Accepted at NeurIPS 2025 Workshop on
Multimodal Algorithmic Reasoning
MAR 2025

SlideAgent: Hierarchical Agentic Framework for Multi-Page Slide Deck Understanding

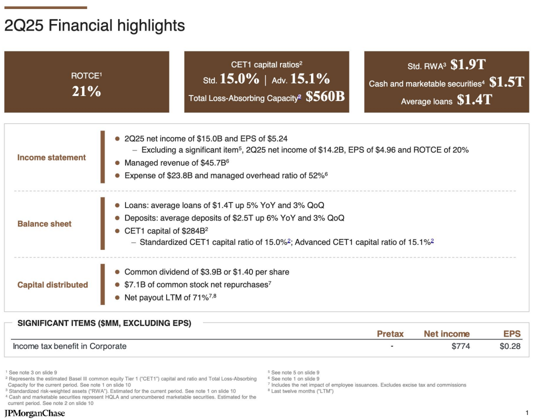
Yiqiao Jin*, Rachneet Kaur, Zhen Zeng, Sumitra Ganesh, Srijan Kumar
Nishan Srishankar and Kelly Patel

J.P. Morgan
Georgia Institute of Technology

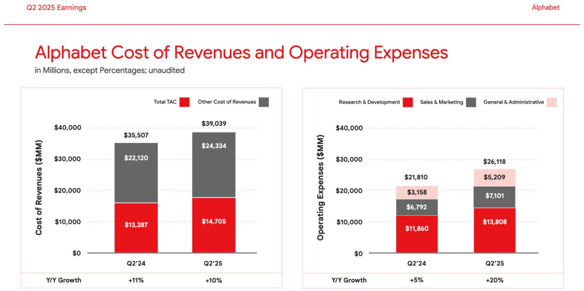
*Work performed as a Summer Intern at J.P. Morgan AI Research

Background & Motivation

- Infographics are **structured visuals** designed to convey complex information.
  - slides, posters, charts, reports
  - Layout, visual hierarchy, and multimodal cues (e.g. color & typography) play a key role in enhancing meaning beyond plain text.
- Multimodal Large Language Models (MLLMs) show strong promise in multimodal understanding.
- LLMs trained on natural images face challenges with spatial relationships, document structure, and narrative flow.



JPMC 2Q25 Earnings



Alphabet 2Q25 Earnings

Challenges

C1: Scalable Fine-Grained Reasoning

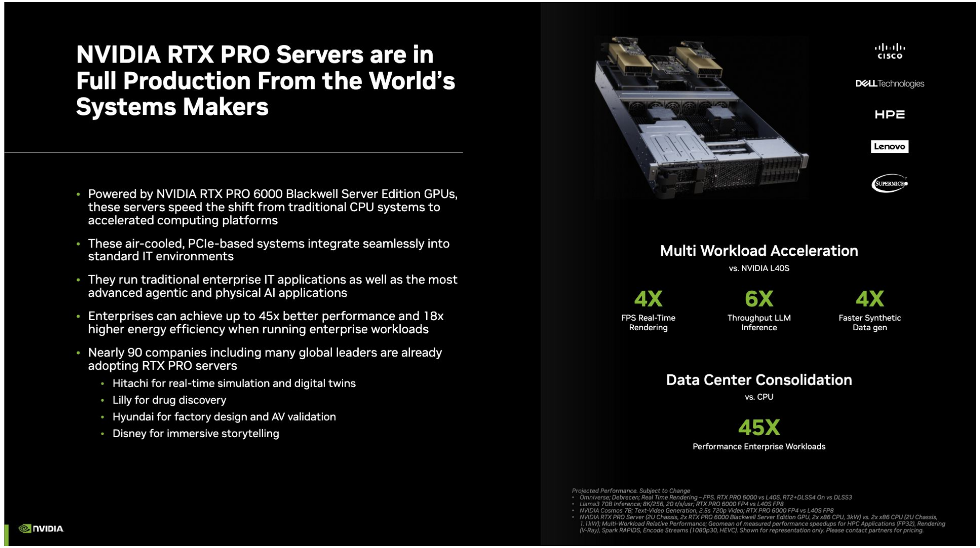
- MLLMs' low-resolution visual encoders miss small fonts, icons, footnotes.

