A10G) or `g5.2xlarge` instance the most cost-effective choice for this project.

2.3 [2 points] Modality Analysis

Using a small sample of the data (e.g. validation splits), we generate statistics and plots for relevant properties of the data.

Property 1: Question Length Distribution

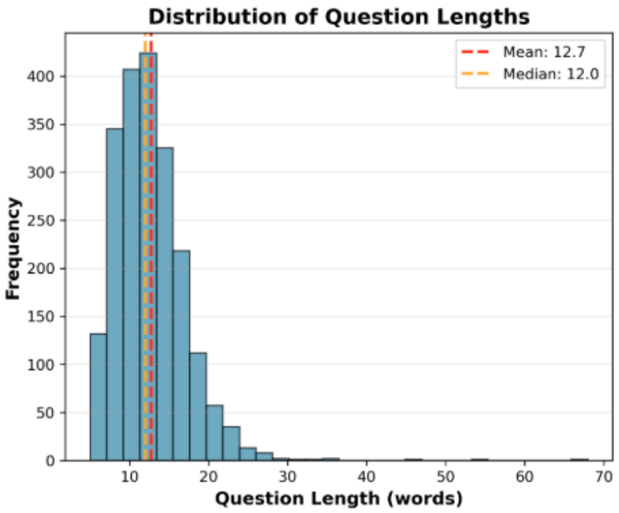

Figure 1: Distribution of question lengths in the validation split.

- Mean: 12.7 words

- Median: 12.0 words
- Range: 5–68 words

Questions are fairly concise, with most clustering around 10–15 words. The tight distribution suggests consistent question formatting.

Property 2: Answer Length Distribution

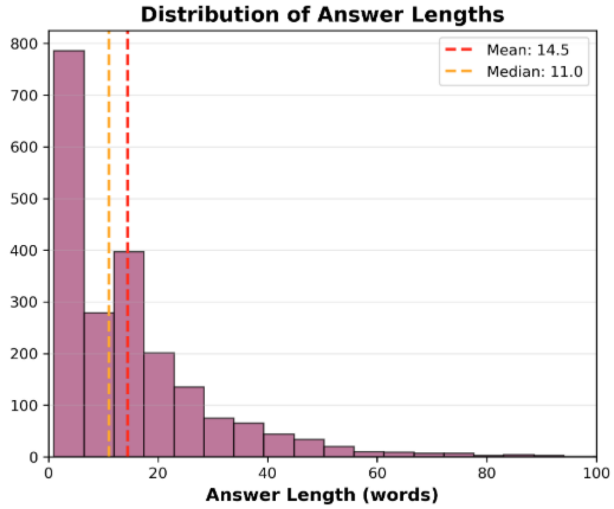

Figure 2: Distribution of answer lengths in the validation split.

- Mean: 14.5 words
- Median: 11.0 words
- Range: 1–220 words

Answer lengths are more variable (std: 16.1). Most answers are short (under 20 words), but there's a long tail of detailed explanations extending to 220 words. The median being lower than the mean confirms this right-skewed distribution.

Property 3: Figure Type Distribution

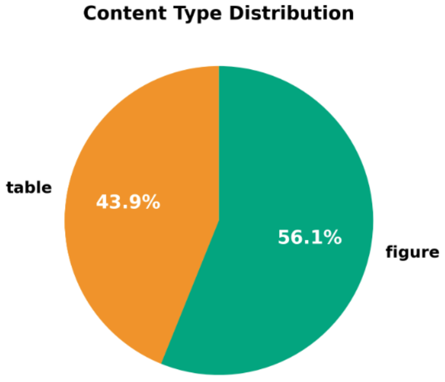

Figure 3: Content type distribution (high-level).

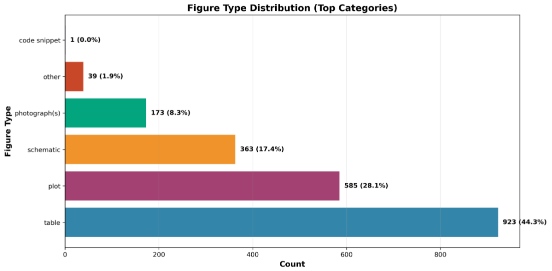

Figure 4: Detailed figure type distribution.

