# Retrieval and Reasoning for Long Multimodal Document QA

2026. 03. 19.

양 준성

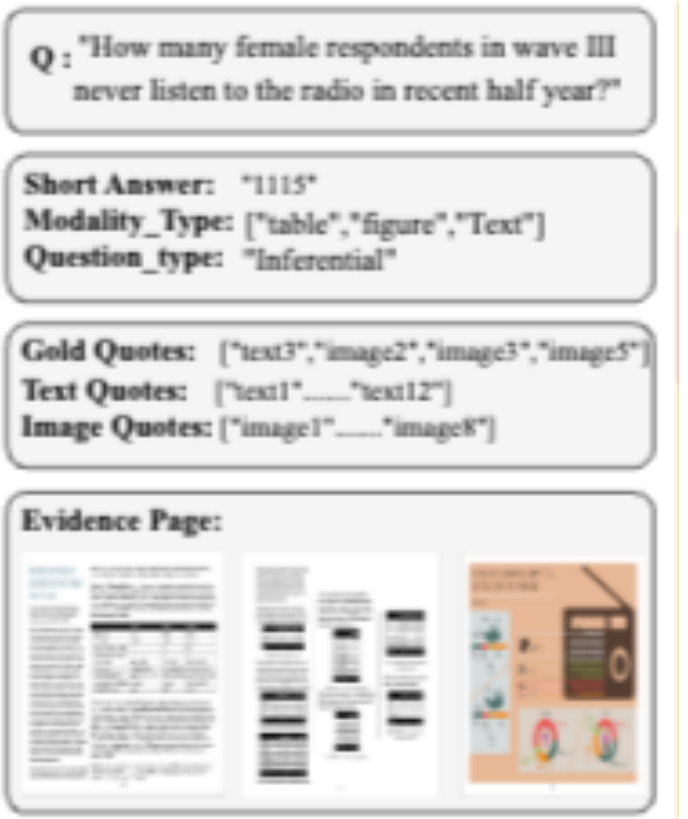
Natural Language Processing & Artificial Intelligence

고려대학교 KOREA UNIVERSITY

# 1. Motivation

❖ Why is multimodal document QA hard?

- Evidence is scattered across many pages
- Information appears in text, tables, charts, and layout
- Reading the whole document is noisy and expensive
- So the key is: retrieve the right pages before reasoning



# MDocAgent: A Multi-Modal Multi-Agent Framework for Document Understanding

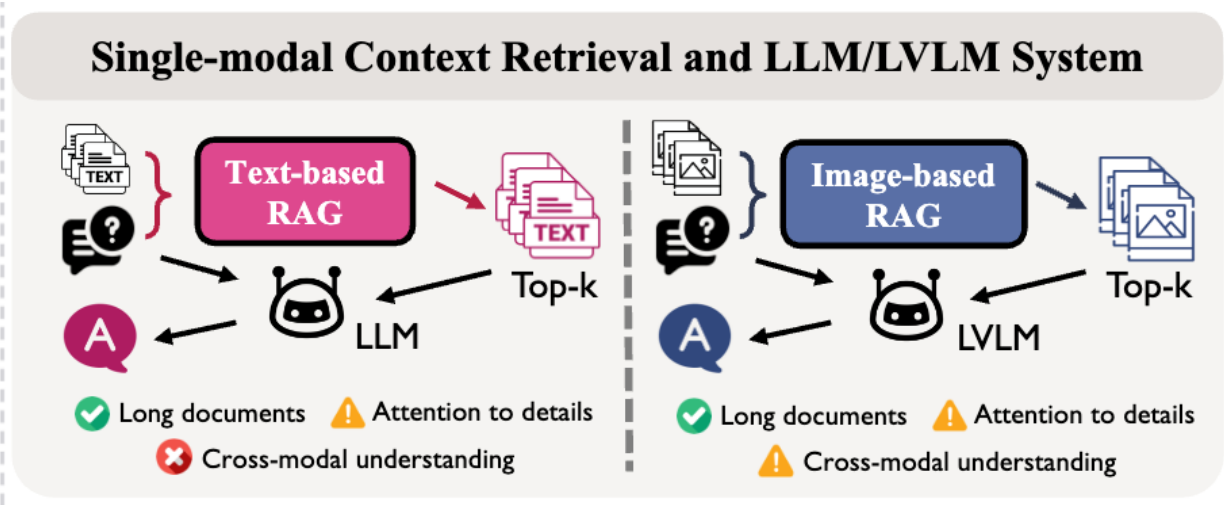
Siwei Han[1], Peng Xia[1], Ruiyi Zhang[2], Tong Sun[2], Yun Li[1], Hongtu Zhu[1], Huaxiu Yao[1]
[1]UNC-Chapel Hill, [2]Adobe Research
{siweih,huaxiu}@cs.unc.edu

