

BizSage: A Self-Evolving Multi-Agent Framework for Business Research with Efficient Knowledge Retrieval

Yuhe Wu, Guangyu Wang, Jiaxin Liu, Guang Zheng*

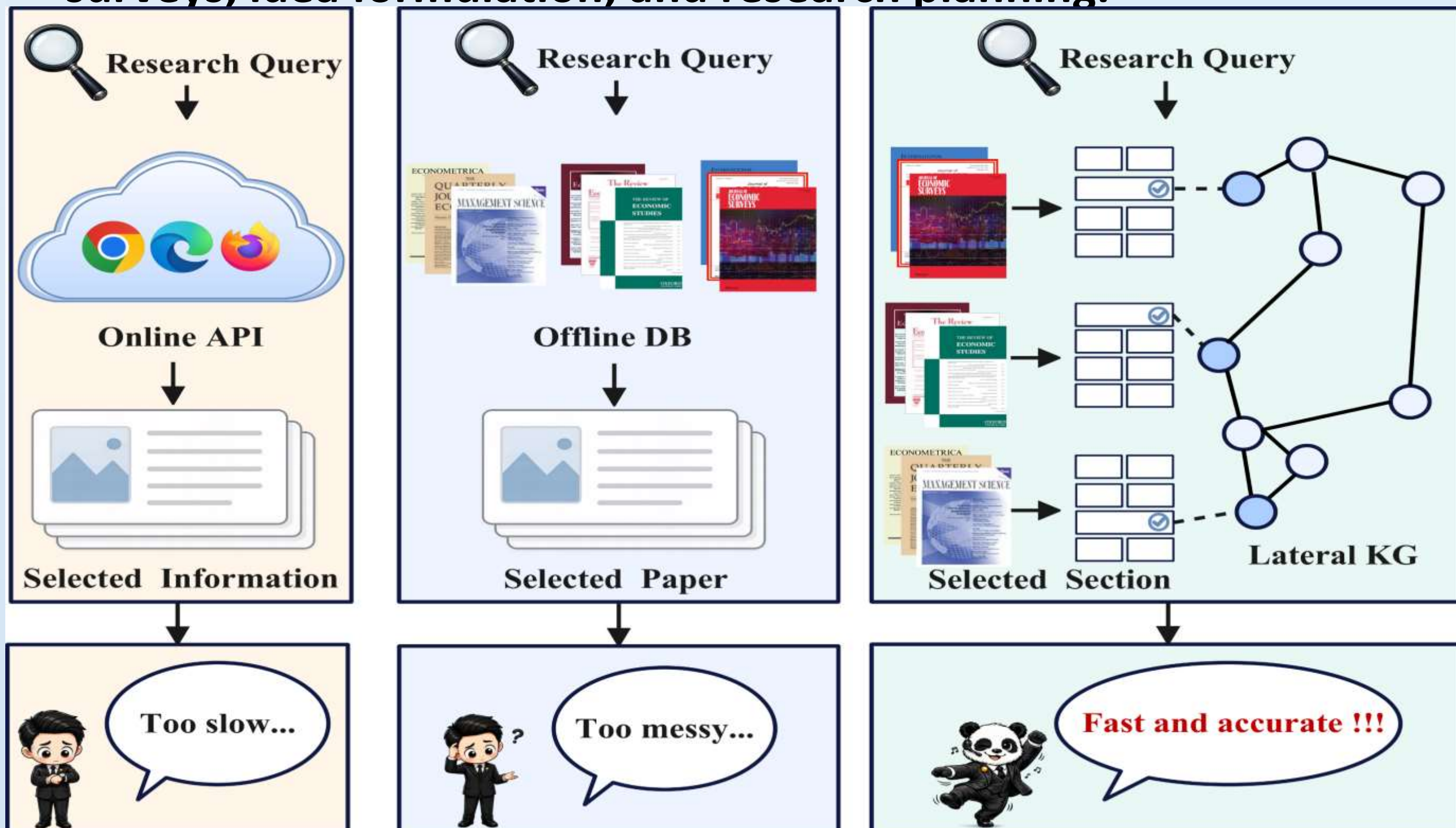


THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

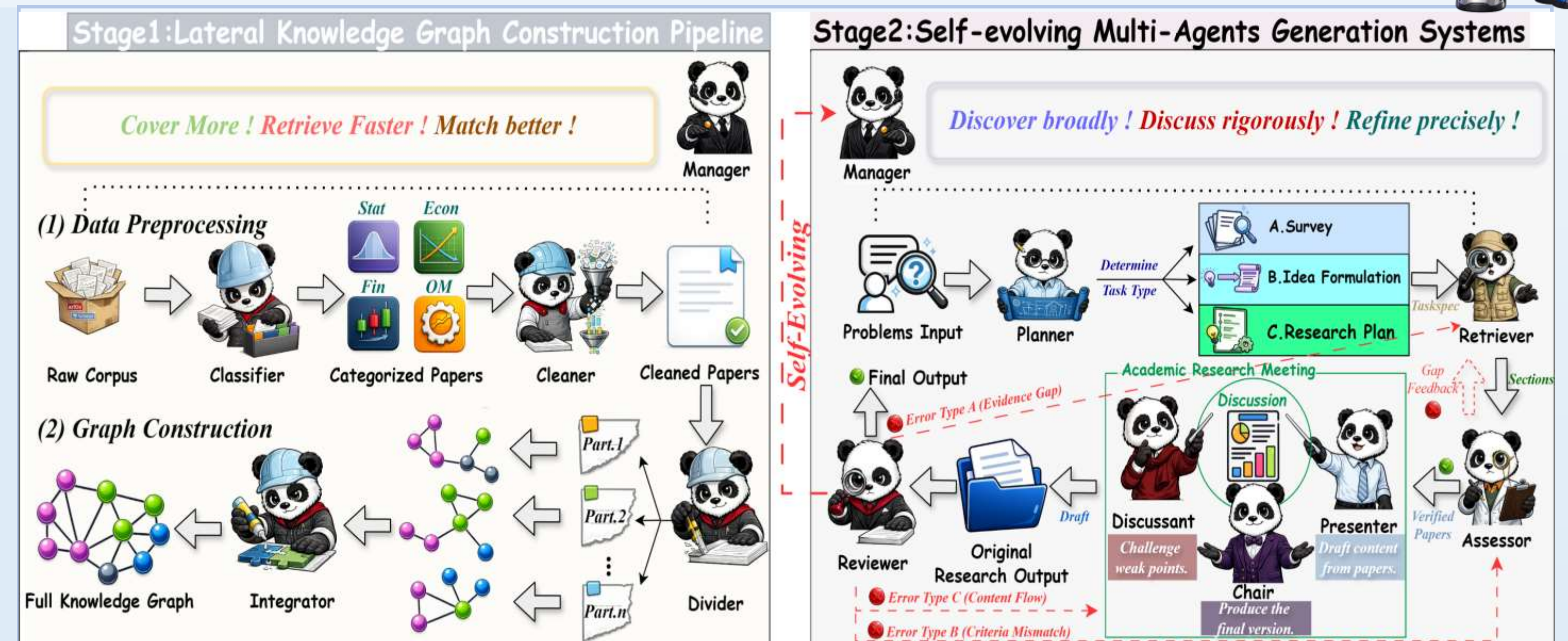


INTRODUCTION & MOTIVATION

- **Retrieval Gap:** Paper-level search struggles to isolate evidence scattered across sections and neighboring disciplines in economics and business research.
- **Quality Adaptation:** Research agents must meet domain-specific standards of grounding and empirical rigor, yet often underuse feedback from previous tasks.
- **Our Approach:** BizSage combines section-level graph retrieval, multi-agent collaboration, and reusable quality strategies for surveys, idea formulation, and research planning.



