NeuSym-RAG: Hybrid Neural Symbolic Retrieval with Multiview Structuring for PDF Question Answering

Ruisheng Cao*, Hanchong Zhang*, Tiancheng Huang*, Zhangyi Kang,
Yuxin Zhang, Liangtai Sun, Hanqi Li, Yuxun Miao, Shuai Fan, Lu Chen, and Kai Yu

SJTU Cross Media Language Intelligence Lab
上海交通大学跨媒体语言智能实验室

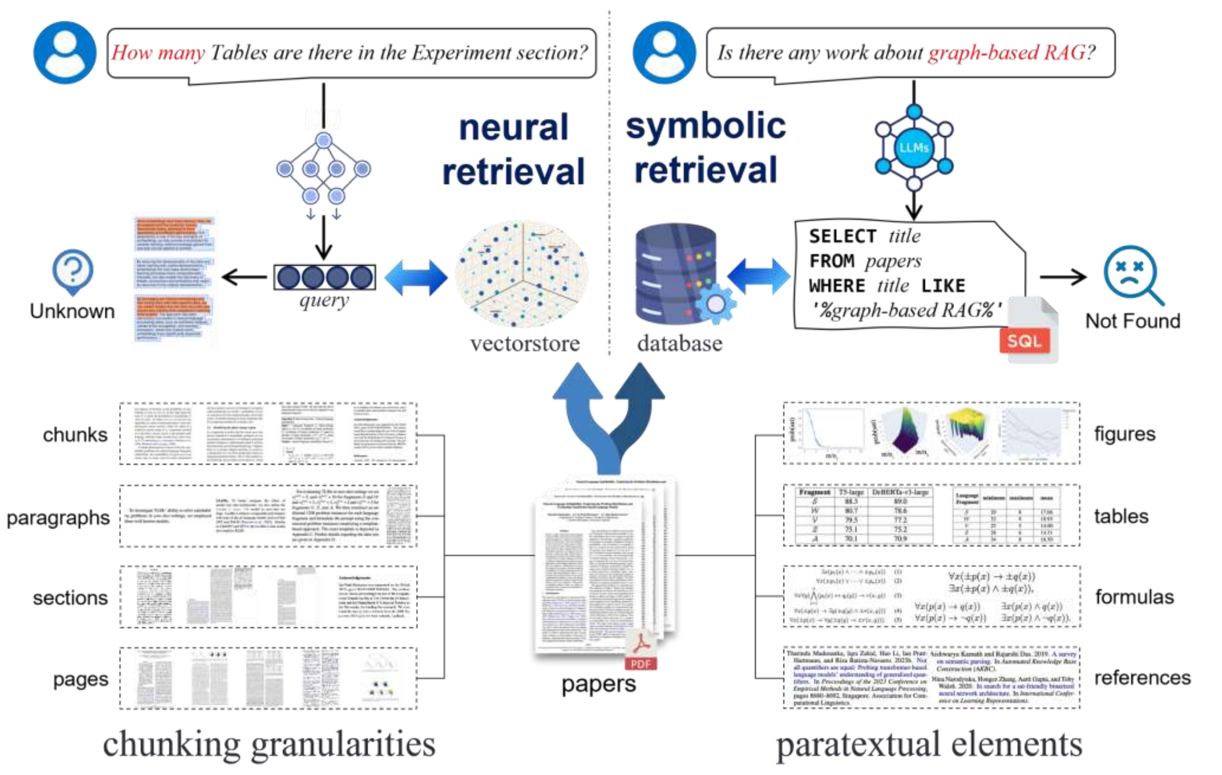
ACL 2025 VIENNA JULY 27 - AUGUST 1

Structural indexing with database schema • Iterative retrieval with hybrid paradigms • Realistic QA dataset *w.r.t.* AI research

Motivation

With the exponential growth in academic papers, RAG-based QA systems show great potential to help researchers extract key details from emerging studies. In this work, we propose:

- **Integration of vector-based neural retrieval and SQL-based symbolic retrieval.** The classic neural retrieval often fails when handling precise queries, while symbolic retrieval breaks down in semantic fuzzy matching or morphological variations.
- **Incorporation of multiple views for parsing and vectorizing PDF documents.** Commonly utilized scheme to segment documents into chunks is based on a fixed length of consecutive tokens, neglecting the intrinsic structure and the salient features of paratextual tables and figures.



