

# DOES CONTEXT MATTER?


Evaluating the role of context in vision language model chart understanding and multimodal retrieval

Marek Lazár

The background features a light gray surface with subtle, flowing white wave-like patterns. A dark gray, almost black, geometric footer is positioned at the bottom, consisting of several angular shapes that create a modern, architectural look.

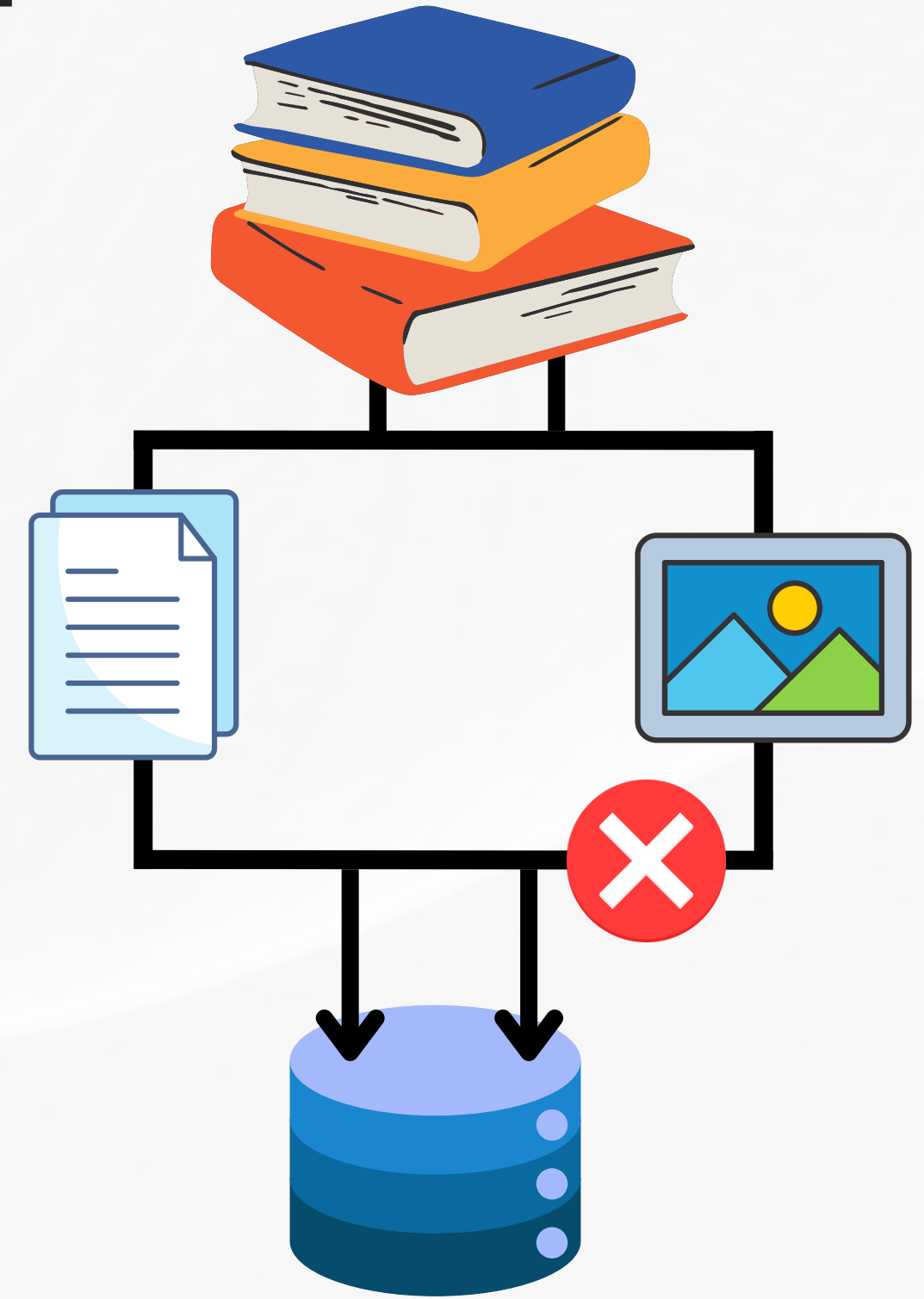
# **Background**

# Background: Motivation

- Project hosted by  **Workspace365**
  - B2B SaaS – digital workspace platform
- Provides a RAG chatbot for querying internal knowledge bases
- Current system: text-only retrieval

## Limitations & Challenge:

- Images missing from retrieval
- Enterprise documents often include images of charts
- Charts encode dense visual info, difficult for AI to interpret

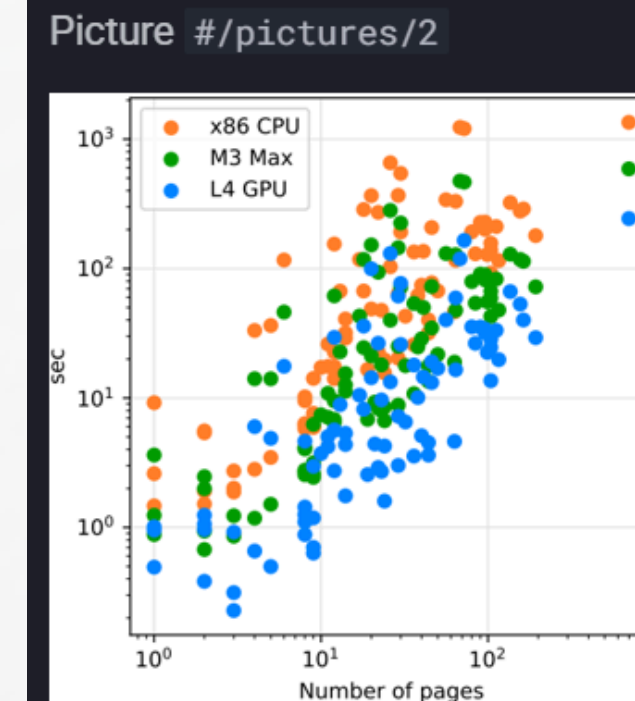


# Background: Potential solutions

- Swap chatbot model → multimodal LLM (can process images + text)
  - Limitation: Proprietary multimodal models are costly & resource-heavy
- Alternative: Small local VLMs are viable but can show poor or inconsistent performance

Challenge: Balancing cost-efficiency with performance

Key Question: What are the current shortcomings in chart understanding of state-of-the-art models?



**Caption**

Figure 3: Distribution of conversion times for all documents, ordered by number of pages in a document, on all system configurations. Every dot represents one document. Log/log scale is used to even the spacing, since both number of pages and conversion times have long-tail distributions.

**Annotations (ibm-granite/granite-vision-3.1-2b-preview)**

In this image we can see a graph. On the x-axis we can see the number of pages. On the y-axis we can see the seconds.