Content Type (high-level):

- Figures: 56.1% (1,170 questions)
- Tables: 43.9% (915 questions)

Detailed Figure Types:

- Tables: 44.3% (923)
- Plots: 28.1% (585)
- Schematics: 17.4% (363)
- Photographs: 8.3% (173)
- Other: 1.9% (39 + 1 code snippet)

The dataset has good diversity across visual modalities, with tables and plots dominating but meaningful representation of diagrams and photographs.

Property 4: Number of Figures/Tables Referenced per Answer

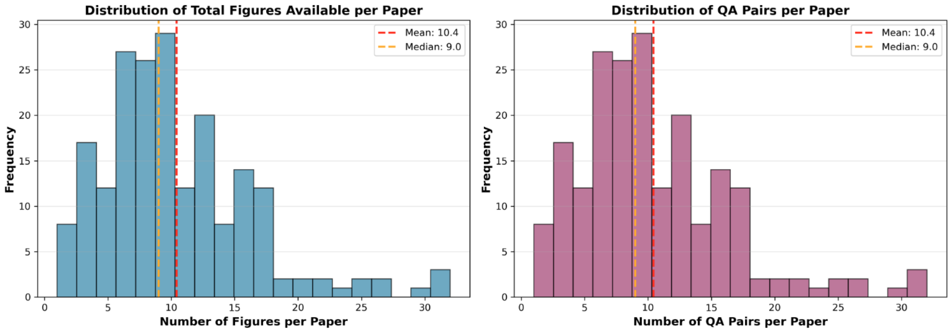

Figure 5: Distribution of figures per paper (left) and QA pairs per paper (right).

Every QA pair references exactly 1 figure/table—there's zero variation in the validation split. This is intentional in SPIQA's design: each question is paired with a single visual element.

The interesting part though is that even though each question only references 1 figure, each paper contains multiple figures (mean: 10.4, range: 1–32). This creates two evaluation scenarios:

"Direct QA" mode (minimal context):

- Input: question + the 1 referenced figure
- This is what the reference field points to
- Context varies much less here since it is only 1 image per question

"Full paper" mode (rich context):

- Input: question + all ∼10 figures from the paper
- This tests whether the model can find the right figure and ignore distractors
- Context varies a lot here depending on the paper

The number of figures per paper exactly equals the number of QA pairs per paper (correlation = 1.000). This means:

- Each figure gets exactly 1 question
- Papers with 20 figures have 20 questions
- Papers with 5 figures have 5 questions

Key Insights for Inference Planning

1. **Context window requirements:** Most questions + answers fit comfortably in ~30 words, so token usage per inference will be dominated by image encoding, not text.
2. **Visual complexity:** Nearly half the questions reference tables (which often have dense text), while the other half reference figures (plots, schematics, photos). This means the chosen models need strong OCR capabilities for tables and spatial reasoning for diagrams.
3. **Answer generation:** Median answer is 11 words, but we will need to handle occasional 100+ word explanations. Budget ~50–100 output tokens per inference to be safe.
4. **Context variance:** Context does not vary much between questions when only passing the reference image, but when passing the whole paper's context, it can vary significantly as there is substantial variance in the number of figures for each paper.

2.4 [0.5 POINTS] METRICS USED

For our answer evaluation task, we use:

1. **Verdict Accuracy:** Three-class classification accuracy (correct/partially correct/incorrect). Also per-class precision, recall, F1.
2. **Error Category Accuracy:** For incorrectly-verdicted samples, does the model identify the right error type? Exact-match on category field.
3. **Figure-Type Breakdown:** Verdict and error category accuracy by figure type (tables, plots, schematics, etc.).
4. **Context Ablation:** Compare accuracy across C1 (figure only), C2 (figure + paragraphs), C3 (full paper).
5. **Explanation Quality:** Manual assessment on ~50 samples for localization accuracy and coaching usefulness.

2.5 [2 POINTS] BASELINES

We select four baselines representing different architectural approaches and capability tiers, reframing their relevance from SPIQA's original answer generation task to our answer evaluation task.