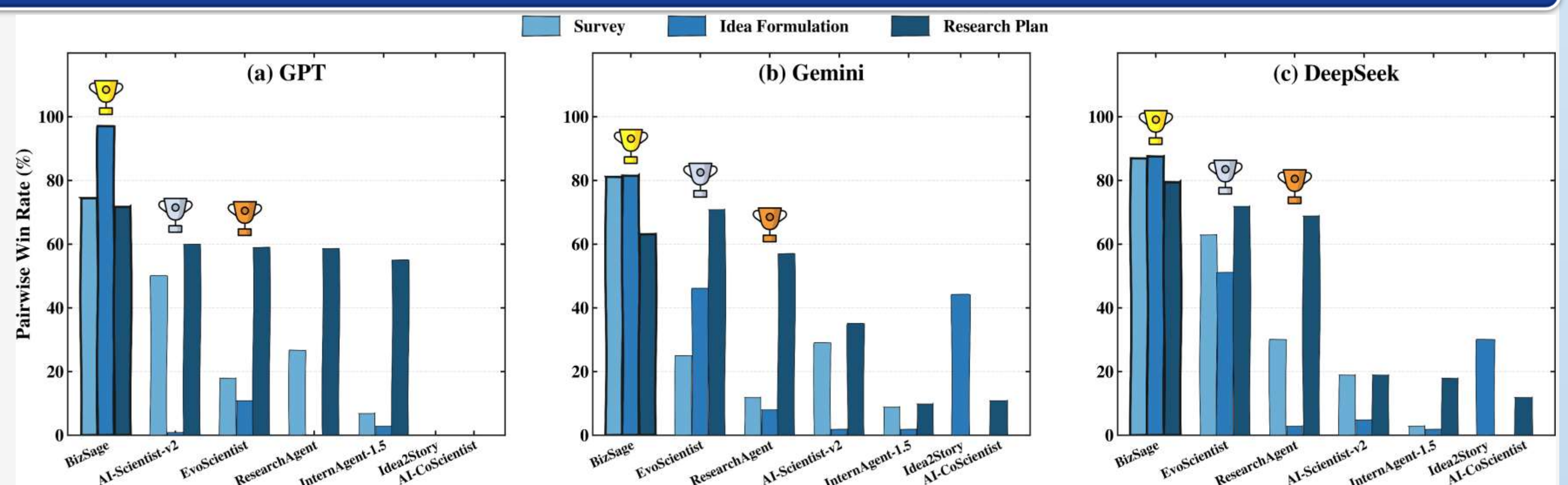
BIZSAGE FRAMEWORK



- **Knowledge Retrieval:** Section-level graphs merge through shared concepts into a Lateral Knowledge Graph. Personalized PageRank ranks sections by semantic relevance and graph importance.
- **Multi-Agent Collaboration:** Seven agents assess evidence, draft, critique, revise, and review outputs under a Manager, which routes feedback for targeted correction.
- **Self-Evolution:** Meta-Review analyzes recurring failures, tests corrective strategies on difficult past cases, and stores validated strategies for future tasks.

EXPERIMENT

Method	Survey				Idea Formulation				Research Plan			
	Rel.	Grd.	Cov.	Syn.	Rel.	Grd.	Nov.	Fea.	Rel.	Grd.	Rig.	Fea.
GPT-5.5												
AI-Scientist-v2	4.68	3.57	3.97	3.94	4.68	3.29	3.69	3.85	4.88	4.22	4.05	3.84
EvoScientist	4.55	2.20	3.97	3.82	4.72	2.28	2.53	3.75	4.43	1.84	2.80	2.76
InternAgent-1.5	3.52	1.97	2.54	2.84	4.19	2.16	3.04	3.36	4.22	1.56	3.87	3.03
ResearchAgent	4.37	2.47	3.81	3.82	4.33	2.87	3.32	3.66	4.83	2.55	3.17	3.52
Idea2Story [†]	—	—	—	—	4.66	2.56	3.14	3.69	—	—	—	—
AI Co-Scientist [†]	—	—	—	—	—	—	—	—	3.91	1.62	2.40	2.47
BizSage (Ours)	4.77	4.14	4.04	3.99	4.71	3.42	3.43	3.95	4.89	4.36	4.07	3.95
Gemini-3-Flash												
AI-Scientist-v2	3.73	2.48	3.00	2.85	4.09	2.25	2.85	2.72	4.25	1.80	2.88	2.83
EvoScientist	3.91	2.27	3.16	2.83	4.23	2.47	2.43	3.32	3.97	2.30	2.28	2.52
InternAgent-1.5	2.69	1.53	2.12	1.90	3.44	1.25	2.15	2.24	3.21	1.51	2.11	2.03
ResearchAgent	3.67	1.82	2.75	2.53	3.65	2.19	2.76	2.74	3.99	1.96	2.89	2.84
Idea2Story [†]	—	—	—	—	4.44	2.03	3.05	3.12	—	—	—	—