# Literature Review

# Literature review

---

- VLM progress: Strong advances, but deep chart understanding remains a challenge
  - ChartQA (Masry et al., 2022): Semi-structured real-world charts; early promise
  - New benchmarks (CharXiv, ChartQAPro, ChartMuseum, 2024–25): Reveal weaknesses in visual + synthesis reasoning, even for top proprietary models
  - Liu et al., 2025: Chart reasoning flagged as key limitation, esp. for multi-step & integrative tasks
- Research focus to date: mostly on Chart QA (fact extraction) → but chart understanding requires richer, qualitative evaluation through interpretive storytelling

# Literature review

---

- Context matters: in real-world documents, charts appear with captions and surrounding text
  - Humans rely on this context to ground and interpret charts
- Gap in research: little work tests whether surrounding text improves chart interpretation and its impact on multimodal retrieval

## This motivates two questions:

- Does adding surrounding textual context improve chart interpretation quality?
- Do richer chart interpretations (with context) improve retrieval in multimodal RAG?

# Methods

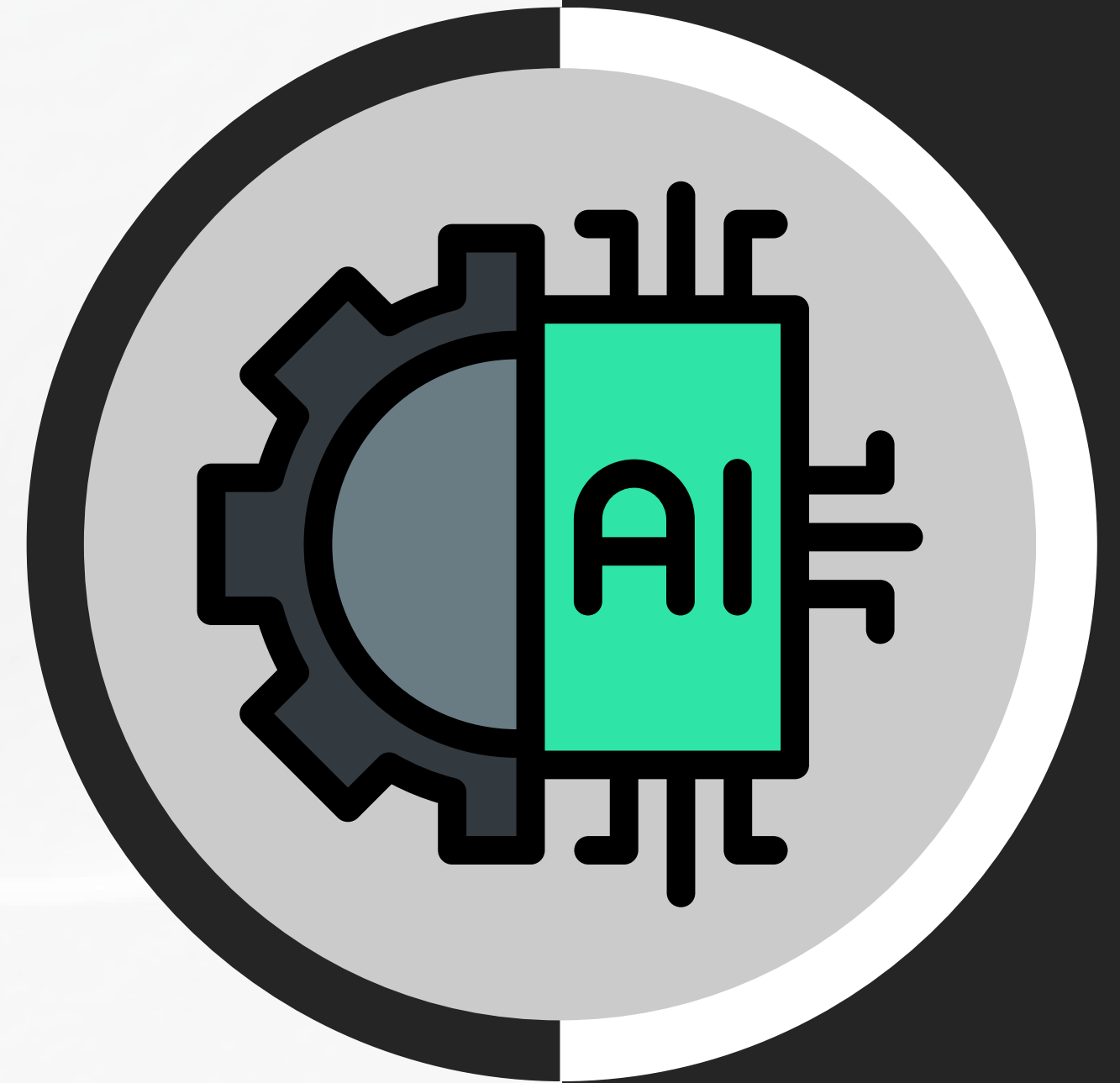


# Methods

**Goal:** Investigate the effect of the surrounding textual context on VLMs' chart interpretive capabilities & multimodal retrieval

## Two Research Questions:

1. Does context improve chart interpretation quality?
2. Does context enhanced chart interpretations increase retrieval performance in multimodal RAG?



# Methods



To assess whether surrounding textual context improves chart interpretation quality, two experimental conditions are defined:

- Image-only
- Image & Context

Using a 7-point Likert Scale, the generated chart interpretations were manually evaluated on:

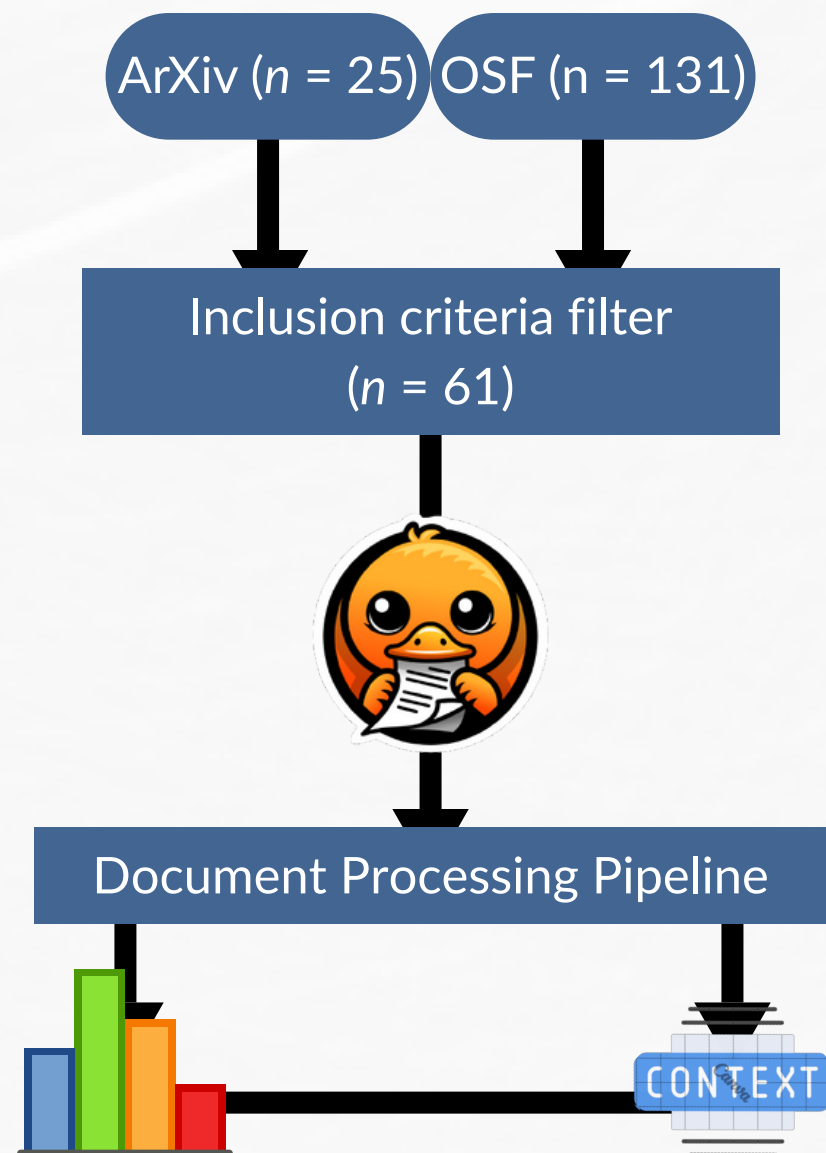
- Factual Accuracy
- Clarity & Coherence
- Relevance
- Completeness