Natural Language Processing & Artificial Intelligence

고려대학교 KOREA UNIVERSITY

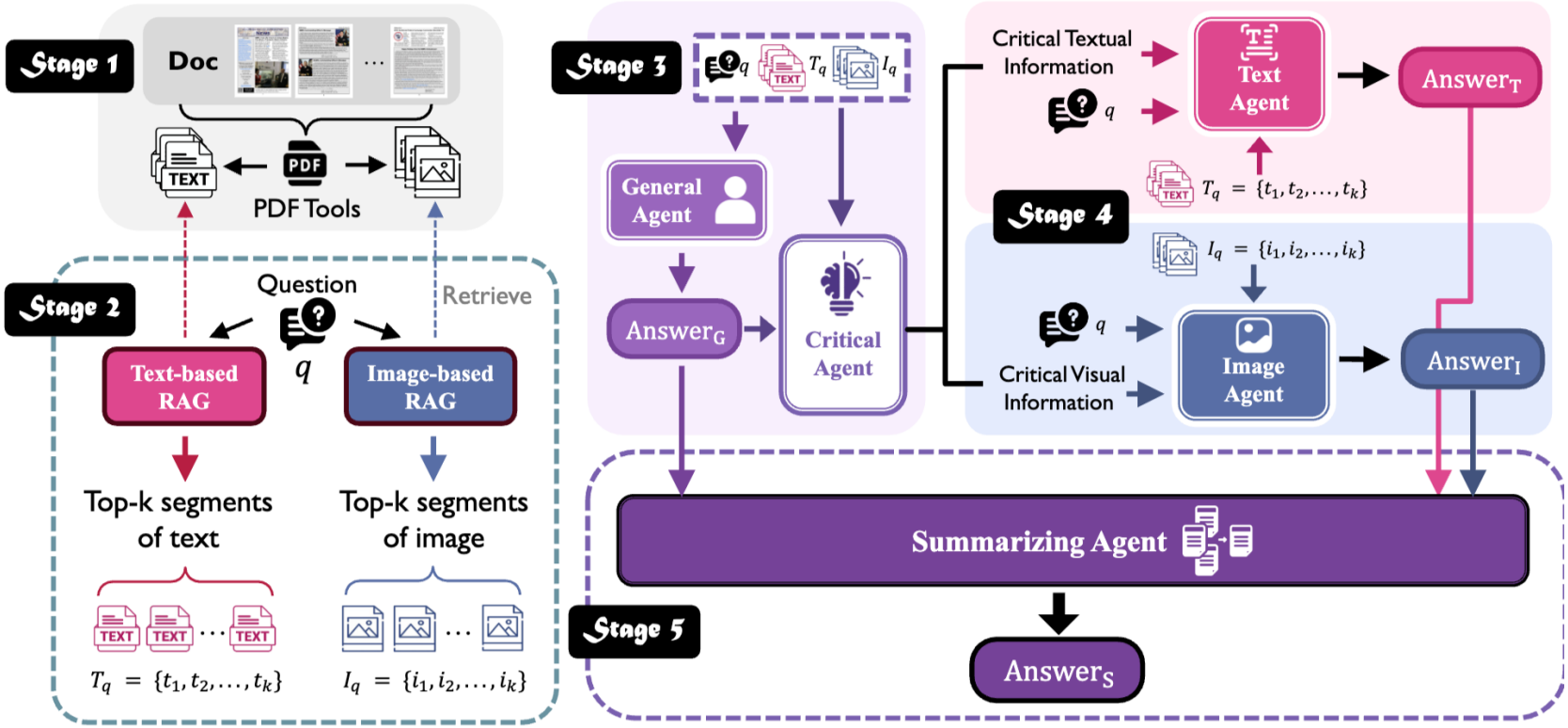
# 2. MDocAgent: Why multi-agent?

