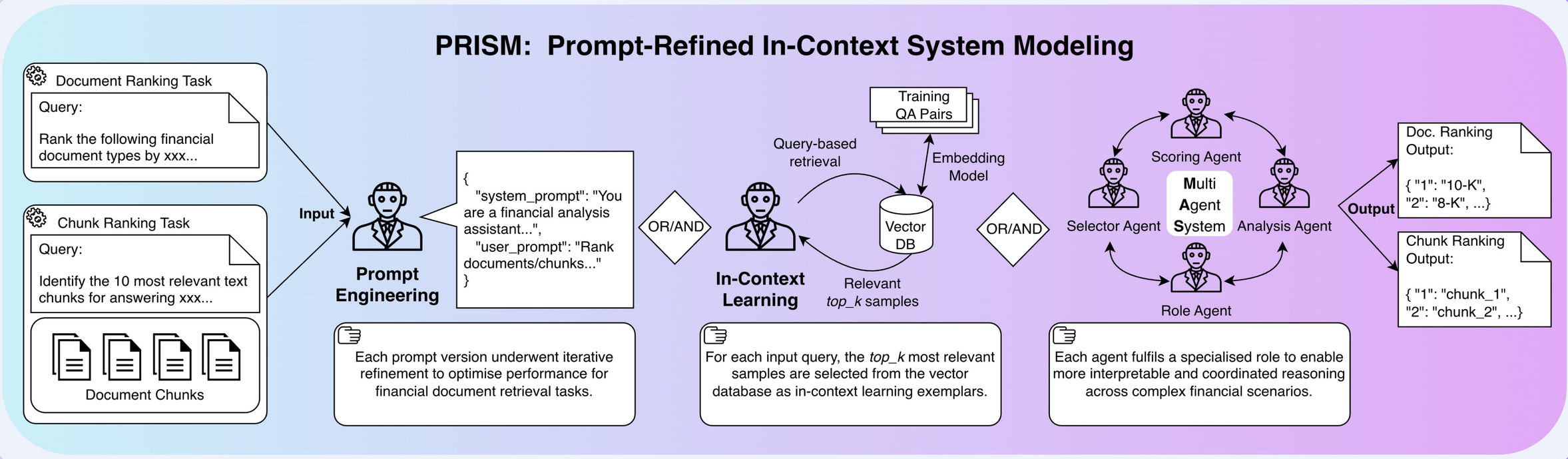
PRISM: Prompt-Refined In-Context System Modeling for Financial Retrieval

Chun Chet Ng*, Jia Yu Lim*,
Wei Zeng Low
AI Lens, Kuala Lumpur, Malaysia
{chunchet.ng}@ailensgroup.com

AILENS

August 9–13, 2026
KDD2026
Jeju, Korea

We present **PRISM, a training-free framework that integrates refined system prompting**, **in-context learning (ICL)**, and **lightweight multi-agent coordination** for **document and chunk ranking tasks**. Our primary contribution is **a systematic empirical study of when each component provides value**: **prompt engineering** delivers consistent performance with minimal overhead, **ICL** enhances reasoning for complex queries when applied selectively, and **multi-agent systems** show potential primarily with larger models and careful architectural design. Extensive ablation studies across **FinAgentBench**, **FiQA-2018**, and **FinanceBench** reveal that **simpler configurations often outperform complex multi-agent pipelines**. Our best configuration achieves an **NDCG@5 of 0.71818 on FinAgentBench**, **ranking third** while being **the only training-free approach** in the top three. We provide comprehensive feasibility analyses covering **latency**, **token usage**, **and cost trade-offs** to support deployment decisions.



Module 1 – Prompt Engineering for Financial Datasets

Encodes domain-specific priors derived from our exploratory data analysis. We introduce logical scaffolding inspired by **ReAct**, **CoT**, and **ToT** to promote **explicit reasoning** before producing final answers. **Four prompt variants** were designed to study the impact of incremental modifications and progressively enhance reasoning structure.