Comparative preference judgement (Text A or Text B) were given.



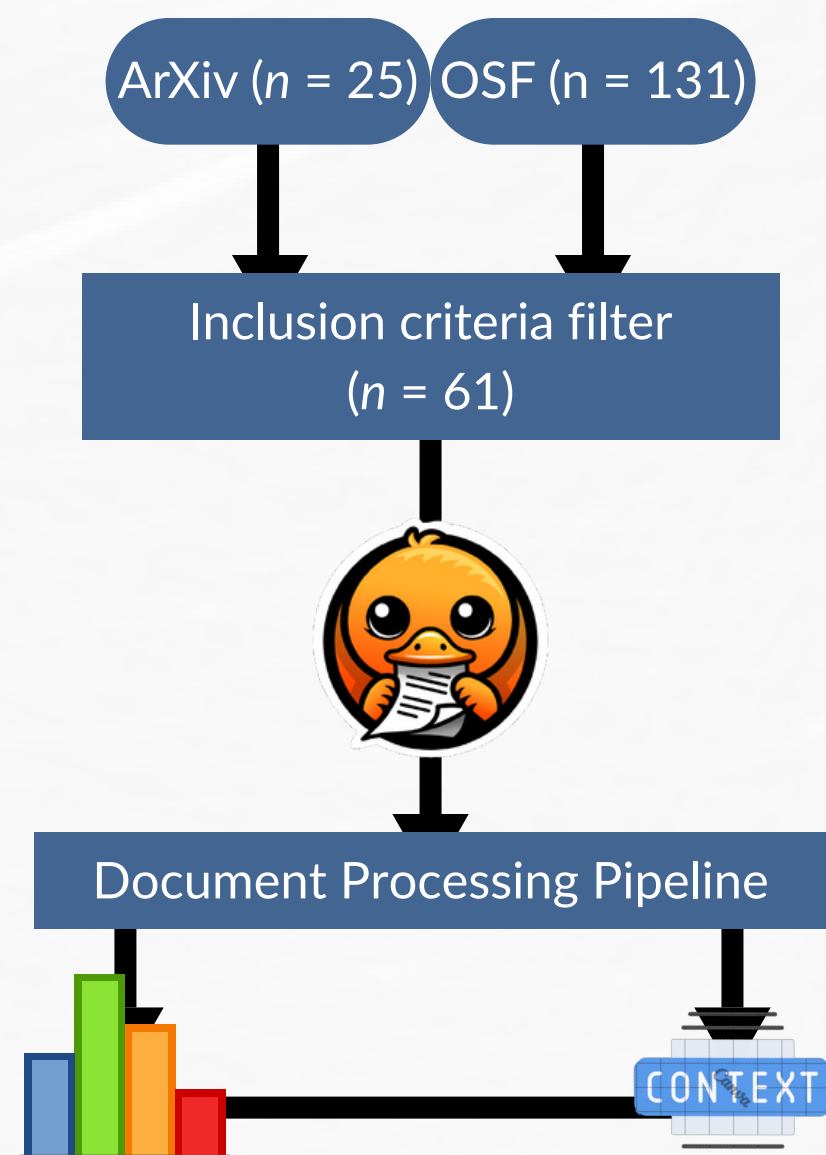
# Data, Materials & Procedure

## Stage 1

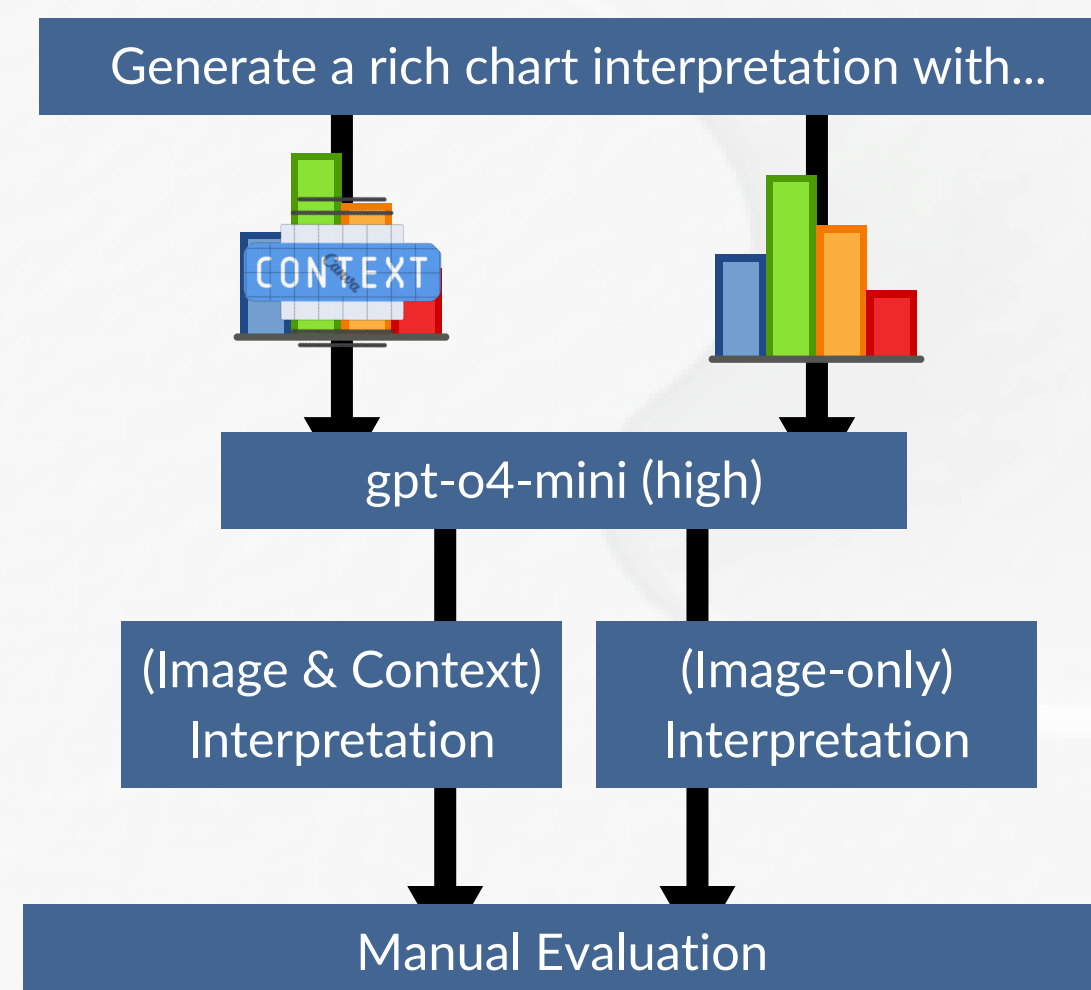


# Data, Materials & Procedure

## Stage 1

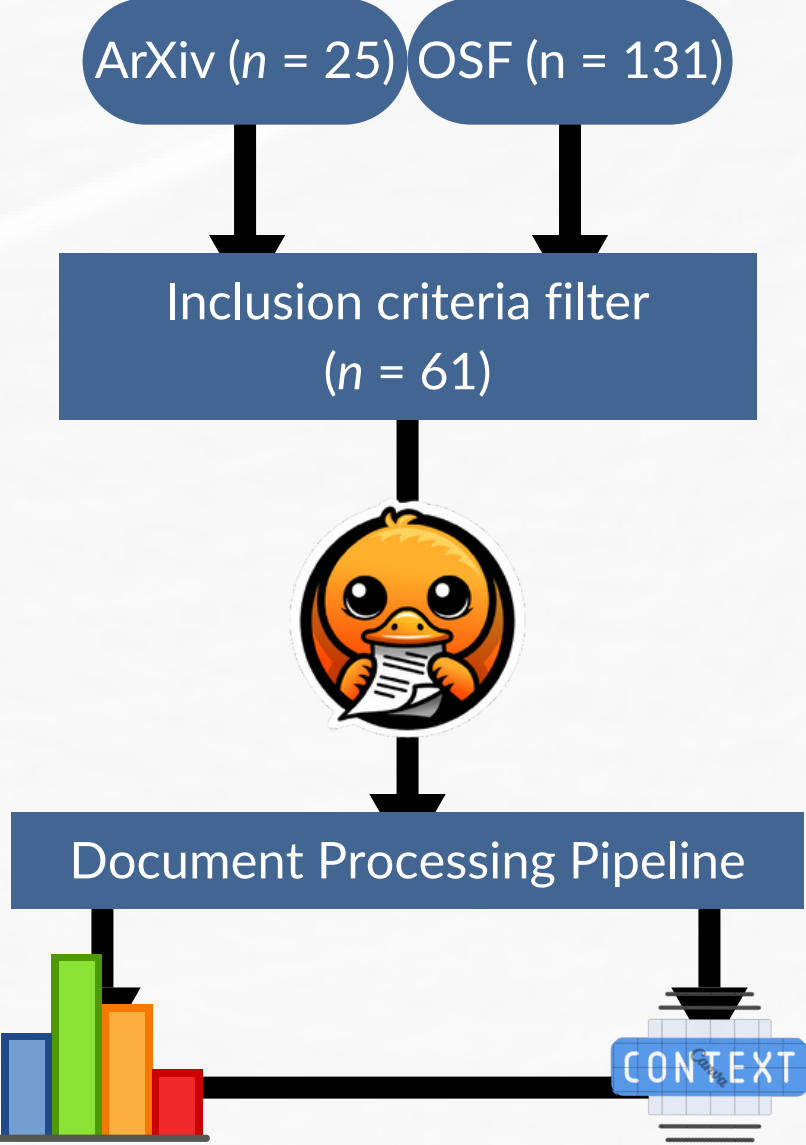


## Stage 2

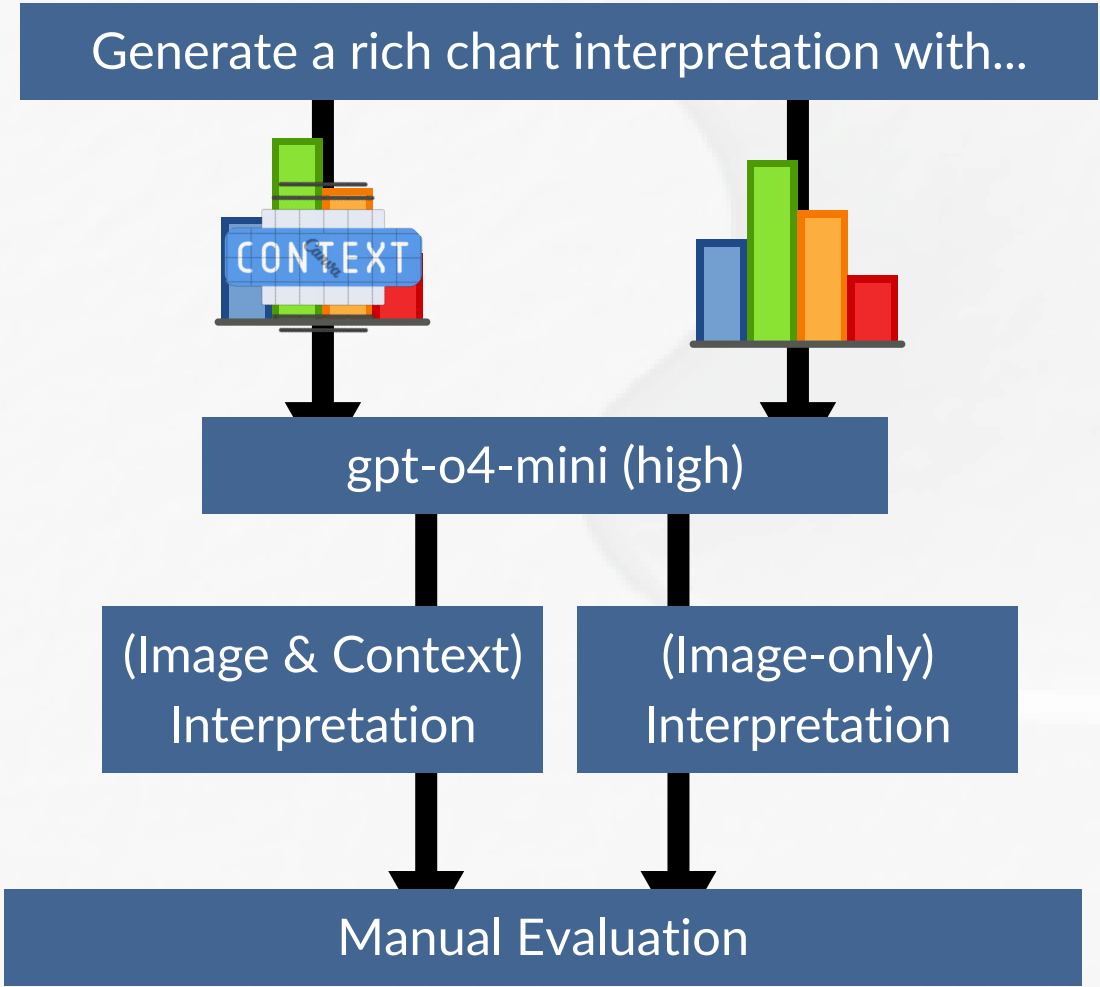


# Data, Materials & Procedure

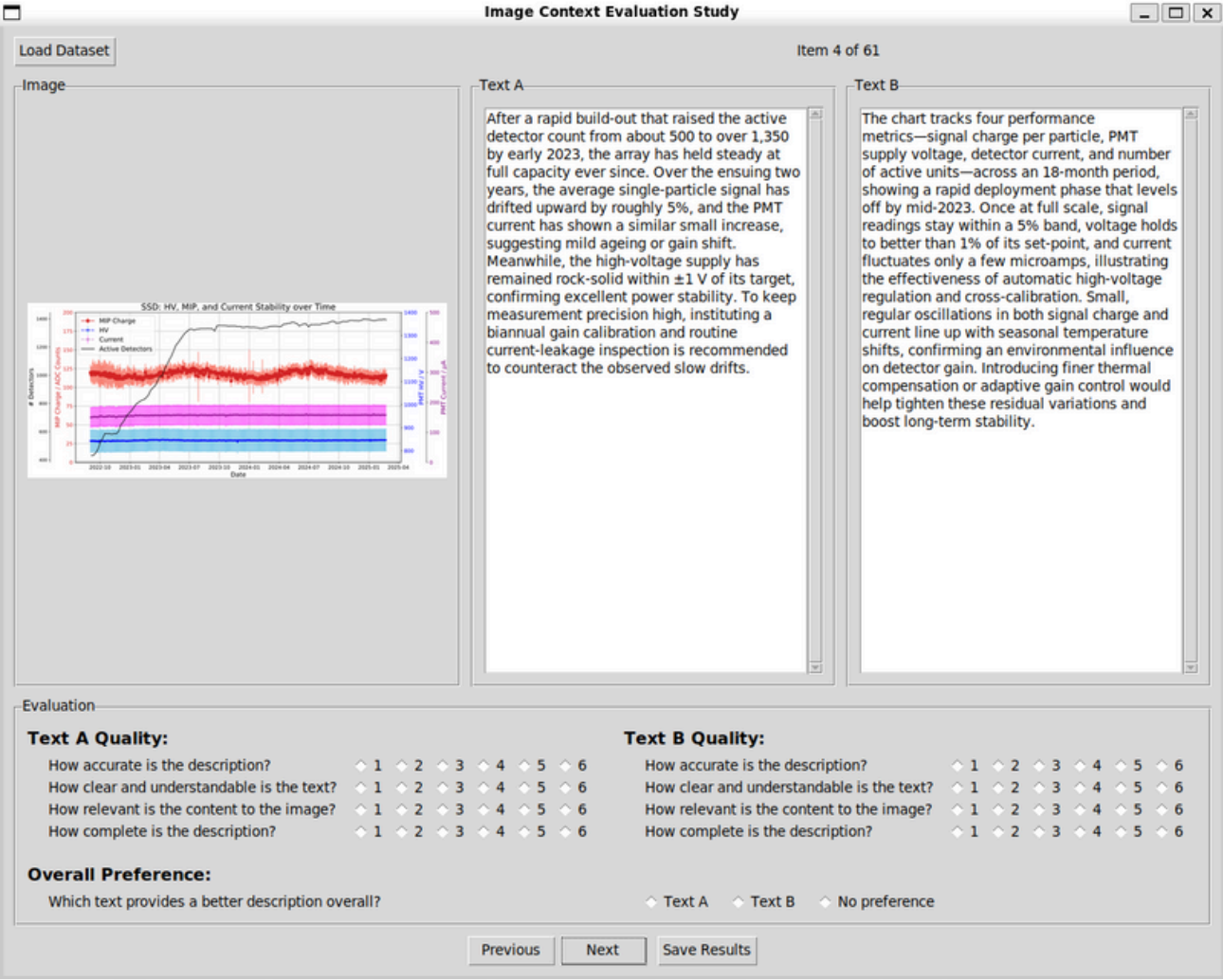
## Stage 1



## Stage 2

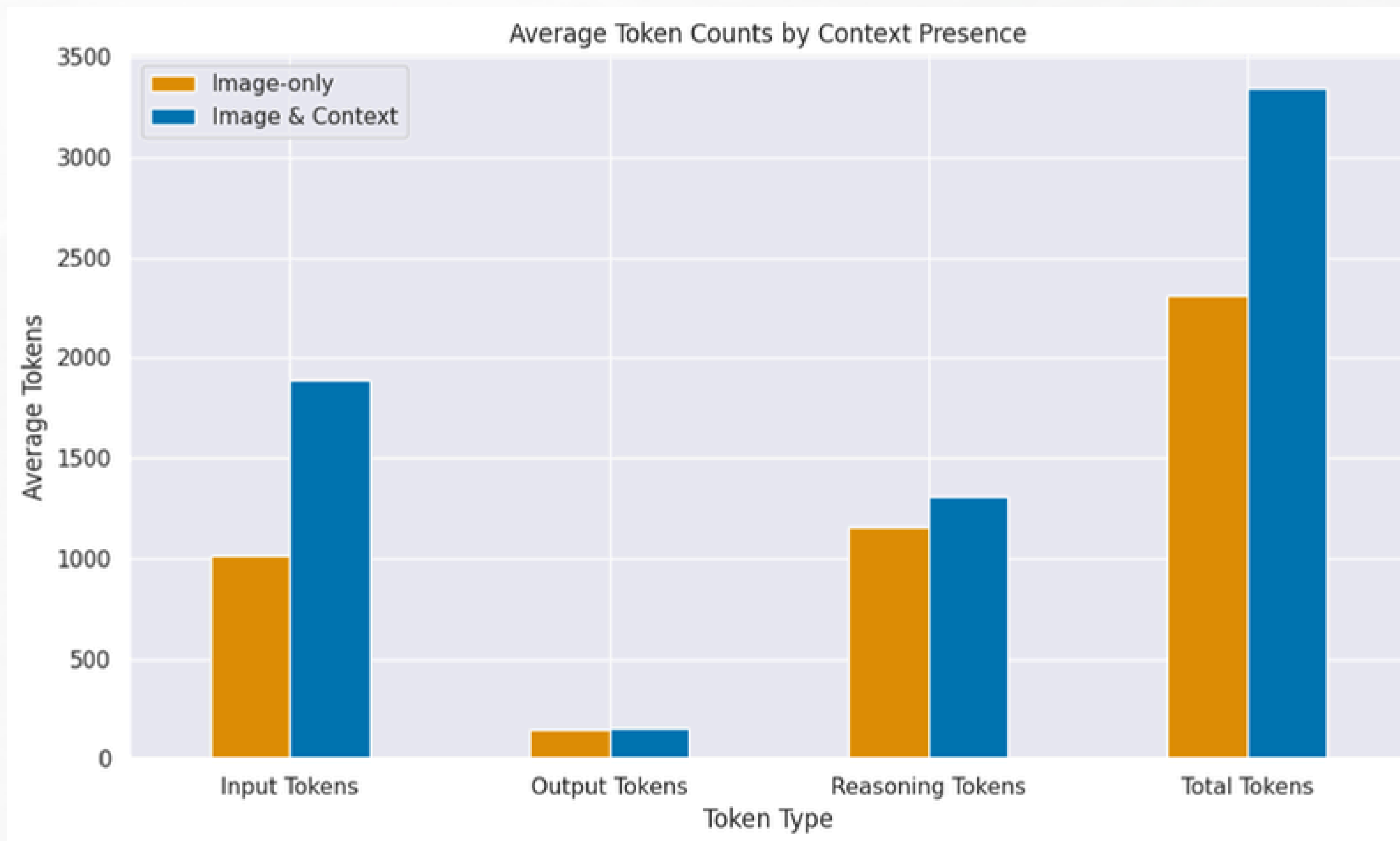


## Stage 3

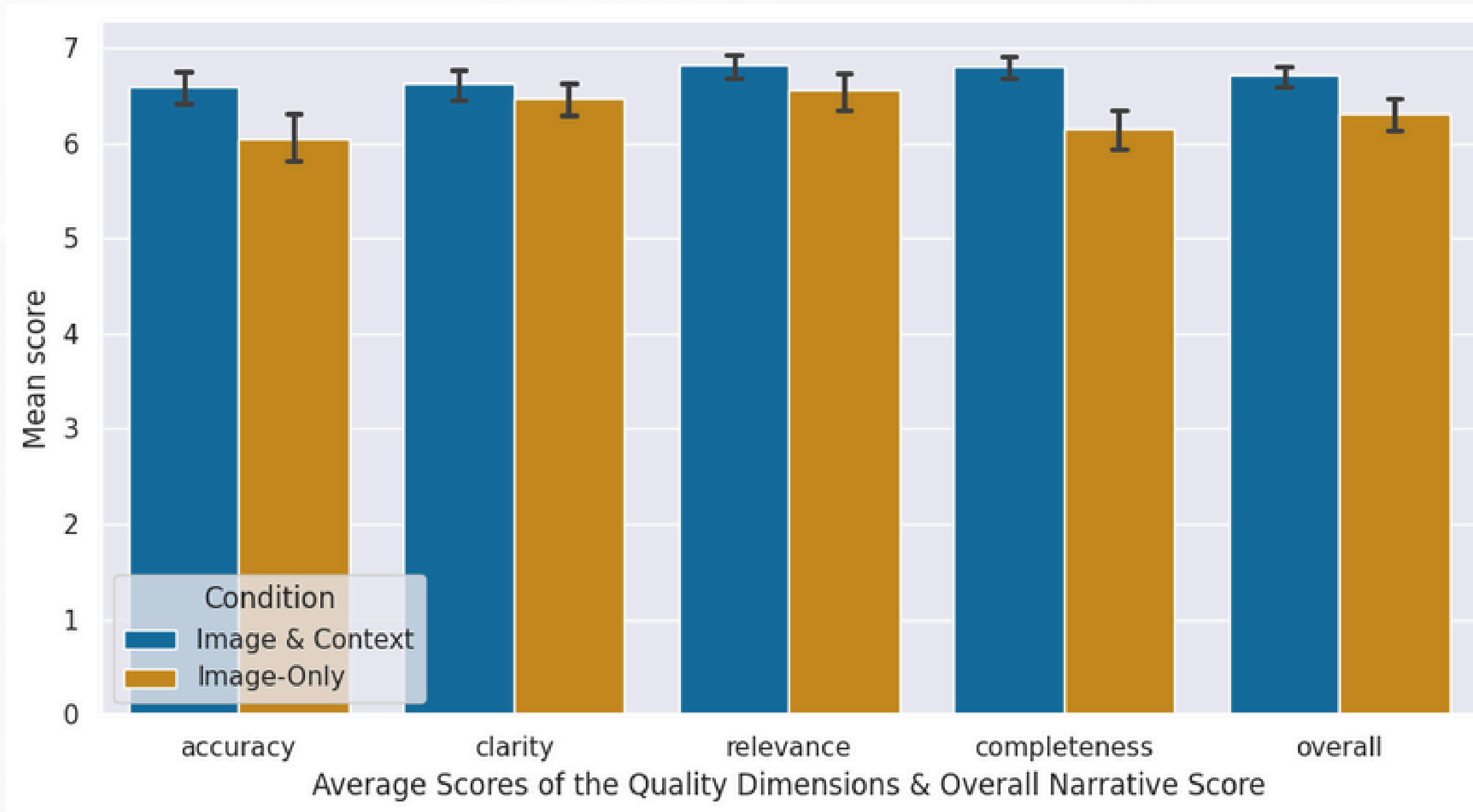


# Results

# Token analysis



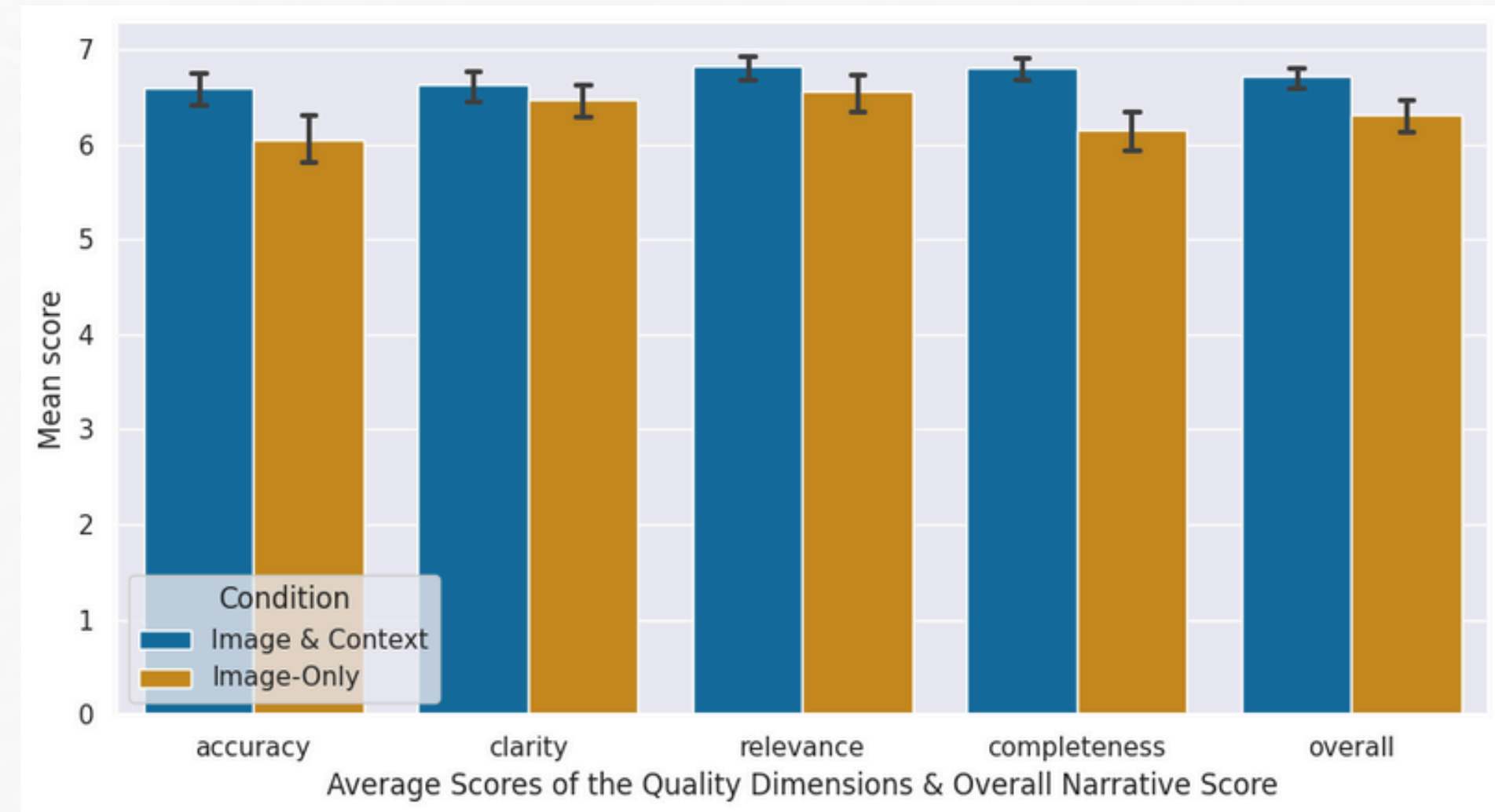
# Key findings



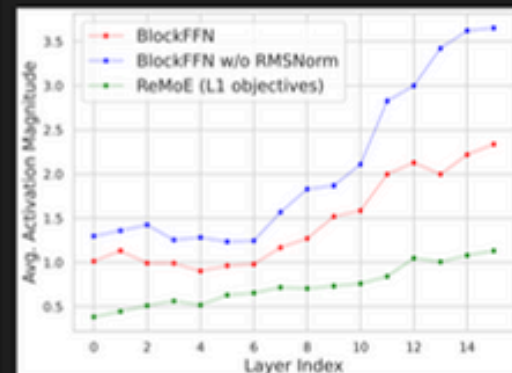


# Key findings

- Overall quality significantly higher with Image & Context than Image-only ( $W = 163.5$ ,  $p < .001$ ,  $rb = .59$ ).
- Factual accuracy, relevance, and completeness all significantly improved with context.