- A single pass often misses cross-modal evidence
- Documents contain both textual and visual clues
- Different subtasks require different reasoning roles
- So MDocAgent decomposes the process into multiple specialized agents

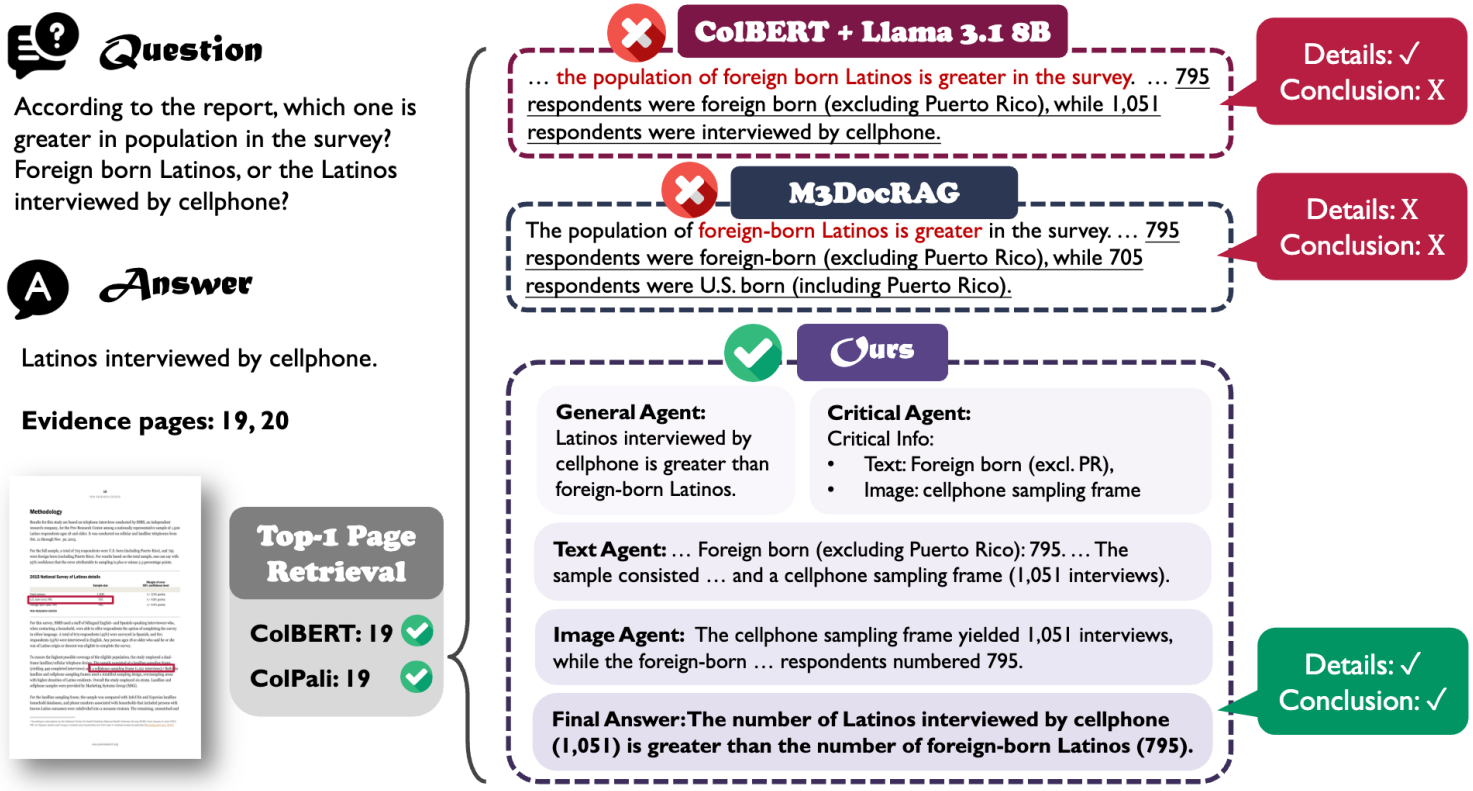


# 2. MDocAgent

MDocAgent: Multi-agent RAG for long document understanding



# 2. MDocAgent




Qualitative example

## Experimental Setup

- **5 benchmarks:** MMLongBench, LongDocURL, PaperTab, PaperText, FetaTab
- **5-agent framework:** general / critical / text / image / summarizing
- **Models:** Llama-3.1-8B-Instruct (text agent), Qwen2-VL-7B-Instruct (other agents)
- **Retrieval:** ColBERTv2 for text, ColPali for image
- **Settings:** top-1 and top-4 retrieval
- **Evaluation:** GPT-4o judge, average accuracy

# QA performance

| Method | MMLongBench | LongDocUrl | PaperTab | PaperText | FetaTab | Avg |
|---|---|---|---|---|---|---|
| *LVLMs* | | | | | | |
| Qwen2-VL-7B-Instruct [38] | 0.165 | 0.296 | 0.087 | 0.166 | 0.324 | 0.208 |
| Qwen2.5-VL-7B-Instruct [2] | 0.224 | 0.389 | 0.127 | 0.271 | 0.329 | 0.268 |