C2: Domain-Specific Visual Semantics

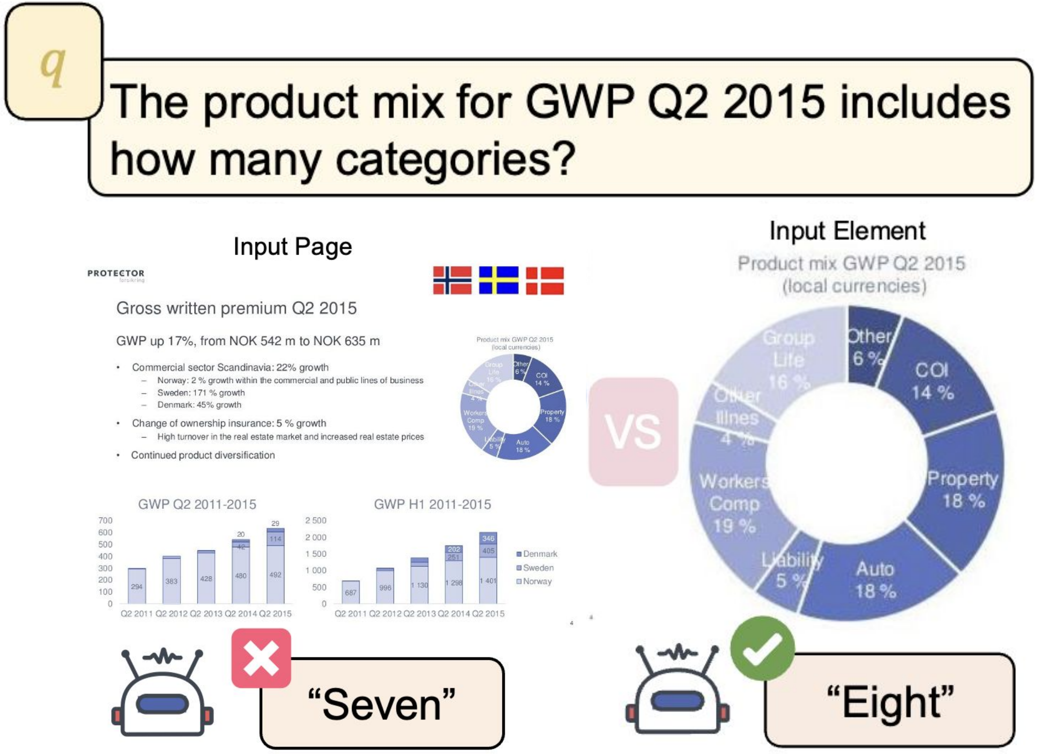
- Identify relevant slides from large slide decks efficiently.
- Charts, logos, colors, icons, spatial layouts.

C3: Metadata-Free Integration

- No reliance on metadata.
- In contrast to PDF-DLA, DocParser, and PDF Plumber.



NVIDIA 2Q25 Earnings



Cropping or highlighting areas of interest significantly improves LLM accuracy.

Contributions

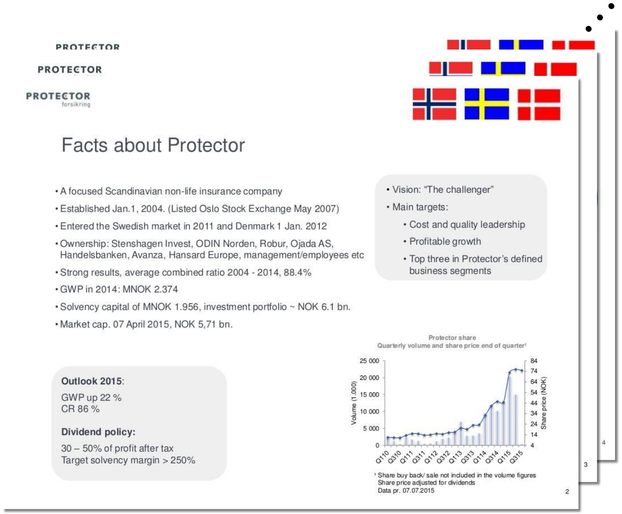
- **SlideAgent: a Hierarchical Agentic Framework**
  - Analyzes multi-page, varying-size visual documents (**C1**)
  - Inspired by the human information processing.
- **Multi-Level Knowledge Construction**
  - Structured knowledge representation of documents for effective retrieval and generation (**C2**, **C3**)
  - **Global**: document-wide topics
  - **Page**: page-specific features and cross-page relations
  - **Element**: fine-grained components, e.g. charts, figures, and text blocks
- **Superior Performance** across open-source & proprietary models
  - Consistent accuracy boost on top of base models
    - +7.9% on GPT-4o (SlideVQA)
    - +9.8% on InternVL3-8B (SlideVQA)
  - Agnostic to model architectures

Method

Problem Statement

Input

- $P = \{p_1, \ldots, p_{|P|}\}$ – A multi-page visual (image formats), e.g. slide deck
- q – A natural language query