- Clarity & coherence → no significant difference ( $p = .105$ ).
- Preference judgments: Image & Context favored in 45 cases, Image-only in 8, no preference in 8.



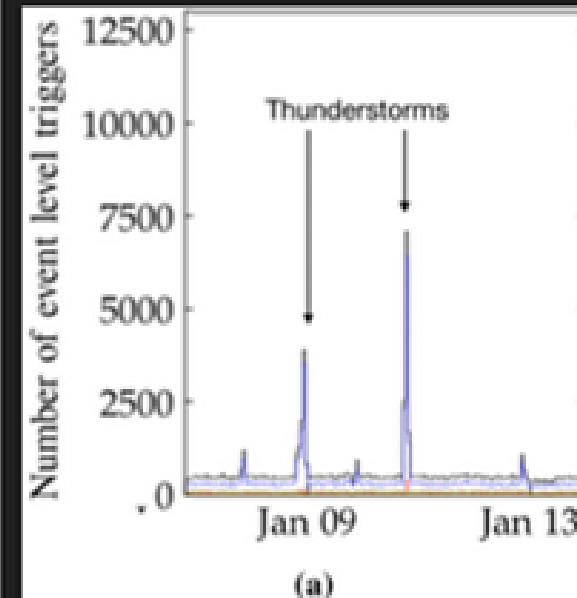
# Error Analysis



## With context

The chart shows that both BlockFFN variants steadily amplify their activation strength the first to the last layer, far surpassing modest growth seen in the competing Removing RMSNorm in BlockFFN triggers most dramatic rise, with activation magnitudes nearly tripling by the final layer, while RMSNorm yields a smoother increase outperforms alternatives. In contrast, the L1-based regularization keeps activations relatively low, suggesting a more uniform but weaker participation of its sub-networks. This demonstrates that tailoring the loss around the classification token inherently