| LLaVA-v1.6-Mistral-7B [22] | 0.099 | 0.074 | 0.033 | 0.033 | 0.110 | 0.070 |
| Phi-3.5-Vision-Instruct [1] | 0.144 | 0.280 | 0.071 | 0.165 | 0.237 | 0.179 |
| LLaVA-One-Vision-7B [19] | 0.053 | 0.126 | 0.056 | 0.108 | 0.077 | 0.084 |
| SmolVLM-Instruct [27] | 0.081 | 0.163 | 0.066 | 0.137 | 0.142 | 0.118 |
| *RAG methods (top 1)* | | | | | | |
| ColBERTv2 [32]+LLaMA-3.1-8B [12] | 0.241 | 0.429 | 0.155 | 0.332 | 0.490 | 0.329 |
| M3DocRAG [5] (ColPali [9]+Qwen2-VL-7B [38]) | 0.276 | 0.506 | 0.196 | 0.342 | 0.497 | 0.363 |
| **MDocAgent (Ours)** | **0.299** | **0.517** | **0.219** | **0.399** | **0.600** | **0.407** |
| *RAG methods (top 4)* | | | | | | |
| ColBERTv2 [32]+LLaMA-3.1-8B [12] | 0.273 | 0.491 | 0.277 | 0.460 | 0.673 | 0.435 |
| M3DocRAG [5] (ColPali [9]+Qwen2-VL-7B [38]) | 0.296 | 0.554 | 0.237 | 0.430 | 0.578 | 0.419 |
| **MDocAgent (Ours)** | **0.315** | **0.578** | **0.278** | **0.487** | **0.675** | **0.465** |

=> MDocAgent improves average accuracy by 12.1% over prior SOTA through five specialized agents.

