CSCE 689 - Special Topics in NLP for Science

LLMs for Research: Idea Generation

Hangxiao Zhu, 04/01/2025

Background

➡ Generating **novel** research ideas is a crucial but challenging step in the scientific process.

➡ Traditionally, ideation relies heavily on human expertise, domain knowledge, and creativity.

➡ With the rapid advancement of Large Language Models (LLMs) like GPT-4 and Claude, researchers have begun exploring their potential to:

- Read and synthesize scientific literature
- Propose novel problem-method-experiment tuples
- Assist or even autonomously generate research ideas

Key Questions

➡ Are these ideas truly novel and useful?

➡ Can LLMs outperform human experts at ideation?

➡ How do we evaluate AI-generated ideas at scale?

Agenda

- ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models [NAACL 2025]
- Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers [ICLR 2025]
- Nova: An Iterative Planning and Search Approach to Enhance Novelty and Diversity of LLM Generated Ideas [arXiv 2024]

Agenda

- ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models [NAACL 2025]
- Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers [ICLR 2025]
- Nova: An Iterative Planning and Search Approach to Enhance Novelty and Diversity of LLM Generated Ideas [arXiv 2024]

Motivation

➡ Scientific research is slow and knowledge-heavy

➡ Research idea generation is critical but under-explored

➡ LLMs have potential to assist ideation, not just validation

Inspiration from Human

Human

Deep understanding of related scientific

An encyclopedic view of concepts and their relations

Feedback and criticism from peers researchers

Inspiration from Human

Human	Research Agent
Deep understanding of related scientific	Begins with a core paper and explores related ones
An encyclopedic view of concepts and their relations	Build an entity-centric knowledge store of concepts
Feedback and criticism from peers researchers	Feedback and criticism from peer LLMs