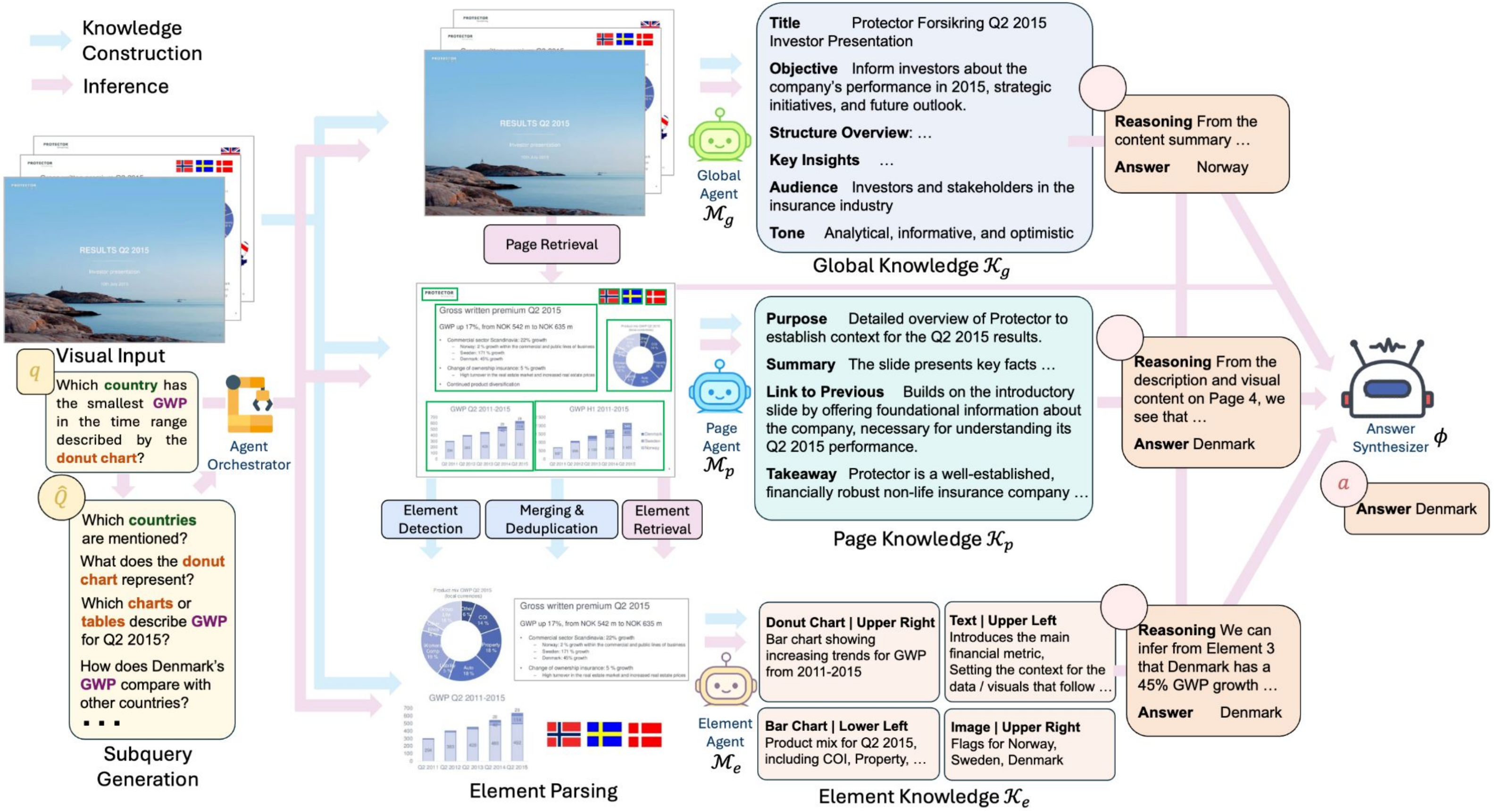
Output

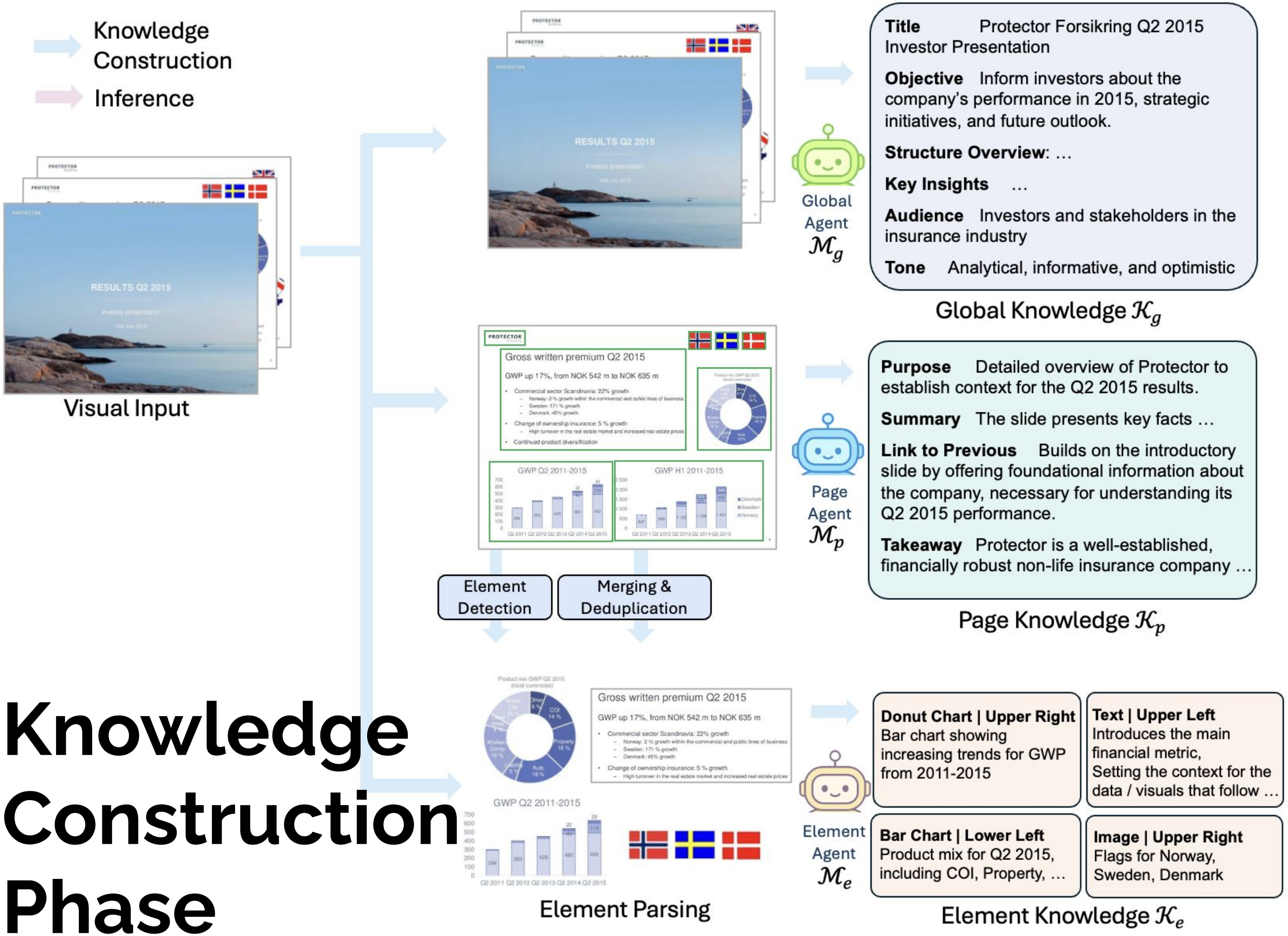
- a – The answer

Goal: Retrieve relevant pages and elements; reason over complex visuals.

Note: We assume NO metadata about the slide deck is provided (e.g. element hierarchy / location).

Method – Overview of SlideAgent





Knowledge Construction Phase

Global Agent: Analyzes entire slide deck for overarching themes & purposes. Generates deck-level summary

Page Agent: creates **per-slide summary**. Analyzes slide-to-slide narrative flow.

Element Agent: Describes individual UI elements

Method – Retrieval & QA Phase