AI Co-Scientist [†]	—	—	—	—	—	—	—	—	2.58	1.36	1.23	1.48
BizSage (Ours)	3.94	2.96	3.41	3.01	4.25	2.99	2.86	3.78	4.02	2.95	3.07	2.86
DeepSeek-V4-Flash												
AI-Scientist-v2	3.93	2.61	3.06	3.03	4.37	2.62	2.91	3.31	4.46	2.06	3.42	3.28
EvoScientist	4.51	2.76	3.97	3.90	4.72	2.91	3.19	3.20	4.87	3.67	3.89	3.68
InternAgent-1.5	2.38	1.42	1.82	1.87	3.95	1.96	2.64	2.75	3.90	1.46	2.68	2.54
ResearchAgent	3.97	2.05	3.12	3.06	3.93	2.21	2.80	3.04	4.44	2.32	3.33	3.26
Idea2Story [†]	—	—	—	—	4.50	2.22	2.92	3.13	—	—	—	—
AI Co-Scientist [†]	—	—	—	—	—	—	—	—	3.59	1.46	2.09	2.23
BizSage (Ours)	4.65	2.96	3.89	4.08	4.74	3.07	3.06	3.79	4.46	3.88	3.93	3.70
Method	Survey				Idea Formulation				Research Plan			
	Rel.	Grd.	Cov.	Syn.	Rel.	Grd.	Nov.	Fea.	Rel.	Grd.	Rig.	Fea.
GPT-5.5												
BizSage (Iteration 1)	4.14	3.05	3.77	3.86	4.52	2.07	3.16	3.54	4.60	3.35	3.94	3.31
BizSage (Iteration 2)	4.30	4.06	3.94	3.89	4.57	3.17	3.40	3.67	4.62	3.75	4.00	3.42
BizSage (Iteration 3)	4.77	4.14	4.04	3.99	4.71	3.42	3.43	3.95	4.89	4.36	4.07	3.95
Gemini-3-Flash												
BizSage (Iteration 1)	3.90	2.67	3.35	2.93	4.10	2.95	2.73	3.63	3.57	2.55	2.96	2.91
BizSage (Iteration 2)	3.92	2.88	3.38	2.99	4.21	2.97	2.82	3.78	3.81	2.76	2.94	2.90
BizSage (Iteration 3)	3.94	2.96	3.41	3.01	4.25	2.99	2.86	3.78	4.02	2.95	3.07	2.86
DeepSeek-V4-Flash												
BizSage (Iteration 1)	4.13	2.92	3.47	3.42	4.22	1.98	2.64	2.77	4.18	3.06	3.33	3.21
BizSage (Iteration 2)	4.15	2.91	3.52	3.43	4.21	1.93	2.85	2.94	4.29	3.11	3.37	3.26
BizSage (Iteration 3)	4.65	2.96	3.89	4.08	4.74	3.07	3.06	3.79	4.46	3.88	3.93	3.70