drives stronger, more concentrated learning in deeper layers, which can speed up convergence and improve inference efficiency. Applying a block-level

(green line) holds activations consistently low, demonstrating a tightly controlled flow of information. These differences highlight that both normalization and targeted sparsity are powerful levers for keeping neural activations in a healthy range and promoting more stable learning.

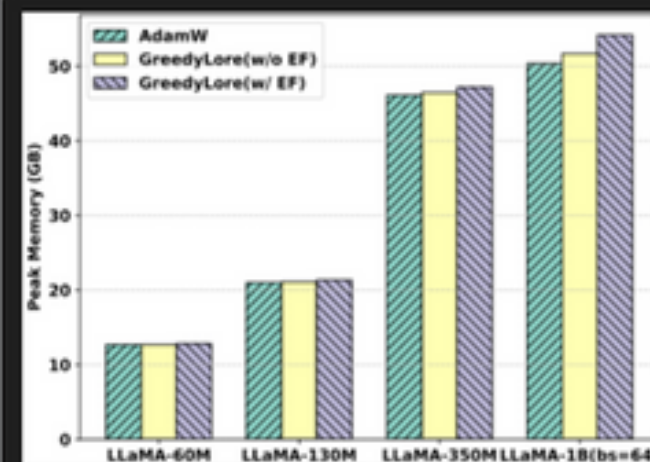


## With context

The chart illustrates the number of basic detector triggers recorded over several days, with two pronounced peaks that coincide with periods of thunderstorm activity. Outside these storm windows, the trigger rate stays low and steady, but during lightning events it soars by more than an order of magnitude, overwhelming the normal data flow. These surges are