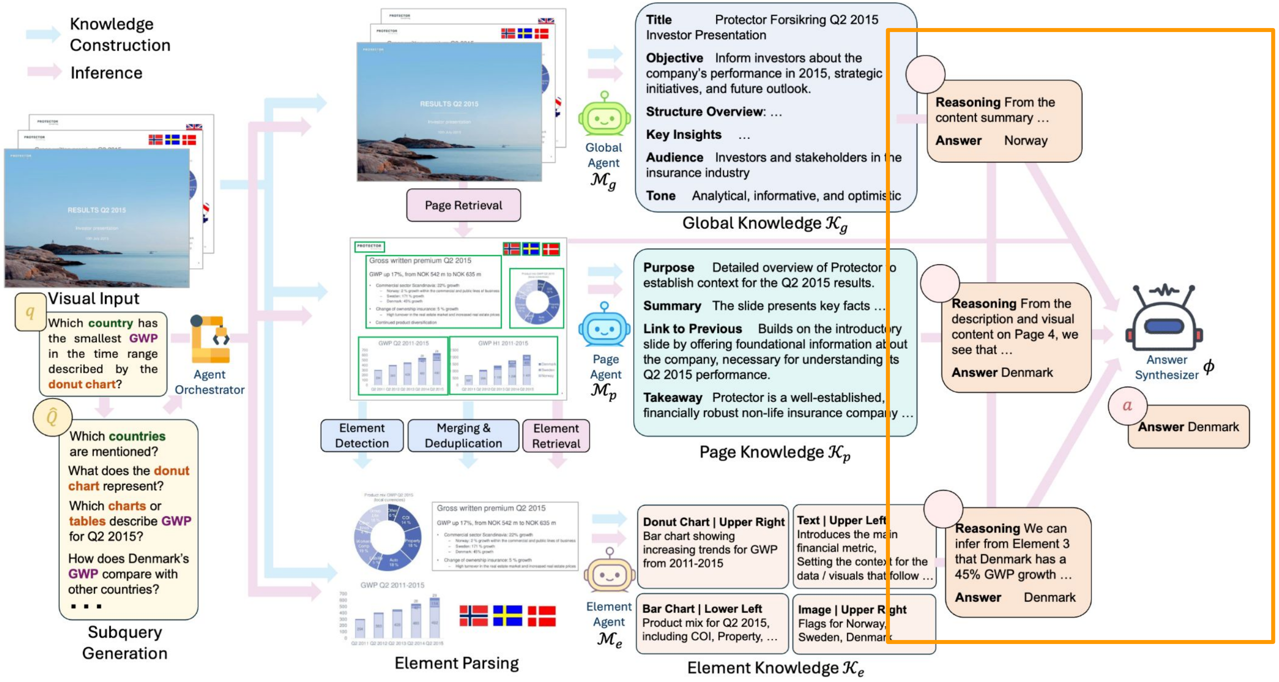
Query Classification

1. Overall Idea / Global-level Query
2. Fact-based Direct Query
3. Multi-hop Reasoning / Comparative Queries
4. Layout / Visual Relationship Queries
5. Cannot decide

Query Enhancement: Generate 5 related subqueries conditioned on both the original query q and the deck summary $\mathscr{K}_g$

Retrieval: Use both the query and subqueries to retrieve top k textual / visual elements

Method – Retrieval & QA Phase



Individual activated agents generate an answer. An answer synthesizer combines answers from all agents based on their reasoning processes.