# SimpleDoc: Multi-Modal Document Understanding with Dual-Cue Page Retrieval and Iterative Refinement

**Chelsi Jain[1,*], Yiran Wu[2,*], Yifan Zeng[1,3,*], Jiale Liu[2],**
**Shengyu Dai[4], Zhenwen Shao[4], Qingyun Wu[2,3], Huazheng Wang[1,3]**

**EMNLP 2025 Main**

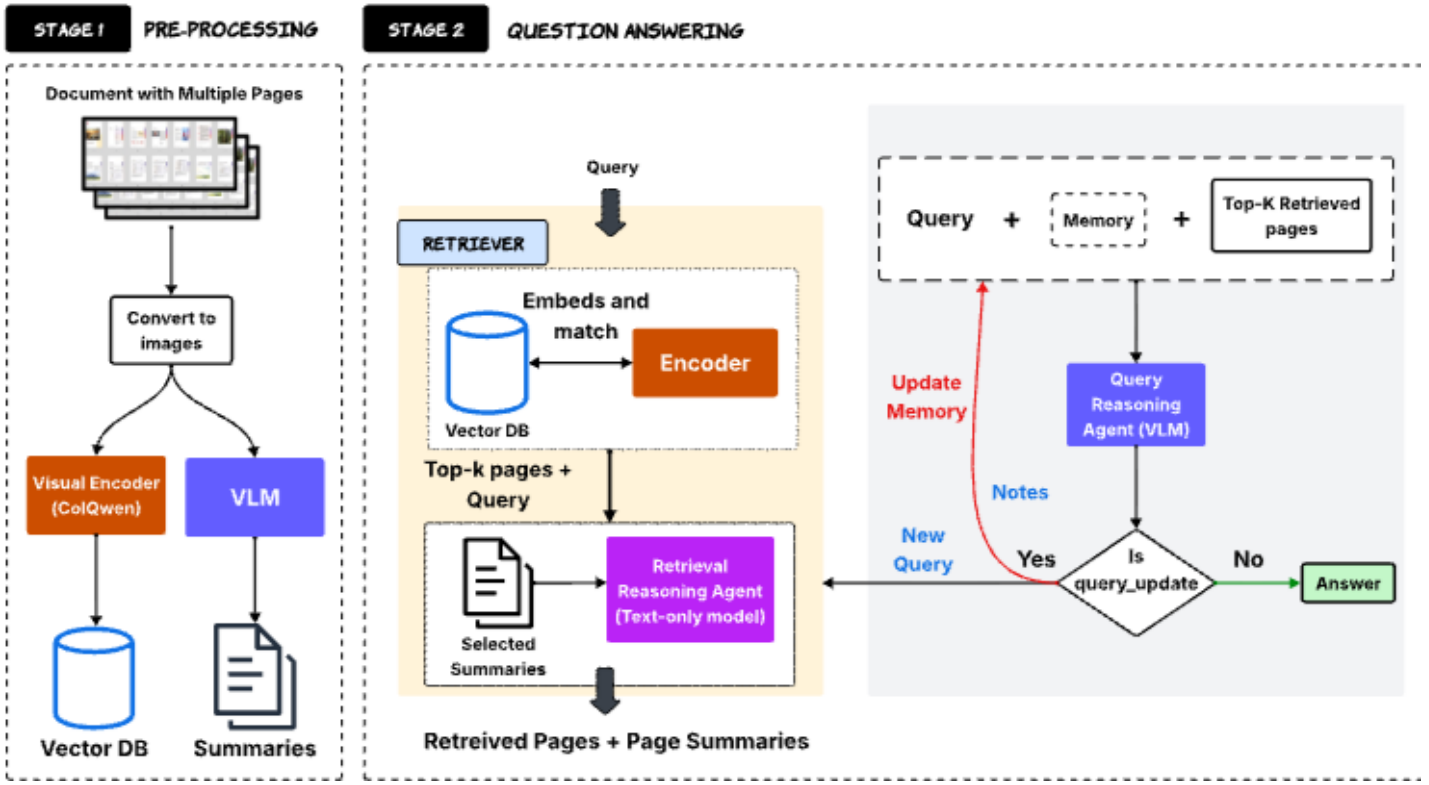
Natural Language Processing & Artificial Intelligence

