UDA: A Benchmark Suite for Retrieval Augmented Generation in Real-world Document Analysis

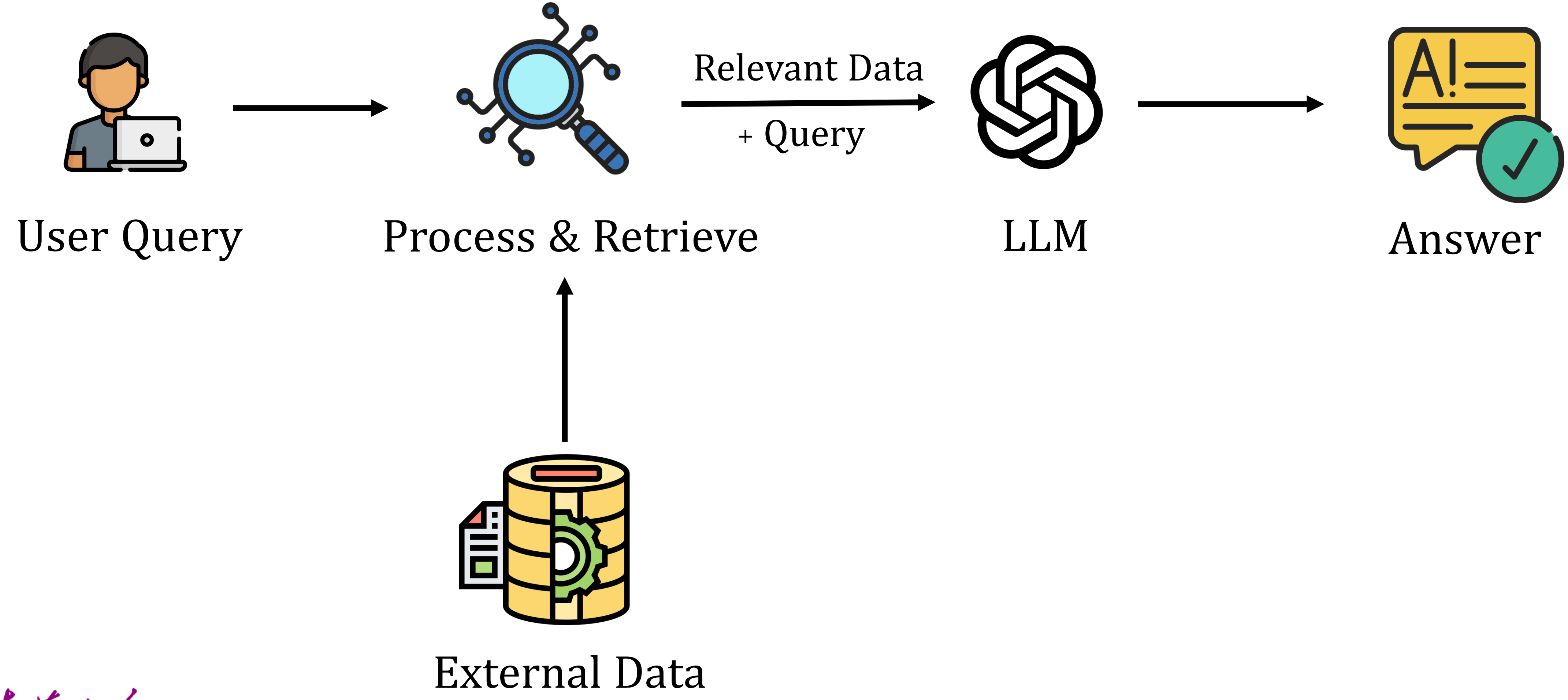
Yulong Hui, Yao Lu[1], Huanchen Zhang

Tsinghua University

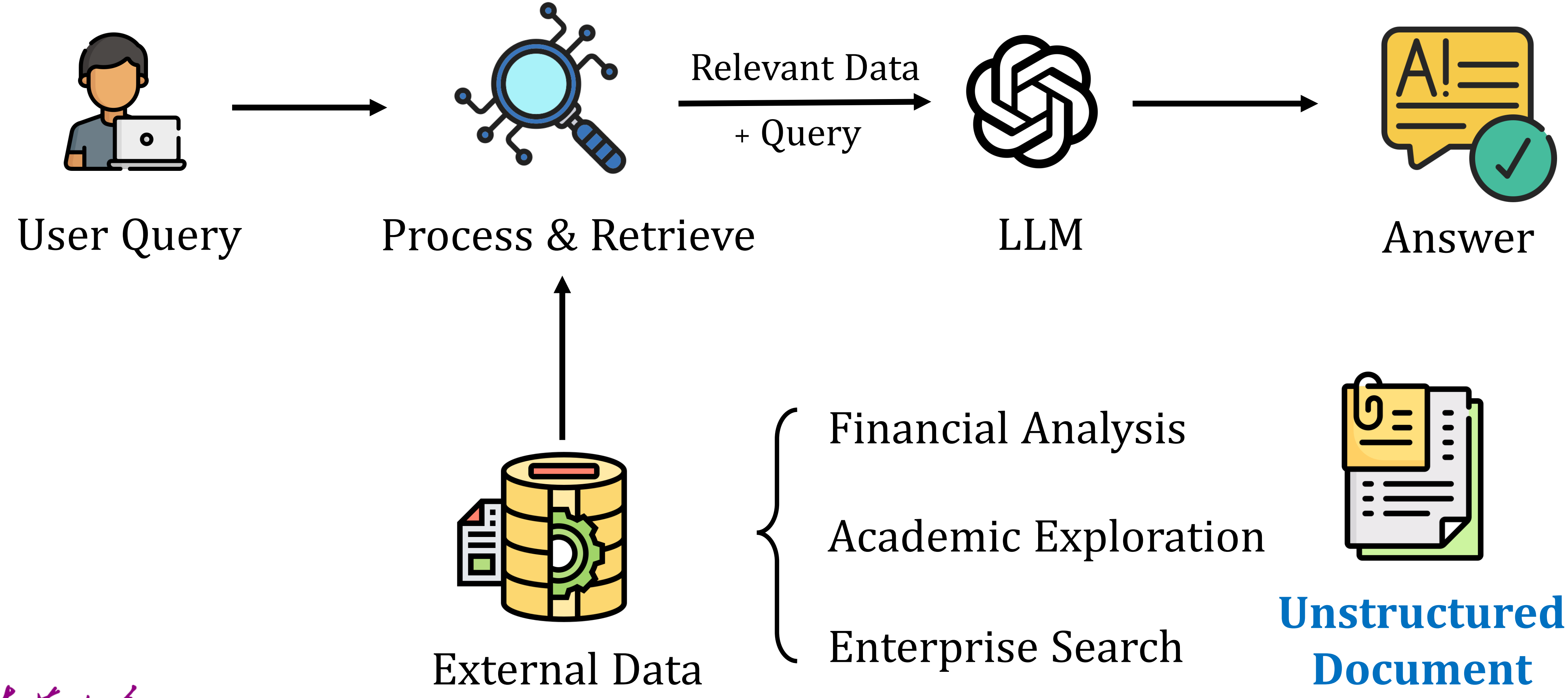
[1] National University of Singapore

NEURAL INFORMATION PROCESSING SYSTEMS

RAG for External Data Understanding



The Ubiquity of Unstructured Document



Challenges in Real-World Document Analysis

Unstructured Pattern

- Intricate layout
- Tabular data
- Noisy symbols

Lengthy Documents

- Redundancy context