caused by electrical noise from lightning, which makes neighboring stations fire in unison as if they saw real air showers. The resulting backlog of spurious events clogs the communication pipeline, delaying genuine cosmic-ray data requests and dropping the purity of collected events. Introducing real-time noise filtering or flagging during storms would help prioritize valid signals and maintain high detector uptime.

## Without context

The chart tracks event-level trigger counts over several days and shows a steady low-level background punctuated by two dramatic spikes peaking at about 4,000 and 7,000 triggers. Each of these surge events aligns with periods of thunderstorm activity, indicating that storms drive a sudden and large increase in detected events. Outside of these bursts, trigger rates remain stable and modest, underlining how distinct the storm-related enhancements are. Recognizing and filtering out these weather-induced spikes can improve the reliability of background measurements and help isolate genuine signals of interest.



## With context

The chart shows that peak GPU memory usage rises from roughly 12 GB for a 60 million-parameter model to about 50 GB for a 1 billion-parameter model as model size increases. Across all sizes, GreedyLore—with or without the error-feedback mechanism—matches the baseline AdamW memory footprint almost exactly, with only a 1–2 GB uptick when error feedback is enabled. This small increase confirms that the extra storage for error feedback and the projection matrix remains negligible compared to activation tensors. In practice, you can adopt GreedyLore for large-scale pre-training without worrying about any significant GPU memory penalty.