1. LLaVA 1.5 Liu et al. (2023)
Paper: Liu et al., "Improved Baselines with Visual Instruction Tuning"
Link: `https://arxiv.org/abs/2310.03744`

LLaVA 1.5 is an open-source multimodal model that we use as a baseline to test whether lightweight vision-language models can detect incongruence between user explanations and figure content. Its substantial fine-tuning gains on SPIQA (+31.59 L3Score) suggest it may learn to distinguish correct from incorrect figure interpretations with domain-specific training.

2. GPT-4o OpenAI (2024)
Paper: OpenAI, "GPT-4o Release"
Link: `https://openai.com/index/hello-gpt-4o/`

GPT-4o achieves the highest performance on SPIQA answer generation (64.00 L3Score on test-A), making it our ceiling reference for answer evaluation. If a model can generate correct answers, it should theoretically also recognize when a user's answer is incorrect—our task tests whether this generation capability transfers to evaluation capability.

3. Gemini 1.5 Pro Reid et al. (2024)
Paper: Reid et al., "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context"
Link: `https://arxiv.org/abs/2403.05530`

Gemini 1.5 Pro shows strong context-dependent gains on SPIQA (+8.31 L3Score with full paper context versus figures alone), making it ideal for testing our H2 (context) hypothesis: that paper context helps detect reasoning-dependent errors more than visually-obvious factual errors. Its long-context capability enables our C3 condition without truncation.

4. TabRAG Si et al. (2025)
Paper: Si et al., "TabRAG: Tabular Document Retrieval via Structured Language Representations"
Link: `https://arxiv.org/html/2511.06582v1`

TabRAG converts table images into structured language representations (JSON/markdown) for improved retrieval and question answering. This directly informs our table-to-structured-text ablation study: if converting tables to structured text improves answer generation, we hypothesize it will also improve error detection—helping Study Coach decide whether to OCR tables before running the coaching agent.

3 TEAM

3.1 EXPERTISE

We have the following expertise in the underlying modalities required by this task:

1. **Dipti Aswath:** Experience with production agentic systems using LangGraph agents (ML-Amp) and curriculum development for AI courses. Completed prior 2 courses in the Generative AI certificate program.
2. **Buddhika Kahawitage:** Prior experience with small language model fine tuning. Familiar with cost estimation and performance analysis. Completed 11967-A Large Language Models: Methods and Applications.
3. **Jiashan Cui:** Completed coursework in Artificial Intelligence and used OpenAI API in class projects to process and analyze data with large language models. Professional experience using Azure OpenAI to support pipelines that transcribe and process audio and video data.
4. **Khubi Srinivasan:** Experience with data analysis and visualization. Conducted the modality analysis for this report including statistical characterization of question/answer distributions and figure type breakdowns. Completed prior 2 courses in the Generative AI certificate program.

REFERENCES

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C. Berg, and Tamara L. Berg. Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 762–772, Montréal, Canada, 2012. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *NeurIPS Workshop on Instruction Tuning and Instruction Following*, 2023. URL `https://arxiv.org/abs/2310.03744`.

OpenAI. Gpt-4o release, 2024. URL `https://openai.com/index/hello-gpt-4o/`.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL `https://arxiv.org/abs/2403.05530`.

Jacob Si, Mike Qu, Michelle Lee, and Yingzhen Li. Tabrag: Tabular document retrieval via structured language representations. *arXiv preprint arXiv:2511.06582*, 2025. URL `https://arxiv.org/abs/2511.06582`.

A APPENDIX

A.1 SCOPE SUMMARY

Dimension	**Full Design**	**Course Project Scope**
Dataset	Full SPIQA (270K QAs)	Test-A only (666 QAs)
Augmentation	All splits	Test-A → SPIQA+ (∼200–400 4-tuples)
Error categories	4 (factual, omission, wrong concl., no relevancy)	2 core + 2 stretch
Figure types	All	All (natural distribution; analyze by type)
Verdicts	3 (correct / partial / incorrect)	3 (same)
Models	12 models (as in SPIQA)	2–3 models
Conditions	4 conditions + table ablation	3 core (C0, C1, C2) + C3 optional
Human annotation	Full manual curation	∼80–130 examples (seed + spot-check)

Table 1: Scope summary comparing full design versus course project constraints.