Our paper on ArXiv

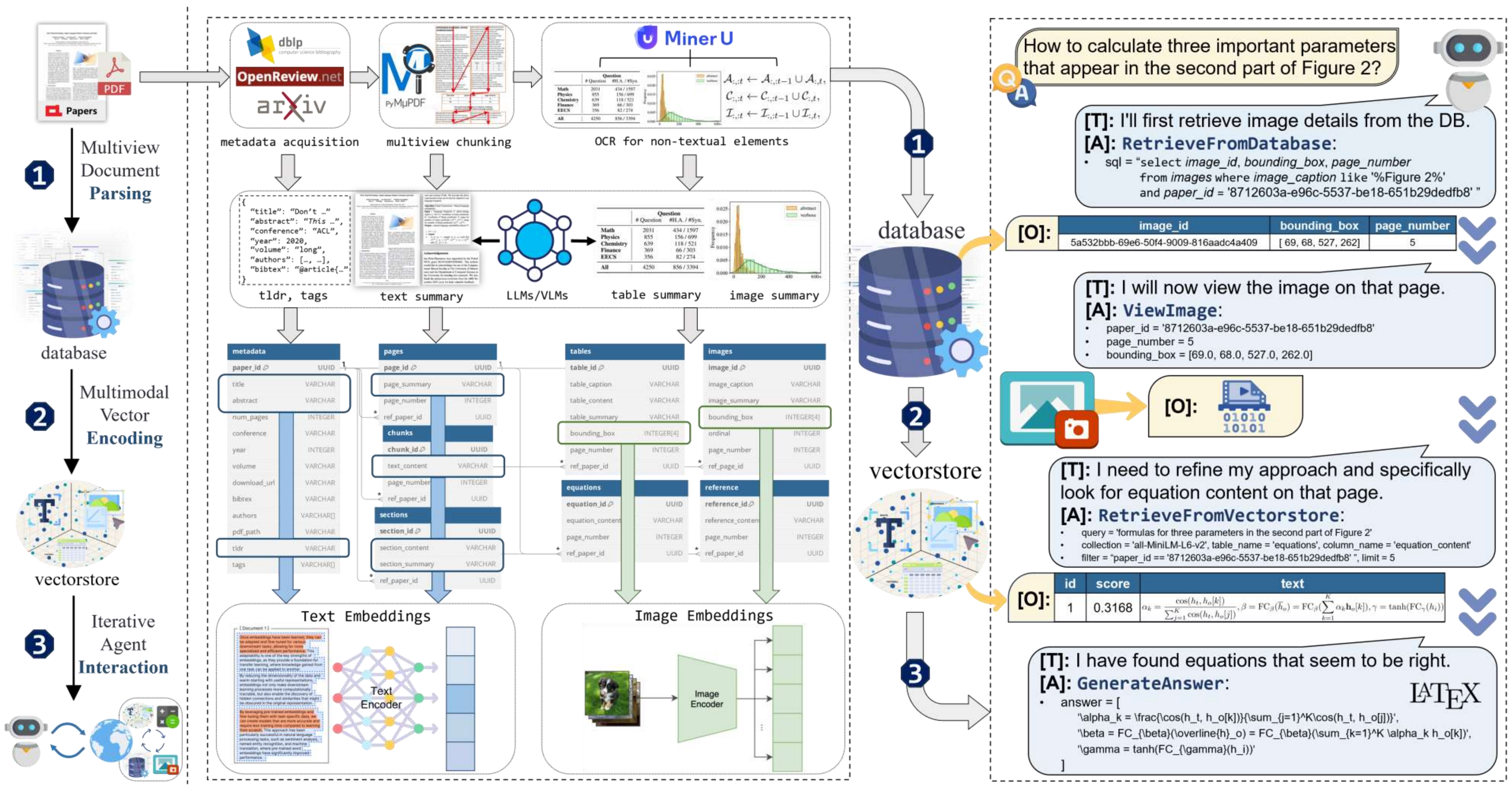
Method

Our entire workflow proceeds as follows:

- **Parsing.** Firstly, we segment the PDF in **multi-view**, extract **non-textual elements**, and store them in a schema-constrained database.
- **Encoding:** Next, we identify those **encodable columns** in the DB, obtain and insert vectors of cell values into the vectorstore.
- **Interaction:** Finally, we build an **iterative** Q&A agent which can predict **executable actions** to retrieve context and answer the input question.

```
RetrieveFromVectorstore(
    # user input can be rephrased
    query: str,
    # select encoding model/modality
    collection_name: str,
    # (table_name, column_name) together
        defines which view to search
    table_name: str,
    column_name: str,
    # allow fine-grained meta filtering
    filter: str = '',
    limit: int = 5
)
```

an example of the parameterized action



3 stages of NeuSym-RAG: multi-view parsing → multi-modal encoding → agentic interaction

Experiment

- Manually annotated PDF-based scholar QA dataset **AirQA-Real**
  - 553 questions + 3 task types + instance-specific evaluation

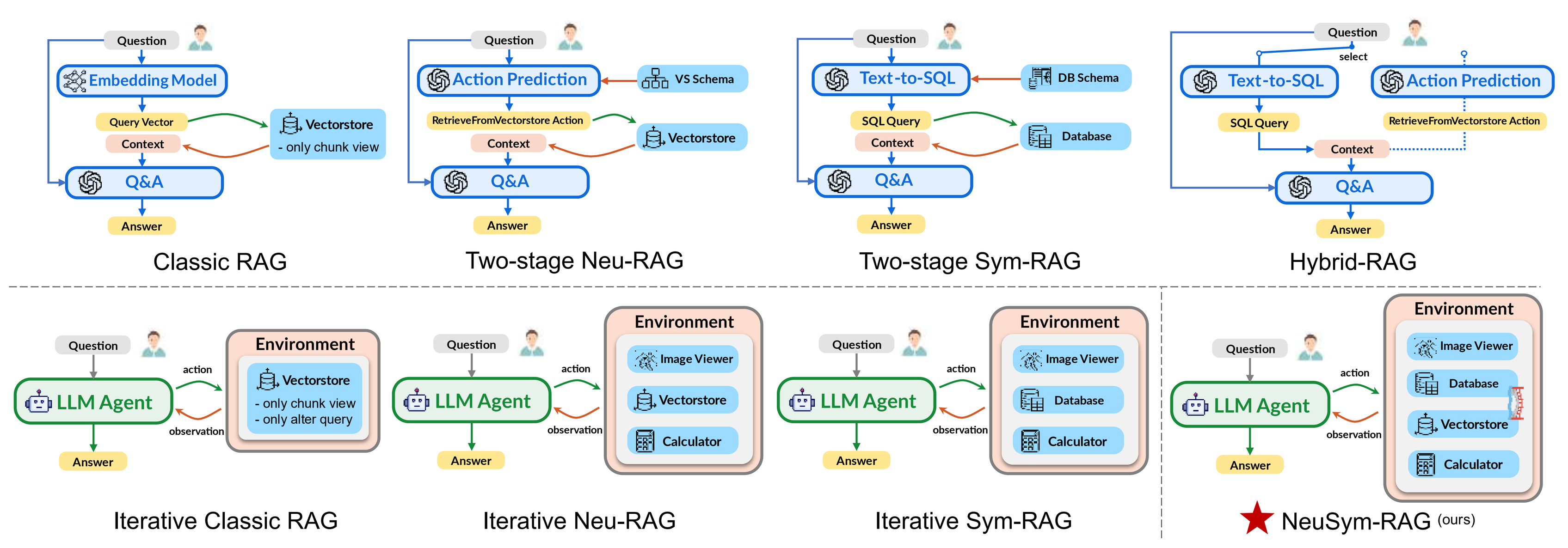
Category	Question	Answer Format
single	On the ALFWorld dataset experiments, how much did the success rate improve when the authors used their method compared to the original baseline model?	Your answer should be a floating-point number with one decimal place.
multiple	I would like to reproduce the experiments of KnowGPT, could you please provide me with the websites of the datasets applied in the experiment?	Your answer should be a Python list of 3 strings, the websites. Note that you should provide the original URL as given in the papers that proposed the datasets.
retrieval	Find the NLP paper that focuses on dialogue generation and introduces advancements in the augmentation of one-to-many or one-to-one dialogue data by conducting augmentation within the semantic space.	Your answer should be the title of the paper WITHOUT ANY EXPLANATION.

```
"eval_func": "eval_structured
        _object_exact_match",
"eval_kwargs": {
    "gold": [
        "SCG-NLI",
        false
    ],
    "ignore_order": false,
    "lowercase": true
}
```