Module 2 – In-Context Learning Sample Retrieval

With **few-shot learning** through a **simplified RAG pipeline**, we construct **an offline exemplar store** from the training dataset and **embed all query-document pairs** using a text embedding model, indexing them in a FAISS vector store. During inference, the **top-*k* most semantically similar exemplars** are **dynamically retrieved** from the vector store.

Module 3 – Multi-Agent System (MAS) Modeling

MAS decomposes ranking into **specialized agent roles** spanning **noise removal**, **candidate selection**, **scoring**, and consensus aggregation to reduce cognitive load, improve calibration, and mitigate hallucinations. We study **how architectural** complexity, **model capacity**, and **graph topology** determine when it provides measurable retrieval value.

ID	Document Ranking Configuration			Chunk Ranking Configuration			Agentic Configuration		NDCG@5 Score	
	Prompt	Model	ICL/Embedding	Prompt	Model	ICL/Embedding	Doc. Agent	Chunk Agent	Public	Private
Non-Agentic Workflow										
1	P_1	GPT-4.1	N/A	P_1	GPT-4.1	N/A	N/A		0.63956	0.66666
2	P_1	GPT-5-mini	N/A	P_1	GPT-5-mini	N/A	N/A		0.65029	0.66628
3	P_1	GPT-5-mini	N/A	P_1	GPT-4.1	N/A	N/A		0.63834	0.66328
4	P_1	GPT-4.1	N/A	P_1	GPT-5-mini	N/A	N/A		0.65151	0.66966
5	P_1	GPT-4.1	N/A	P_2	GPT-5-mini	N/A	N/A		0.63817	0.67353
6	P_1	GPT-5-mini	N/A	P_3	GPT-5-mini	N/A	N/A		0.64711	0.67021
7	P_2	GPT-5-mini	N/A	P_1	GPT-4.1	N/A	N/A		0.63835	0.66088
8	P_2	GPT-5-mini	N/A	P_1	GPT-5-mini	N/A	N/A		0.63694	0.67014
9	P_2	GPT-5-mini	N/A	P_2	GPT-5-mini	N/A	N/A		0.63695	0.66774
10	P_3	GPT-5-mini	N/A	P_1	GPT-4.1	N/A	N/A		0.64075	0.66673
11	P_3	GPT-5-mini	N/A	P_1	GPT-5-mini	N/A	N/A		0.6527	0.66973
12	P_3	GPT-5-mini	N/A	P_3	GPT-5-mini	N/A	N/A		0.64297	0.69019
13	P_4	GPT-5-mini	N/A	P_4	GPT-5-mini	N/A	N/A		0.65559	0.68895
14	P_4	GPT-5-mini	N/A	P_4	GPT-5	N/A	N/A		0.67855	0.70575
15	P_4	GPT-5	N/A	P_4	GPT-5	N/A	N/A		0.66236	0.70977
16	P_4	GPT-5-mini	ICL-5/TE3-S	P_4	GPT-5-mini	N/A	N/A		0.66685	0.69325
17	P_4	GPT-5-mini	ICL-5/TE3-S	P_4	GPT-5-mini	ICL-5/TE3-S	N/A		0.64841	0.68252
18	P_4	GPT-5-mini	ICL-5/TE3-S	P_4	GPT-5	N/A	N/A		0.67373	0.71446
19	P_4	**GPT-5**	**ICL-5/TE3-S**	P_4	**GPT-5**	**N/A**	**N/A**		**0.67444**	**0.71818**
Agentic Workflow										
20	N/A	GPT-4o-mini	N/A	N/A	GPT-4o-mini	N/A	A_1		0.59171	0.57276
21	P_1	GPT-4o-mini	N/A	P_1	GPT-4o-mini	N/A	A_1		0.58714	0.58234
22	P_1	GPT-4o-mini	N/A	P_1	GPT-4o-mini	N/A	A_2		0.52028	0.51518
23	P_1	GPT-5-mini	N/A	P_1	GPT-5-mini	N/A	A_2		0.56317	0.55256
24	P_1	GPT-4o-mini	N/A	P_1	GPT-4o-mini	N/A	A_3		0.53861	0.53400
25	P_1	GPT-5-mini	N/A	P_1	GPT-5-mini	N/A	A_3		0.63515	0.66291
26	P_1	GPT-4o-mini	N/A	P_1	GPT-4o-mini	N/A	A_4		0.52395	0.49782
27	P_1	GPT-5-mini	N/A	P_1	GPT-5-mini	N/A	A_4		0.61226	0.63654
28	P_1	GPT-5-mini	ICL-5/TE3-S	P_1	GPT-5-mini	ICL-5/TE3-S	A_4		0.64721	0.62775
29	P_1	GPT-5-mini	ICL-10/TE3-S	P_1	GPT-5-mini	ICL-10/TE3-S	A_4		0.63089	0.62771
30	P_1	GPT-5-mini	ICL-10/TE3-L	P_1	GPT-5-mini	ICL-10/TE3-L	A_4		0.60038	0.62294
31	P_1	GPT-5-mini	ICL-15/TE3-L	P_1	GPT-5-mini	ICL-15/TE3-L	A_4		0.58428	0.59906
32	P_4	GPT-5-mini	ICL-10/TE3-S	P_1	GPT-5-mini	ICL-10/TE3-S	A_4		0.58328	0.55757
Hybrid Workflow										
33	P_4	GPT-5-mini	N/A	P_4	GPT-5	N/A	A_3	N/A	0.67825	0.70685
34	P_4	GPT-5-mini	N/A	P_4	GPT-5	N/A	A_4	N/A	0.67234	0.69439