Evaluation

Experimental Setup – Baselines

Type 1: Multimodal LLMs

- 15 LLMs from 8 model families
- **Proprietary Models:** GPT-4o, Gemini-2.5/2.0, Claude-4.1/3.5
- **Open-source Models:** Llama-3.2, InternVL3-8B, Phi-3-vision, Qwen2.5-VL, LLaVA-1.5 / 1.6

Type 2: Multimodal RAG

- VisRAG, VDocRAG, COLPALI

Type 3: Multi-agent Systems

- ViDoRAG

Datasets

Multi-page Understanding

- SlideVQA [1]
- TechSlides and FinSlides [2]

Single-page Understanding

- InfoVQA [3]

Question: Why did the author invest $1M in the Seed for greenhouse?

Answer: Strong team and market conviction and early customer validation.

[1] Tanaka et al. SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images. AAAI 2023.

[2] Wasserman et al. REAL-MM-RAG: A Real-World Multi-Modal Retrieval Benchmark. ACL 2025.

[3] Mathew et al. InfographicVQA. CVPR 2022.

Experimental Setup – Metrics

Num: Numeric Comparison using Exact Match

- When numeric values are present in the answer
- Normalize the answer and compare with GT
- 8k / 8,000 / 8 thousand / 8.0×10^3 → 8000

F1: String Comparison w.r.t. F1 score

- Tokenize into individual words
- The capital of France is Paris → ["The", "capital", "of", "France", "is", "Paris"]
- Calculate precision / recall / F1-score
  - Precision: % overlapping tokens in predicted
  - Recall: % overlapping tokens in GT

Overall: Average between Num & F1

Quantitative Results

Performance – Proprietary

When GT pages NOT provided

- SlideAgent consistently outperforms Type 2/3 baselines (GPT-4o based) and base models across all metric.
- SlideVQA: +7.9 Overall, +8.3 Num, +6.5 F1
- SlideAgent's structured reasoning pipeline can fill the gap (77.0 → 84.9) w.r.t. strong models like Gemini-2.5 (83.8).

When GT pages provided

- Improves over the proprietary model (+7.7 overall and +12.5 numeric on SlideVQA

w/ Retrieved Pages

Model	SlideVQA			TechSlides			FinSlides		
	Overall	Num	F1	Overall	Num	F1	Overall	Num	F1
Multimodal LLMs (Type 1)									
Gemini 2.0	75.0	71.3	79.8	50.4	67.6	41.6	70.8	70.6	77.8
Gemini 2.5	<u>83.8</u>	<u>78.3</u>	**91.8**	51.1	71.4	41.2	76.2	75.8	**100.0**
Gemini 2.5-lite	71.2	60.8	87.0	47.3	58.1	41.9	57.0	56.6	68.3
Claude 4.1	78.4	74.3	82.3	61.0	<u>81.4</u>	52.3	56.5	54.8	73.3
Claude 3.5	62.5	68.3	54.6	52.5	80.2	39.5	48.5	49.5	29.6
GPT-4o	77.0	72.1	84.0	63.4	78.3	53.9	80.0	80.8	62.1
Multimodal RAG (Type 2) and Agentic Methods (Type 3)									
COLPALI	78.8	73.7	83.4	64.1	73.2	54.5	80.9	81.5	62.7
VisRAG	78.2	73.1	85.4	64.7	72.6	54.7	79.2	81.1	75.8
VDocRAG	80.0	75.0	87.8	67.0	80.5	57.0	<u>83.5</u>	<u>83.8</u>	64.2
ViDoRAG	81.1	76.4	88.1	<u>68.7</u>	78.2	<u>59.4</u>	82.2	83.3	65.1
SlideAgent	**84.9**	**80.4**	<u>90.5</u>	**70.9**	**82.5**	**66.2**	**85.5**	**85.9**	<u>79.6</u>
Impr.	+7.9	+8.3	+6.5	+7.5	+4.2	+12.3	+5.5	+5.0	+17.5

w/ Ground-truth Pages

Model	SlideVQA			TechSlides			FinSlides		
	Overall	Num	F1	Overall	Num	F1	Overall	Num	F1
Raw Models									
Gemini 2.0	86.3	81.0	90.3	59.7	60.0	<u>59.6</u>	78.1	77.8	**88.9**
Gemini 2.5	**89.0**	**85.7**	**93.2**	61.5	65.0	59.5	76.6	76.4	83.3
Gemini 2.5-lite	81.8	75.5	90.1	56.3	55.0	57.0	78.1	78.4	66.7
Claude 4.1	85.7	82.4	89.7	58.0	77.5	48.3	52.6	52.0	60.8
Claude 3.5	58.2	64.0	50.9	55.8	**83.7**	42.5	47.8	48.0	40.3
GPT-4o	79.4	71.9	86.4	64.5	77.1	58.0	83.0	84.3	80.1
Multimodal RAG and Agentic Methods									
ViDoRAG	81.8	73.8	87.0	<u>65.8</u>	78.0	58.7	<u>84.2</u>	<u>84.7</u>	81.5
SlideAgent	<u>87.1</u>	<u>84.4</u>	<u>90.6</u>	**68.7**	<u>82.5</u>	**61.5**	**85.8**	**85.9**	<u>85.6</u>
Impr.	+7.7	+12.5	+4.2	+4.1	+5.4	+3.5	+2.8	+1.5	+5.5