KOREA UNIVERSITY 고려대학교 KOREA UNIVERSITY

# 3. SimpleDoc: Why simplify the pipeline?

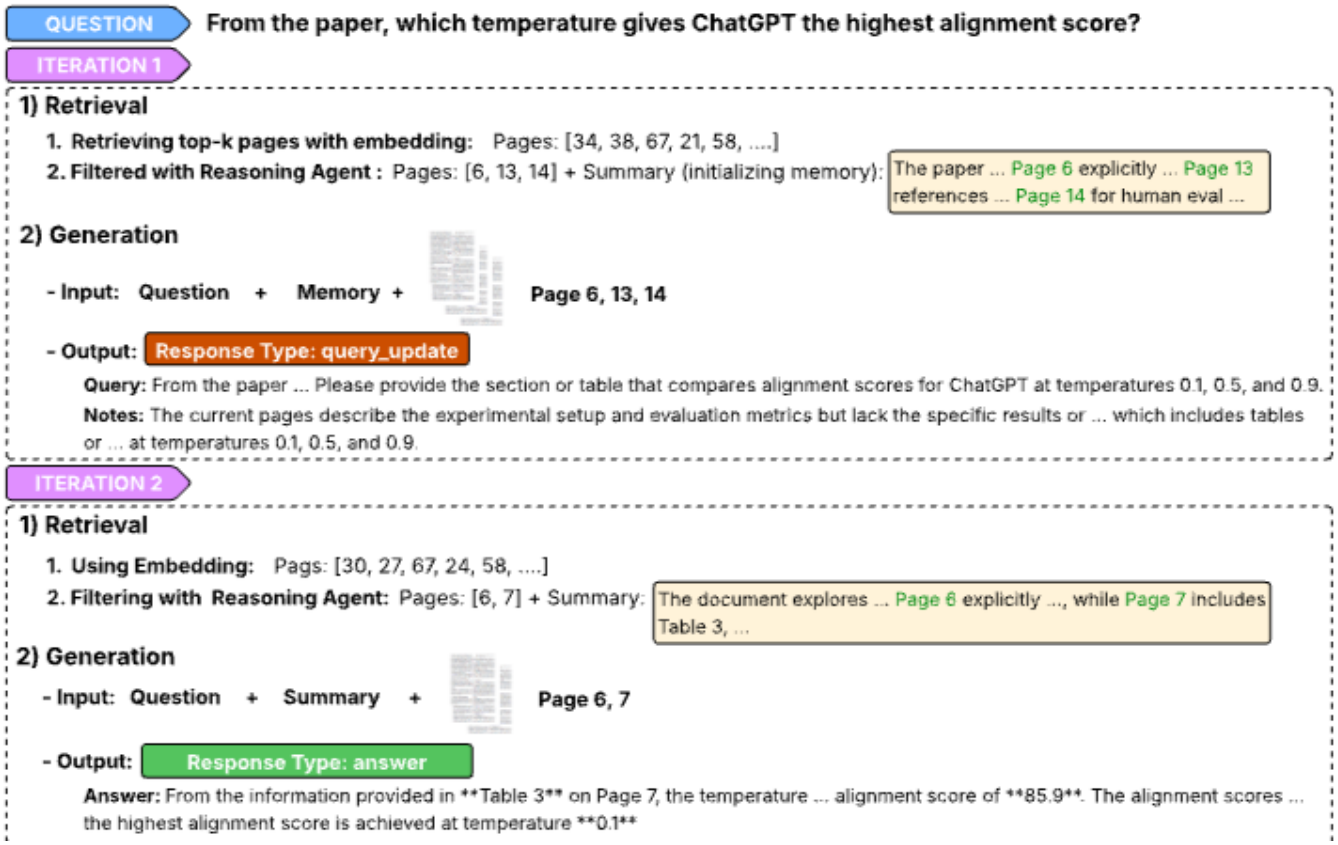
- Multi-agent systems are powerful but complex
- Initial retrieval is often noisy
- We need a lightweight way to select better pages
- So SimpleDoc uses dual-cue retrieval + iterative refinement

# 3. SimpleDoc




Offline page summaries + dual-cue retrieval + iterative reasoning

# 3. SimpleDoc




Iterative retrieval

*if the first retrieval is insufficient => the model refines the query and retrieves again*

# Experimental Setup

**Dataset**:

- **MMLongBench**: 1,073 queries/135 docs, 47.5 pages/doc
- **LongDocURL**: 33,000 pages, 2,325 queries.
- **PaperTab**: 393 queries / 307 paper.
- **FetaTab**: 1,023 queries / 878 document/table.