- Hard to retrieve

Diverse Query Types

- Different answering strategies
- Mechanism decision

Challenges in Real-World Document Analysis

Unstructured Pattern

- Intricate layout
- Tabular data
- Noisy symbols

Lengthy Documents

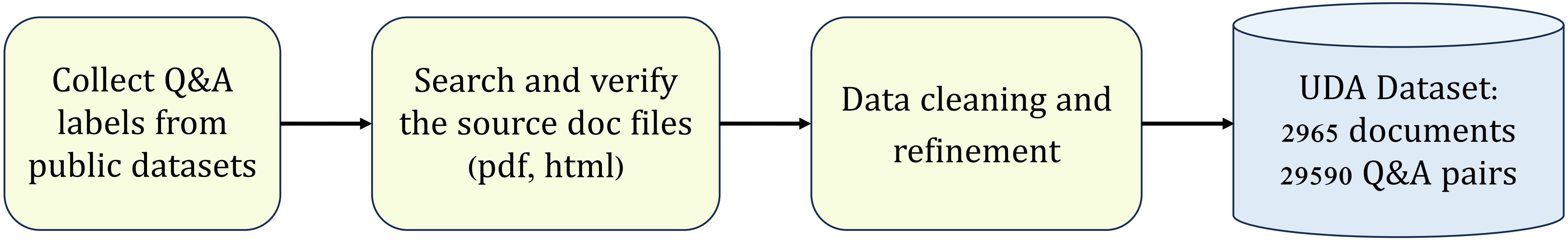
- Redundancy context
- Hard to retrieve

Diverse Query Types

- Different answering strategies
- Mechanism decision

Prior works overlook the challenges of real-world scenarios, providing:
clean or segmented input, homogeneous source domains (e.g. Wikipedia) and similar question types.