Performance – Open-source

- **Strong Gains**: +9.8 overall and +11.7 numeric over InternVL3-8B.
- Outperforms open-source models (except for Qwen2.5)
- **Model-agnostic**: can further advance LLMs like Gemini-2.5 and Qwen2.5.

w/ Retrieved Pages

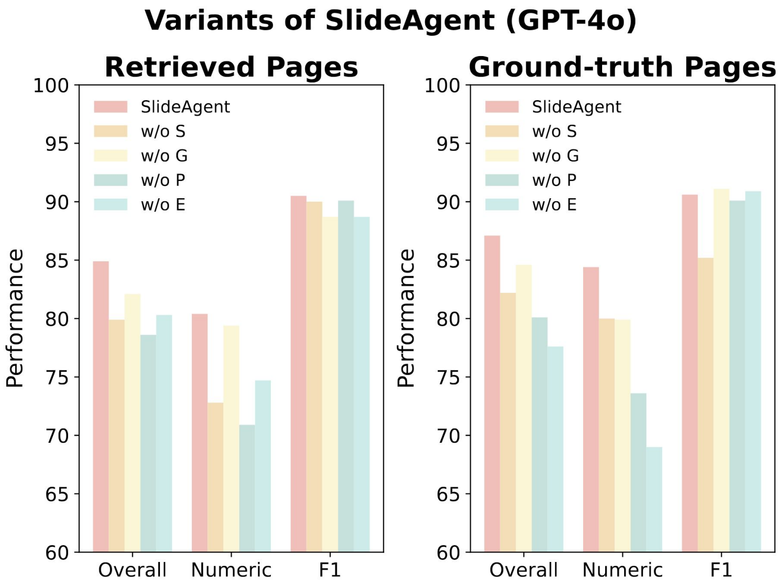
Model	SlideVQA			TechSlides			FinSlides		
	Overall	Num	F1	Overall	Num	F1	Overall	Num	F1
Multimodal LLMs (Type 1)									
Llama 3.2 11B	42.9	43.3	42.3	41.4	52.5	36.2	23.3	23.2	26.2
Phi3	72.3	61.8	90.6	59.4	60.0	59.1	48.8	48.5	64.3
Qwen2.5 7B	**79.5**	70.5	**94.3**	59.3	52.5	**65.9**	53.6	52.5	85.7
Qwen2.5 32B	79.2	**71.1**	92.2	**67.5**	**87.5**	60.6	57.4	56.6	**87.5**
LLaVA 1.5 7B	36.8	22.4	79.0	23.3	12.5	38.7	10.7	10.1	16.6
LLaVA 1.5 13B	44.9	25.1	81.8	28.1	17.5	45.0	20.6	16.5	36.7
LLaVA 1.6 7B	50.9	37.3	82.6	34.4	37.5	32.2	12.2	12.1	17.8
LLaVA 1.6 13B	16.7	10.2	81.5	45.2	40.0	49.1	32.0	31.3	64.3
InternVL3 8B	63.0	56.5	74.1	55.4	57.5	54.4	49.8	49.5	64.3
Multimodal RAG (Type 2) and Agentic Method (Type 3)									
COLPALI	63.4	56.7	73.8	57.1	60.9	55.2	50.4	49.3	65.7
VisRAG	63.6	56.5	75.5	56.8	57.7	55.4	51.1	49.6	65.2
VDocRAG	65.2	59.7	77.0	59.2	60.7	58.3	51.8	50.1	65.9
ViDoRAG	68.8	61.9	77.3	61.4	61.9	59.3	52.7	55.4	66.6
SlideAgent	72.7	68.2	79.4	63.1	78.0	61.7	**63.3**	**62.8**	68.3
Impr.	+9.8	+11.7	+5.4	+7.7	+20.5	+2.3	+13.5	+13.3	+4.0

w/ Ground-truth Pages

Model	SlideVQA			TechSlides			FinSlides		
	Overall	Num	F1	Overall	Num	F1	Overall	Num	F1
Raw Models									
Llama 3.2 11B	44.6	52.1	34.6	47.1	62.8	39.3	39.1	39.2	33.5
Phi3	78.3	69.1	91.6	53.9	67.4	47.2	63.8	63.7	65.1
Qwen2.5 7B	85.1	77.7	**95.3**	57.7	69.8	51.8	52.7	52.0	**77.8**
Qwen2.5 32B	**87.4**	**82.6**	93.7	49.6	62.8	43.2	**69.5**	**69.6**	65.1
LLaVA 1.5 7B	42.9	27.9	79.9	24.6	15.0	39.4	14.2	14.4	20.9
LLaVA 1.5 13B	46.7	29.5	83.1	29.2	17.5	46.6	23.8	20.3	41.2
LLaVA 1.6 7B	59.0	45.8	84.2	36.2	40.0	34.0	16.0	15.2	21.3
LLaVA 1.6 13B	62.3	48.2	87.1	54.7	62.5	49.5	35.2	30.7	67.6
InternVL3 8B	73.3	65.4	85.7	58.4	72.1	51.5	56.3	55.9	65.6
Baseline methods based on InternVL3-8B									
ViDoRAG	76.3	68.1	89.9	61.3	75.1	53.9	58.4	58.7	67.3
SlideAgent	82.8	75.3	93.3	**64.6**	**79.5**	**58.0**	62.8	62.6	68.1
Impr.	+9.5	+9.8	+7.6	+6.2	+7.4	+6.4	+6.5	+6.7	+2.5