**Baseline:** M3DocRAG, MDocAgent

**Visual retrieval backbone**: ColQwen-2.5

**VLM generation:** Qwen2.5-VL-32B-Instruct

**Evaluate QA:** GPT-4.1

**Algorithm 1** SimpleDoc

**Require:** query $q$, per–page embeddings $E$ and summaries $S$, cutoff $k$, max iterations $L$

**Ensure:** answer $a$ or failure notice

$q_{\text{cur}} \leftarrow q$
$M \leftarrow \varnothing$
**for** $\ell \leftarrow 1$ **to** $L$ **do**
  $s_{\text{DOC}}, C \leftarrow \mathit{RetrievePages}(q_{\text{cur}}, E, S, k)$
  $I_C \leftarrow \{\, i_c \mid c \in C \}$; $T_C \leftarrow \{\, t_c \mid c \in C \}$
  $M \leftarrow M \cup s_{\text{DOC}}$
  $(\mathit{is\_solved}, a, m', q') \leftarrow$
    $\textsc{Reasoner}(q, I_C, T_C, M)$
  **if** *is_solved* **then**
    **return** $a$
  **else**
    $M \leftarrow M \cup \{m'\}$
    $q_{\text{cur}} \leftarrow q'$
**return** FAIL

## Retrieval performance

| Method | Avg Ret. Pages | All Hit % | F1 Score |
|---|---|---|---|
| ColQwen-2.5 | 2 | 64.12 | 38.75 |
| ColQwen-2.5 | 6 | 76.42 | 24.36 |
| ColQwen-2.5 | 10 | **83.60** | 18.38 |
| **Ours** (top-10) | 3.19 | 65.72 | 61.42 |
| **Ours** (top-30) | 3.46 | 67.37 | **62.22** |

⇒ summary-based re-ranking improves retrieval. precision while keeping the number of read pages low.

| Model | Top-K | Match Rate % |
|---|---|---|
| ColQwen-2.5 | 2 | 54.55 |
| ColPali | 2 | 28.74 |
| ColQwen-2.5 | 6 | 70.13 |
| ColPali | 6 | 44.35 |
| ColQwen-2.5 | 10 | 79.22 |
| ColPali | 10 | 55.15 |

=> Better evidence selection leads to better answer quality.

## Generate performance

| Method | Pg. Ret. | MMLongBench | LongDocUrl | PaperTab | FetaTab | Avg. Acc |
|---|---|---|---|---|---|---|
| | | *LVMs* | | | | |
| Qwen2.5-VL-32B-Instruct | – | 22.18 | 19.78 | 7.12 | 16.14 | 16.31 |
| Qwen2.5-VL-32B-Instruct + Ground-Truth pages | – | 67.94 | 30.80 | - | - | - |
| | | *RAG methods (top 2)* | | | | |
| M3DocRAG (Qwen2.5-VL-32B) | 2 | 41.8 | 50.7 | 50.1 | 75.2 | 54.4 |
| MDocAgent (Qwen3-30B + Qwen2.5-VL-32B) | 4 | 50.6 | 56.8 | 50.9 | 80.3 | 59.6 |
| | | *RAG methods (top 6)* | | | | |
| M3DocRAG (Qwen2.5-VL-32B) | 6 | 41.8 | 53.1 | 60.1 | 79.8 | 58.7 |
| MDocAgent (Qwen3-30B + Qwen2.5-VL-32B) | 12 | 55.3 | 63.2 | 64.9 | **84.5** | 66.9 |
| | | *RAG methods (top 10)* | | | | |
| M3DocRAG (Qwen2.5-VL-32B) | 10 | 39.7 | 52.2 | 56.7 | 78.6 | 56.8 |
| MDocAgent (Qwen3-30B + Qwen2.5-VL-32B) | 20 | 54.8 | 61.9 | 63.1 | 84.1 | 65.9 |
| | | *Ours (top-10 and top-30)* | | | | |
| **SimpleDoc** (Qwen3-30B + Qwen2.5-VL-32B) | 3.2 | 59.55 | 72.26 | 64.38 | 80.31 | 69.12 |
| **SimpleDoc** (Qwen3-30B + Qwen2.5-VL-32B) | 3.5 | **60.58** | **72.30** | **65.39** | 82.19 | **70.12** |