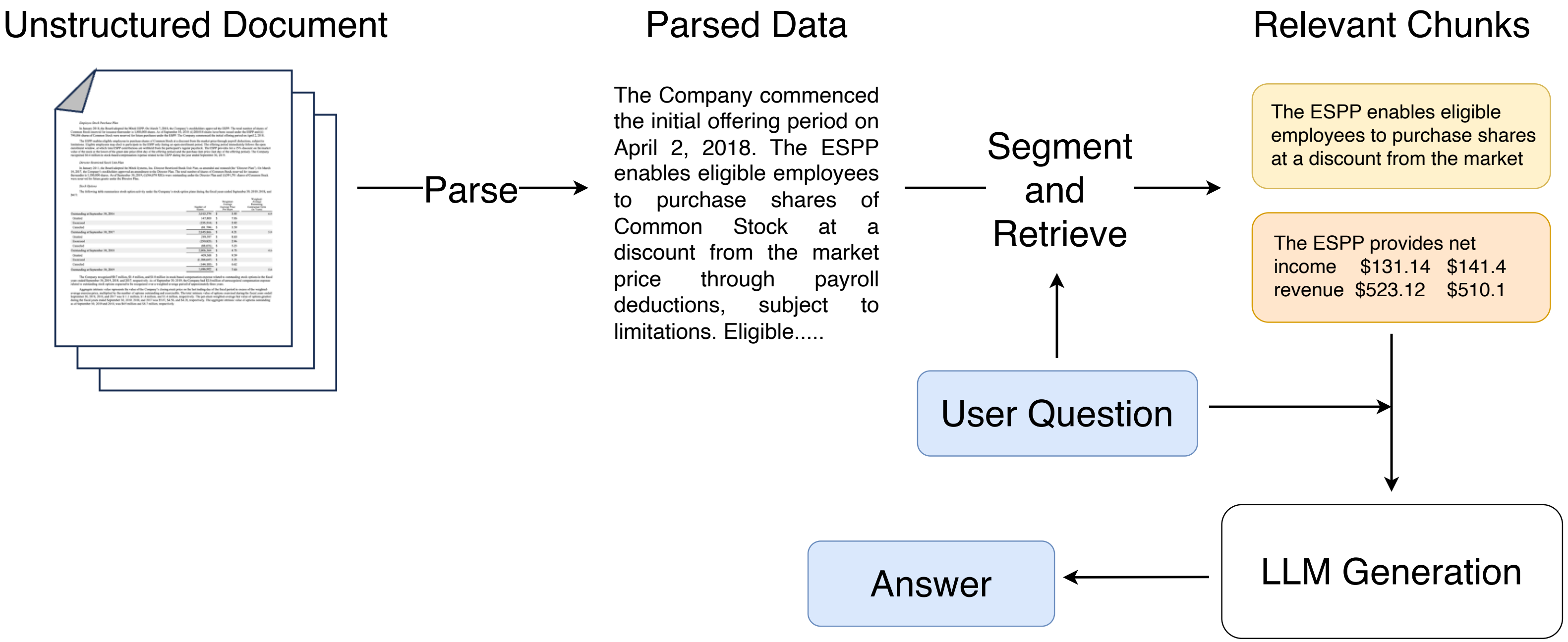
Our UDA Dataset



Each data item: (doc, question, ground-truth-answer)

- Integrated unstructured patterns
- Both tabular and textual data
- Un-segmented long context
- Diverse query types
- Multiple source domains (finance, academia, world-knowledge)

The Focus of UDA Benchmark



- Parsing approaches
- Long-context mechanism
- End-to-end performance
- Retrieval strategies
- LLM generation policies

Evaluation: Table Parsing

Table 1: Performance scores of LLMs on table-based Q&As, using varying parsing strategies.

Dataset	LLM Name	Well Parsed	GPT-4-Omni	Raw Text	CV	CV + LLM
Tabular FinHybrid (EM)	GPT-4-Turbo	71.9	**72.4**	68.0	61.3	52.4
	Llama-3-8B	**59.5**	56.3	51.6	44.6	40.2
PaperTab (F1)	GPT-4-Turbo	42.8	**44.3**	42.4	38.6	40.7
	Llama-3-8B	35.8	**37.7**	36.5	34.6	32.1

- Traditional CV-based method may be unreliable due to edge cases.
- Raw-text extraction yields decent results with structural markers (e.g. line-breakers and spaces).

Evaluation: RAG and Long-context

Table 2: Performance scores between RAG and the long-context mechanism.

LLM Name	Input Type	FinHybrid	TatHybrid	PaperTab	PaperText	FetaTab	NqText
Qwen-1.5-7B	OpenAI Retrieval @5	**21.0**	**26.6**	**31.4**	**39.1**	58.1	**32.4**
	Long Context	3.0	20.9	26.3	33.1	**58.7**	30.2
GPT-4-Turbo	OpenAI Retrieval @5	**43.4**	**46.3**	**43.5**	47.1	61.8	**35.8**
	Long Context	37.4	36.9	43.3	**47.4**	**63.3**	35.4

- Long-context processing falls short in financial tasks that require precise information and calculations.
- The smaller model prefers RAG due to the limited long-context capability.

More Evaluations and Observations

- Model scaling laws may not apply to retrieval scenarios.
- Sparse retriever outperforms in specific tasks.
- Chain-of-Thought excels in solving analytical problems.

……

More details in the paper!
Project Link: https://github.com/qinchuanhui/UDA-Benchmark