Ablation Studies

- **w/o P (Page)**: –6.3 overall (–9.5 numeric) and InternVL3-8B –8.8 overall.
- Page knowledge is important as it integrates global themes $\mathcal{K}_g$ and sequential context $\mathcal{K}_p^{i-1}$
- **w/o E (Element)**: –4.6 overall for GPT-4o; –6.3 for InternVL3-8B.
- **w/o G (Global)**: minimal impact
- **w/o S (Subquery)**: larger losses under retrieval (–5.0 GPT-4o; –11.3 InternVL3-8B) than with ground-truth pages



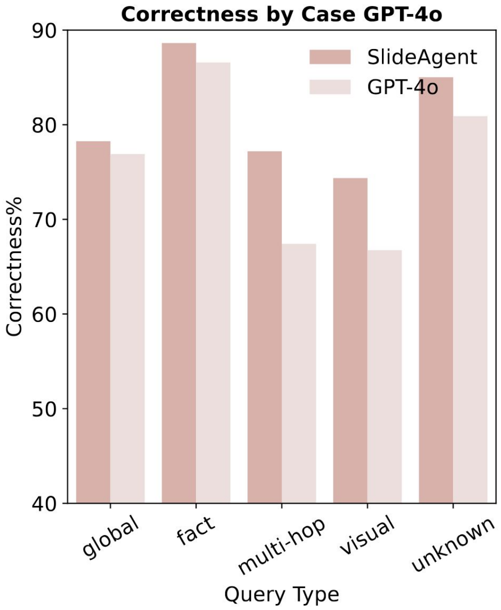
Performance – Retrieval

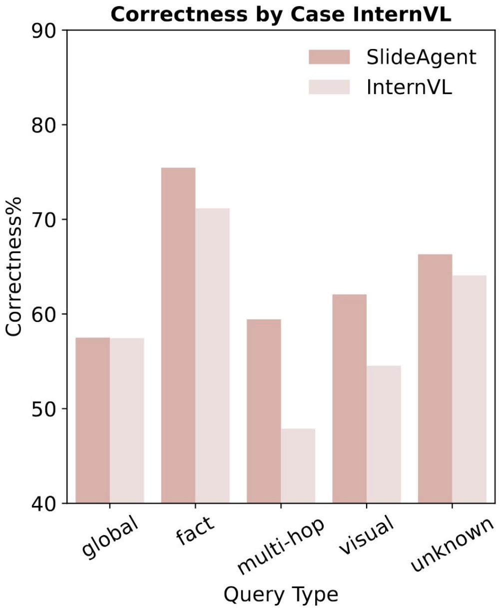
Text-based Retrievers	MRR	Recall@1	nDCG@1	Recall@3	Hit@3	nDCG@3
BM25 (Robertson et al., 2004)	59.0	51.5	52.0	63.0	65.3	57.7
w/ SA	63.9 +4.9	54.9 +3.4	56.6 +4.6	67.1 +4.1	68.8 +3.5	62.6 +4.9
BGE (Xiao et al., 2023)	70.1	56.1	60.9	75.8	78.1	69.2
w/ SA	72.3 +2.2	58.4 +2.3	63.2 +2.3	77.4 +1.6	80.3 +2.2	71.3 +2.1
SFR (Meng et al., 2024)	70.1	56.1	60.9	75.8	78.1	69.2
w/ SA	76.5 +6.4	59.9 +3.8	69.4 +8.5	77.3 +1.5	81.8 +3.7	73.2 +4.0
Multimodal Retrievers	**MRR**	**Recall@1**	**NDCG@1**	**Recall@3**	**Hit@3**	**NDCG@3**
SigLIP2 (Tschannen et al., 2025)	26.7	15.9	18.0	31.0	34.0	25.0
w/ SA	28.0 +1.3	16.3 +0.4	18.0 +0.0	32.3 +1.3	35.5 +1.5	26.1 +1.1
COLPALI (Faysse et al., 2024)	82.1	68.2	75.5	78.9	84.0	77.4
w/ SA	82.9 +0.8	70.4 +2.2	76.2 +0.7	88.6 +9.7	90.1 +6.1	83.1 +5.7
VisRAG (Yu et al., 2025)	76.0	63.3	68.6	82.3	84.1	76.0
w/ SA	79.7 +3.7	66.3 +3.0	71.6 +3.0	85.5 +3.2	87.7 +3.6	79.4 +3.4