FinanceBench Evaluation Results

Prompt	Model	ICL	Correct	Incorrect	Unable to Answer
Baseline	GPT-5	N/A	141	9	0
P_1	GPT-5	ICL-9	147	3	0
P_2	GPT-5	ICL-9	142	8	0
P_3	GPT-5	ICL-9	140	10	0
P_4	GPT-5	ICL-9	146	4	0

FiQA-2018 Evaluation Results

Prompt	Model	ICL	NDCG@10	Recall@100
BM25+CE	N/A	N/A	0.3470	0.5390
P_1	GPT-5	ICL-5	0.6104	0.8083
P_2	GPT-5	ICL-5	0.6027	0.7941
P_3	GPT-5	ICL-5	0.6197	0.8206
P_4	GPT-5	ICL-5	0.6097	0.8021

FinAgentBench Reproducibility Analysis

Descriptive statistics across multiple runs

Run ID	n	Mean	SD	CV (%)	95% CI-Low	95% CI-High
12	5	0.68005	0.01023	1.50489	0.66734	0.69276
15	4	0.70246	0.00661	0.94078	0.69194	0.71297
18	4	0.70660	0.00546	0.77319	0.69790	0.71529
19	9	0.71163	0.00433	0.60861	0.70830	0.71496

Welch's t-tests of Run 12

Comparison	t-stat	p-val	Signif.
12 vs 15	−3.969	0.0057	Yes
12 vs 18	−4.981	0.0022	Yes
12 vs 19	−6.582	0.0014	Yes

In conclusion, this study presents **PRISM**, **a training-free framework leveraging system prompting**, **in-context learning**, and **multi-agent coordination for financial document and chunk ranking**. We demonstrate that **simpler non-agentic configurations with document-level ICL** achieve the strongest performance, while **multi-agent coordination provides value only with larger models and careful scope limitation**, challenging common assumptions about agentic architectures. Evaluation across **FinAgentBench**, **FiQA-2018**, and **FinanceBench** confirms **PRISM's cross-dataset adaptability**, while our feasibility analysis establishes the **first latency-cost baselines for training-free financial information retrieval**. These findings offer practical guidance for deploying LLM-based retrieval in high-stakes financial applications.