A.2 DATASET CONSTRUCTION PIPELINE

We augment SPIQA Test-A into SPIQA+ using In-Context Learning (ICL) for synthetic data generation, with a two-stage validation pipeline and a quality gate.

Phase 1: Prepare

1. **Download Test-A:** Fetch SPIQA_testA.json, test-A images, and extracted paragraphs from Hugging Face. Yields 118 papers, 666 curated QA pairs.
2. **Filter for Figure-Grounded QAs:** Select QA pairs where the answer depends on reading the figure. Target ∼200–300 figure-grounded QAs.
3. **Annotate Figure Types:** Classify each figure as table, plot/chart, schematic, or visualization using an LLM classifier.

Phase 2: Generate

4. **Hand-Curate ICL Seed Set:** Manually create 20–30 high-quality 4-tuple examples (5–8 per error category across figure types).

5. **Generate Congruent Samples:** Use SPIQA ground truth answers as "correct" user answers.
6. **ICL-Based Synthetic Generation:** Prompt an LLM with 2–3 seed examples to generate 1–2 incongruent user answers + structured model responses per QA.

Phase 3: Validate

7. **LLM Validator:** Send each synthetic example to a different model to verify error presence and category.
8. **Human Spot-Check:** ~50–100 samples, stratified across error categories and figure types.
9. **Quality Gate:** LLM validator must agree with human judgment $\geq$90% on spot-check subset.
10. **Final SPIQA+ Dataset:** ~200–400 validated 4-tuples with figure-type annotations.

Total human annotation effort: ~80–130 examples (20–30 seed set + 50–100 spot-check).

A.3 SPIQA+ 4-Tuple Structure

Each example in our augmented SPIQA+ dataset is a 4-tuple:

Field	**Source**	**Description**
Context	SPIQA	Visual and textual context from the paper
Question	SPIQA	The figure-grounded question
User Answer	Generated (synthetic)	A student's explanation (correct/partial/incorrect)
Model Response	Generated (synthetic GT)	Verdict + error category + explanation

A.4 Structured Model Response Format

Component	**Values**	**Purpose**
Verdict	Correct / Partially Correct / Incorrect	Tests H1 (detection)
Error Category	Factual / Wrong Conclusion / None (core)	Tests H4 (error type)
Explanation	Free-text coaching feedback	Qualitative evaluation

A.5 Experimental Conditions

These conditions follow the spirit of SPIQA's three tasks but are adapted for our answer evaluation task. The key question is: what information helps the model evaluate a user's explanation? These conditions systematically ablate input modalities to answer this.

Condition	**Input to Model**	**What It Tests**
C0: Text Only	Question + user answer + caption (no figure)	Is the task fundamentally multimodal?
C1: Figure Only	Figure image + question + user answer	Can vision alone detect errors?
C2: Figure + Paragraphs	C1 + caption + relevant paragraphs	Does textual context help?
C3: Full Paper (stretch)	C2 + full paper text	Does broader context help further?

Table 2: Experimental conditions for modality and context ablation.

Comparing C0 vs. C1 tests whether the task is fundamentally multimodal (Section 1.2), as C0 helps test if language (text) alone can detect errors. Following Dodge et al. Dodge et al. (2012), who found humans achieve 91% accuracy on visual text detection without seeing images, we expect C0 to perform above chance but below C1 - validating that cross-modal verification is required.

A.6 Table-to-Structured-Text Ablation

The SPIQA authors noted that tables are hard for current MLLMs because models struggle with spatial cell relationships in images. We propose a focused ablation on the table subset: convert table images to structured text (JSON or markdown) and re-run conditions C1 and C2.

Research Questions:

- Does structured text improve error detection accuracy on tables?
- Does structured text reduce the value of paper context?

Practical Implication: If structured text dramatically improves evaluation accuracy on tables, it tells Study Coach to OCR/parse tables before running the coaching agent rather than relying on vision alone.