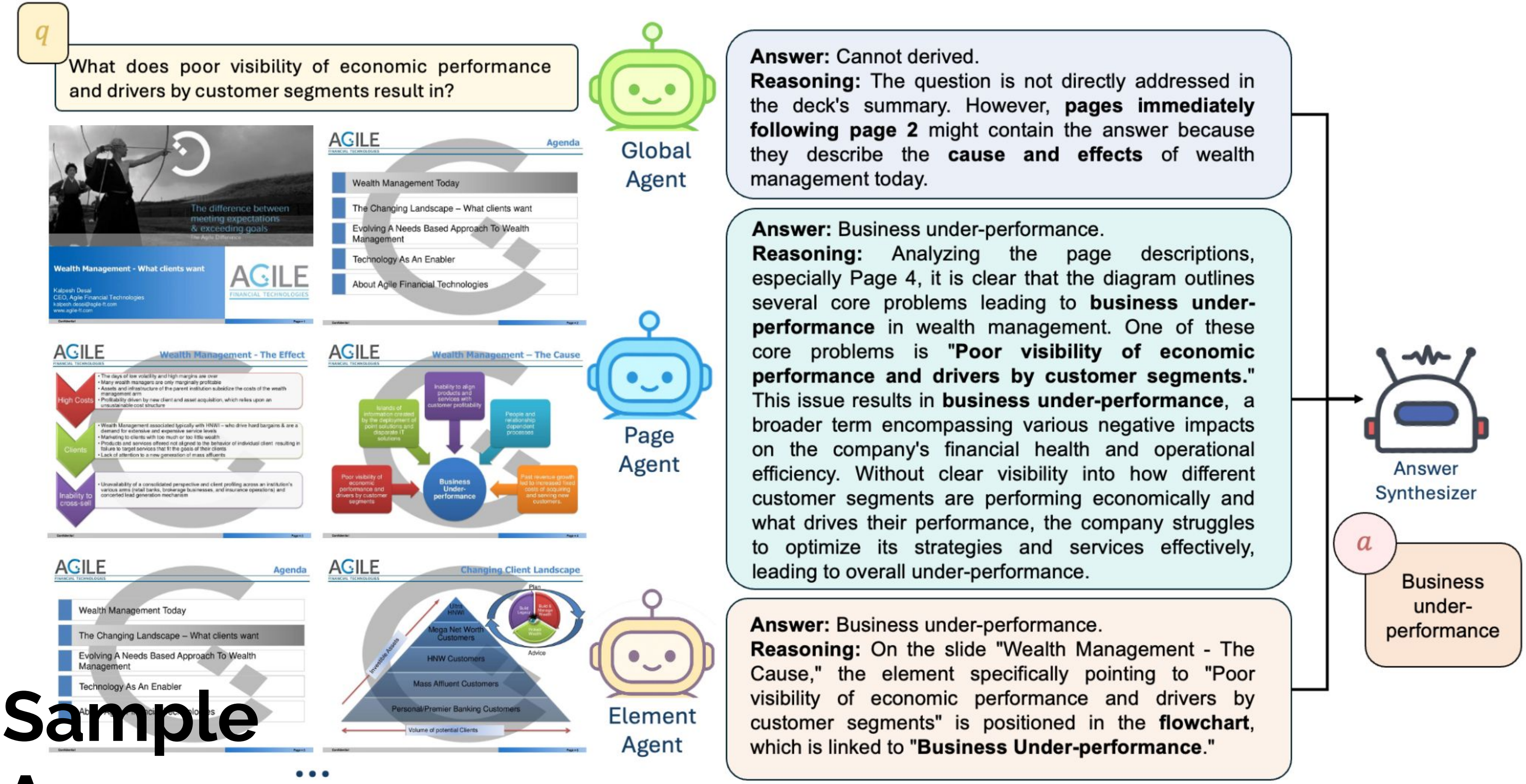
- Can the constructed knowledge $\mathscr{K}_p$ enhance text-based / multimodal retrieval?
- Text-based retrievers show largest gains, rivaling multimodal retrievers.
- Multimodal Retrievers have smaller but consistent gains.

Correctness by Cases

- Particular boost in multi-hop reasoning (Case 3, +9.8 from 67.4 to 77.2) and visual/layout questions (Case 4, +7.7).
- Modest gain in global understanding (Case 1, +2.1)







Sample Answer

Global agent gives a general scope. Page/element agent performs finer-grained reasoning.

Conclusion

- **Framework**: SlideAgent, a hierarchical agentic framework
- **Insight**: Element-level parsing & reasoning is crucial
- **Future Work**: address user queries in an iterative manner.
- Accepted at NeurIPS 2025 Workshop on Multimodal Algorithmic Reasoning (MAR 2025)
- Under review at ACL ARR October.
- Thanks to my mentor Rachneet, my manager Zhen, Sumitra, and my collaborators Nishan and Kelly.
- Special thanks to J.P. Morgan AI Research and to everyone in the team!!

Accepted at NeurIPS 2025 Workshop on
Multimodal Algorithmic Reasoning
MAR 2025

Thanks

SlideAgent: Hierarchical Agentic Framework for Multi-Page Slide Deck Understanding

J.P.Morgan
Georgia Institute of Technology