- **Main Experiment:** Across 300 test queries, four domains, three tasks, and three backbones, BizSage leads most metrics against six baselines: 10/12 on GPT-5.5, with a 0.57-point lead in Survey Grounding.

Method	Survey				Idea Formulation				Research Plan			
	Rel.	Grd.	Cov.	Syn.	Rel.	Grd.	Nov.	Fea.	Rel.	Grd.	Rig.	Fea.
BizSage (full)	4.77	4.14	4.04	3.99	4.71	3.42	3.43	3.95	4.89	4.36	4.07	3.95
Ablations												
without Section-Level Knowledge Graph	4.32	3.19	3.99	3.92	4.48	2.26	3.35	3.61	4.73	3.56	4.01	3.55
without Reviewer	4.16	4.06	3.89	3.83	4.36	3.08	3.35	3.50	4.62	3.27	3.97	3.32
without Discussant	4.20	4.03	3.79	3.68	4.26	3.07	3.19	3.36	4.57	3.26	3.87	3.18
without Iterative Refinement	4.23	3.15	3.93	3.94	4.36	2.23	3.34	3.55	4.71	2.41	3.97	3.46

- **Self Evolution:** Strategy refinement yields the largest gains in grounding, with benefits varying across backbones. This suggests that accumulated strategies help most when agents can reliably apply them.
- **Component Contributions:** Components play complementary roles: structured retrieval supports grounding, critique strengthens synthesis and iterative refinement helps resolve weaknesses that retrieval alone leaves unaddressed.

DISCUSSION

- **Graph Construction:** GPT-4o offers a practical balance between extraction quality and cost. Stronger models improve concept and relation extraction, but provide limited gains in cross-document merge consistency.

Extraction Model	NV	RS	MC
Qwen3-8B	65.0	44.4	42.9
GPT-4o	90.0	88.2	92.3
GPT-5.5	95.2	94.4	93.3
Hallucination Rate			
Method	#Citations	Hallucination Rate	
BizSage (Ours)	894	0.0%	
InternAgent-1.5	98	5.1%	
EvoScientist	75	2.7%	
AI-Scientist-v2	0	—	
ResearchAgent	0	—	
Idea2Story*	0	—	
AI Co-Scientist*	0	—	

- **Citation Integrity:** BizSage maintains citation integrity in the sampled audit, verified through CrossRef checks and human review. Baseline outputs also reveal domain drift: references may point to real publications yet remain unrelated to the research topic. These findings highlight that citation reliability requires both accurate bibliographic details and meaningful alignment with the research question.

CONCLUSION & LIMITATIONS

- **Conclusion:** BizSage combines evidence retrieval and reusable feedback strategies for business research.
- **Limitations:** English journal literature, 300 test queries, and three evolution iterations limit generalization.

Explore BizSage. Let's talk research!



Wechat



Website



Linkedin