## Without context

The chart compares peak GPU memory use of three training approaches—AdamW, GreedyLore without an extra feature (EF), and GreedyLore with EF—across four LLaMA model sizes. For the smallest 60 M-parameter model, GreedyLore without EF uses slightly less memory than AdamW, while adding EF nudges it just above the baseline. As models grow to 350 M and 1 B parameters, both GreedyLore variants begin to exceed AdamW's footprint, with the EF-enabled version showing the largest overhead. This pattern highlights that the extra feature introduces a moderate but compounding memory cost at scale. In practice, teams working with very large models should plan for this additional GPU memory requirement or consider sticking with the lighter-weight version when resources are limited.

# Error Analysis

## Key Takeaways:

- Context enhanced the chart interpretations for more contextually grounded, accurate, relevant, and complete narrative.
- Provided domain-relevant terminology.

However, ...

- At times, extraneous details not derivable from the chart were added to the interpretation.
  - Reduced clarity & understanding; obscured the central message.

Less common errors: omission / false information

# Methods 2

# Methods

## ☑ **Model & Conditions**

- **Model:** InternVL3.5-1B
- **Embedding:** BAAI/bge-base-en-v1.5
- **Vector store:** Chroma vector DB
- **Conditions:**
  - Image-only interpretations
  - Image & Context interpretations

## ☑ **Data & Procedure**

- **Procedure:**
  - Chart interpretations generated by small VLM
  - Natural-language queries (blind to outputs) were manually generated
  - Retrieval tested separately on both collections

## ☑ **Evaluation**

- **Evaluation:**
  - Metrics: Top-1 hit-rate & mean cosine similarity
  - Statistical test: paired-samples t-test / Wilcoxon signed-rank test

# Results 2

# Results

## ☑ Top-1 hit rate

- Image & Context → 93.4% (57/61)
- Image-only → 90.2% (55/61)

## ☑ Cosine similarity

- Image & Context (M = 0.34)
- Image-only (M = 0.31)
- Difference statistically non-significant

## ☑ Statistical tests

- Paired-samples t-test:  $t(60) = 1.04$ ,  $p = .304$
- Wilcoxon signed-rank:  $W = 785$ ,  $p = .249$

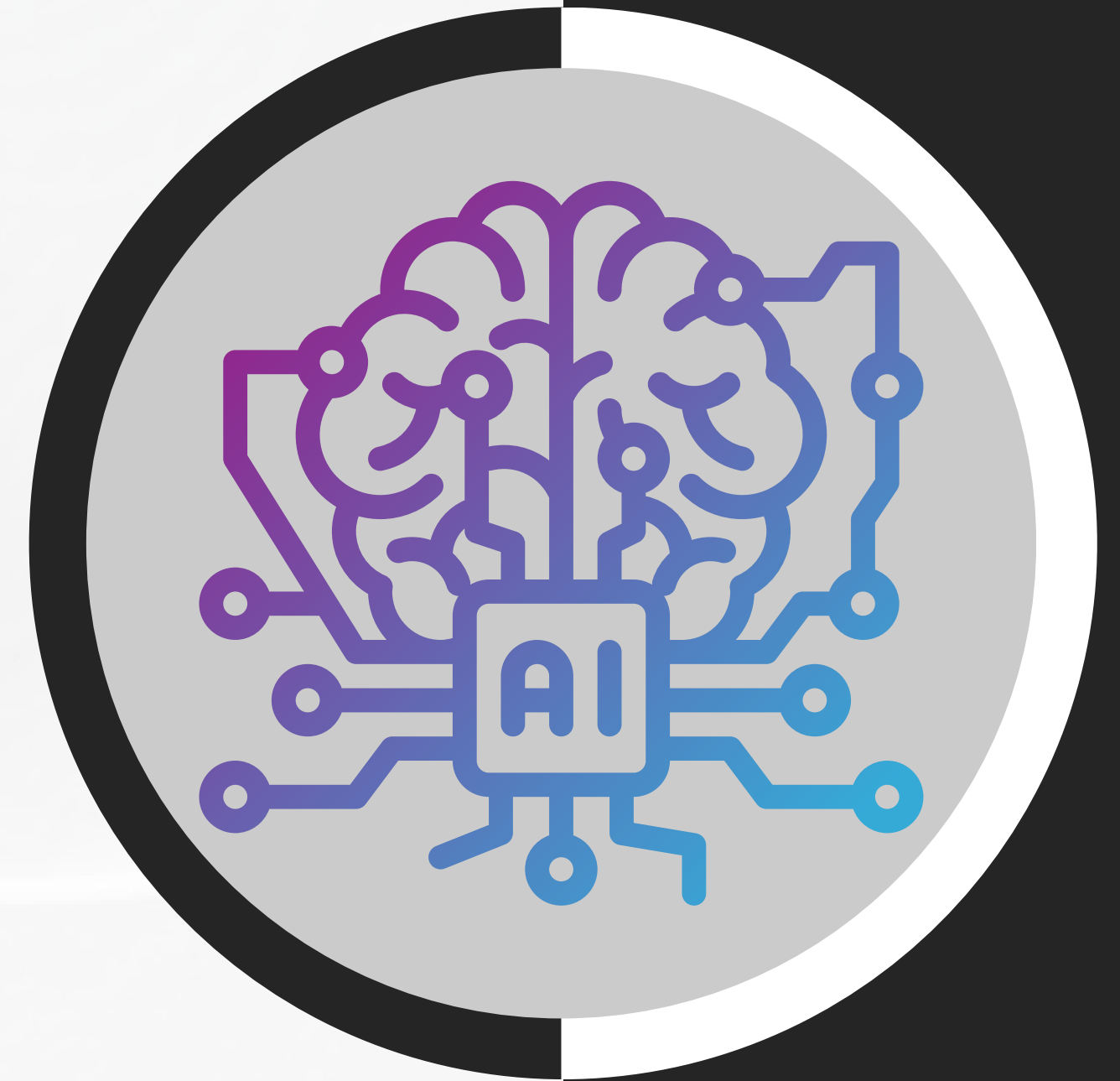
**Conclusion:** Context yielded descriptively better performance, but no significant improvement in retrieval.

# Conclusions



# Conclusions

- H1 (Narrative Quality) → confirmed (except clarity).
- H2 (Retrieval Performance) → direction observed, not significant.
- Context → more grounded, domain-relevant, helps humans & models interpret charts.
  - Sometimes context included extraneous information, confusing the model.
- Better interpretations ≠ better retrieval.



# What do these findings mean?

## Limitations

- Possible researcher bias (easy to guess “context condition”).
- Longer interpretations may inflate perceived quality.
- Retrieval tested only with smaller VLM, not larger models.

## Implications

- Context boosts interpretive quality, but does not guarantee retrieval gains.
- Highlights trade-offs in contextual grounding.

## Future Work

- Test multimodal retrieval with more sophisticated context integration (semantic filtering).
- Explore larger models and more precise context definitions.

---

# Questions?

Thank you for attending

Code available at <https://github.com/lazarmarek/UvA-Thesis>

---