examples of questions and evaluation from AirQA-Real dataset

Model	AirQA-Real text	table	image	formula	metadata	AVG	M3SciQA table	image	AVG	SciDQA table	image	formula	AVG
Classic-RAG													
GPT-4o-mini	12.3	11.9	12.5	16.7	13.6	13.4	17.9	10.6	15.6	59.4	60.4	59.3	59.8
GPT-4V	13.2	13.9	10.0	13.9	13.6	14.7	12.1	8.8	11.1	56.6	56.8	58.1	57.4
Llama-3.3-70B-Instruct	8.7	7.9	9.5	16.7	0.0	10.0	12.7	8.1	11.3	56.8	58.8	58.9	58.0
Qwen2.5-VL-72B-Instruct	9.6	5.9	11.9	11.1	13.6	10.5	11.6	11.6	11.6	54.8	56.9	56.3	56.2
DeepSeek-R1	11.7	13.9	9.5	30.6	9.1	13.9	11.9	9.5	11.2	63.9	61.3	61.7	62.4
NeuSym-RAG													
GPT-4o-mini	33.0	12.9	11.9	19.4	18.2	30.7	18.7	16.6	18.0	63.0	63.6	62.5	63.0
GPT-4V	38.9	**18.8**	**23.8**	**38.9**	**27.3**	37.3	13.7	13.4	13.6	62.6	63.5	63.2	63.1
Llama-3.3-70B-Instruct	30.6	11.9	16.7	16.7	**27.3**	29.3	**26.3**	17.6	**23.6**	55.5	57.3	56.6	56.4
Qwen2.5-VL-72B-Instruct	**43.4**	15.8	11.9	25.0	**27.3**	**39.6**	20.2	**22.7**	21.1	60.2	60.6	61.8	60.5
DeepSeek-R1	33.2	16.8	11.9	27.8	18.2	32.4	19.0	13.7	17.4	**64.3**	**64.6**	**63.9**	**64.5**

- **NeuSym-RAG remarkably outperforms** Classic RAG on all datasets.
- **VLMs** perform better in tasks that require vision capability.
- **Open-source LLMs** are capable of handling this interactive procedure in a zeroshot paradigm, and **even better than some closed-source models**.



comparisons between our NeuSym-RAG and other agentic baselines

Method	Neural	Symbolic	Multi-view	# Interaction(s)	sgl.	multi.	retr.	subj.	obj.	AVG
Question only	✗	✗	✗	1	5.7	8.0	0.4	9.4	2.7	4.0
Title + Abstract				1	5.7	14.0	0.0	13.1	3.6	5.4
Full-text w/. cutoff				1	28.3	10.7	0.4	26.2	7.6	11.2
Classic RAG	✓	✗	✗	1	18.2	4.0	9.4	8.4	11.0	10.5
Iterative Classic RAG				≥ 2	8.2	10.0	15.2	5.6	13.2	11.8
Two-stage Neu-RAG	✓	✗	✓	2	19.5	10.0	5.3	15.9	9.4	10.7
Iterative Neu-RAG				≥ 2	**37.7**	18.7	48.4	**32.7**	38.3	37.3
Two-stage Sym-RAG	✗	✓	✓	2	12.2	5.4	9.4	10.6	8.7	9.1
Iterative Sym-RAG				≥ 2	32.1	14.7	33.6	27.1	28.3	28.0
Graph-RAG	✓	✗	✓	2	22.2	11.1	0.0	21.1	11.5	15.6
Hybrid-RAG	✓	✓	✓	2	23.3	9.3	5.7	16.8	10.5	11.8
NeuSym-RAG (ours)				≥ 2	28.3	**32.3**	**58.2**	27.1	**42.6**	**39.6**

- Two-stage Neu-RAG (**multi-view**) beats Classic RAG.
- Hybrid RAG (**more views**) improves further.
- **Iterative** methods outperforms **two-stage** ones.
- As turn increases, **objective** score rises faster.