=> SimpleDoc achieves the best average accuracy with only ~3.5 pages per question

## Generate performance

| Method | Evidence Source | | | | | Evidence Page | | | ACC |
|---|---|---|---|---|---|---|---|---|---|
| | TXT | LAY | CHA | TAB | FIG | SIN | MUL | UNA | |
| *Qwen2.5-VL-7B-Instruct + Qwen-3-8B* | | | | | | | | | |
| VLM + GT pages | 51.32 | 45.38 | 37.71 | 40.09 | 47.83 | 58.90 | 35.01 | 77.97 | 54.99 |
| M3DocRAG (top-6) | 43.21 | 39.98 | 36.05 | 31.60 | 42.01 | 55.46 | 24.78 | 8.72 | 35.50 |
| MDocAgent (top-6) | 47.04 | 38.98 | **47.09** | **41.04** | 39.93 | **59.45** | 28.57 | 33.49 | 43.80 |
| **Ours** | **49.67** | **42.02** | 44.57 | 37.79 | **42.14** | 58.69 | **31.65** | **62.11** | **50.42** |
| *Qwen2.5-VL-32B-Instruct + Qwen-3-30B-A3B* | | | | | | | | | |
| VLM + GT pages | 63.25 | 66.39 | 58.86 | 65.44 | 57.53 | 72.60 | 55.46 | 77.53 | 67.94 |
| M3DocRAG | 46.69 | 41.53 | 45.35 | 39.15 | 43.75 | 58.61 | 30.61 | 22.48 | 41.80 |
| MDocAgent | 57.49 | 50.00 | 54.65 | **56.13** | **52.78** | 68.70 | **42.86** | 45.41 | 55.30 |
| Chain-of-Notes† | 36.75 | 35.29 | 38.46 | 32.26 | 33.44 | 49.59 | 21.69 | 50.00 | 40.45 |
| Plan*RAG† | 46.03 | 36.13 | 43.75 | 38.71 | 37.12 | 54.88 | 25.35 | 23.89 | 38.58 |
| **Ours** | **59.93** | **51.26** | **54.86** | 51.15 | 51.17 | **70.76** | 39.22 | **67.40** | **59.55** |

=> SimpleDoc remains competitive even with smaller models.

## Generate performance

| Top-k | Avg. Page Used | Acc. |
|---|---|---|
| 2 | 2.15 | 56.66 |
| 6 | 2.75 | 58.25 |
| 10 | 3.19 | 59.55 |
| 30 | 3.46 | 60.58 |

| Setting | QA Accuracy |
|---|---|
| SimpleDoc (with memory) | 59.55 |
| SimpleDoc (without memory) | 59.27 |

| Iteration | 1 | 2 | 3 |
|---|---|---|---|
| Accuracy | 58.62 | 59.27 | 59.55 |
| # Query Update | 182 | 121 | 97 |

❖ Same task, different design philosophy

**MDocAgent**

- multi-agent collaboration
- separate text/image reasoning roles
- stronger modularity, higher complexity

**SimpleDoc**

- one reasoner + dual-cue retriever
- summary-based reranking
- lighter pipeline, fewer pages read

# Discussion and Conclusion

- Long document QA is mainly a retrieval problem before it is a generation problem
- MDocAgent improves reasoning by decomposing roles across agents
- SimpleDoc improves efficiency by making retrieval more selective and iterative
- The key trade-off is not accuracy alone, but reasoning structure versus system simplicity.

# Thank you