ResearchAgent

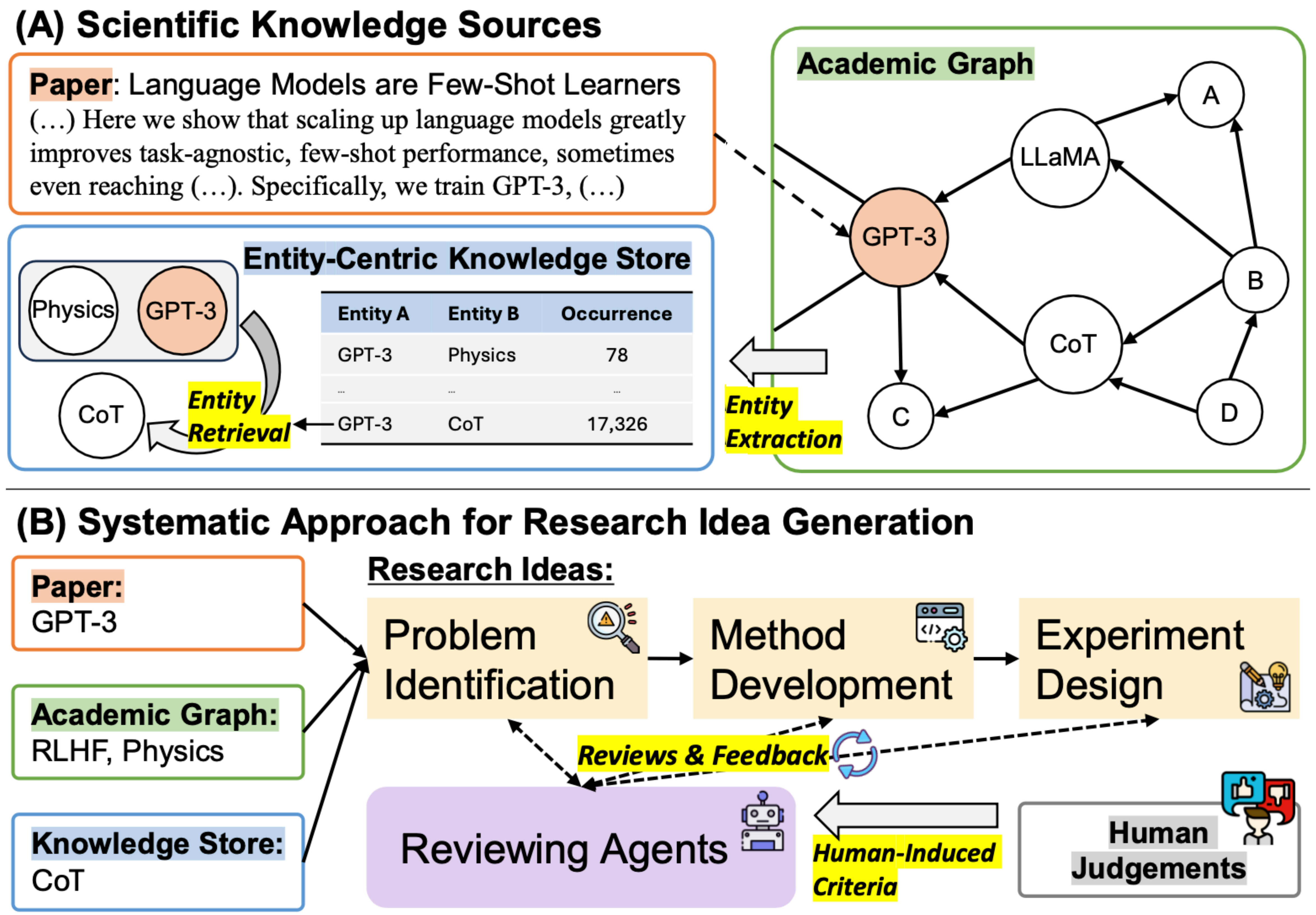
Core Paper

+

Incorporates citation graphs and entity-centric knowledge

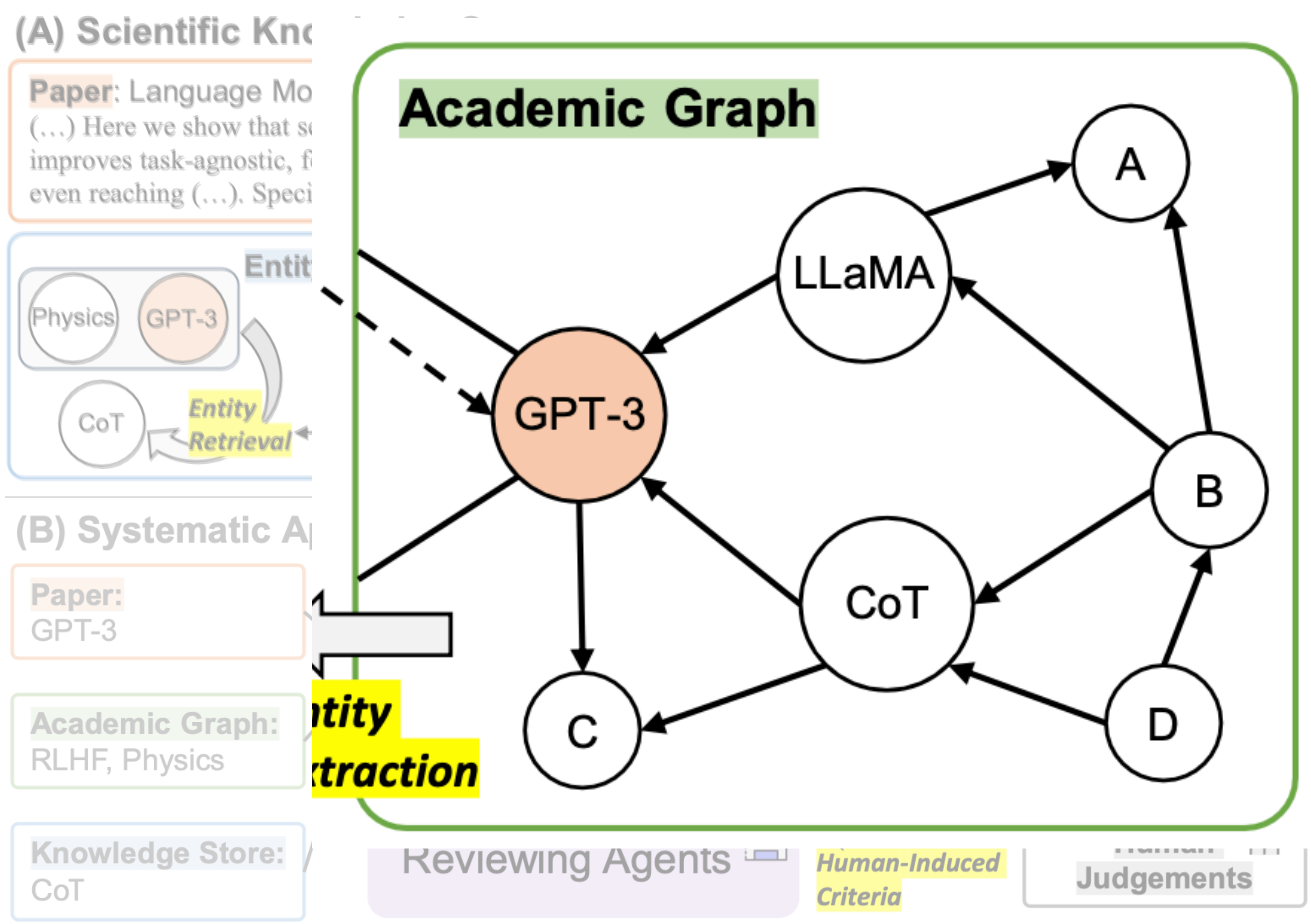
+

Uses ReviewingAgents for iterative refinement



ResearchAgent

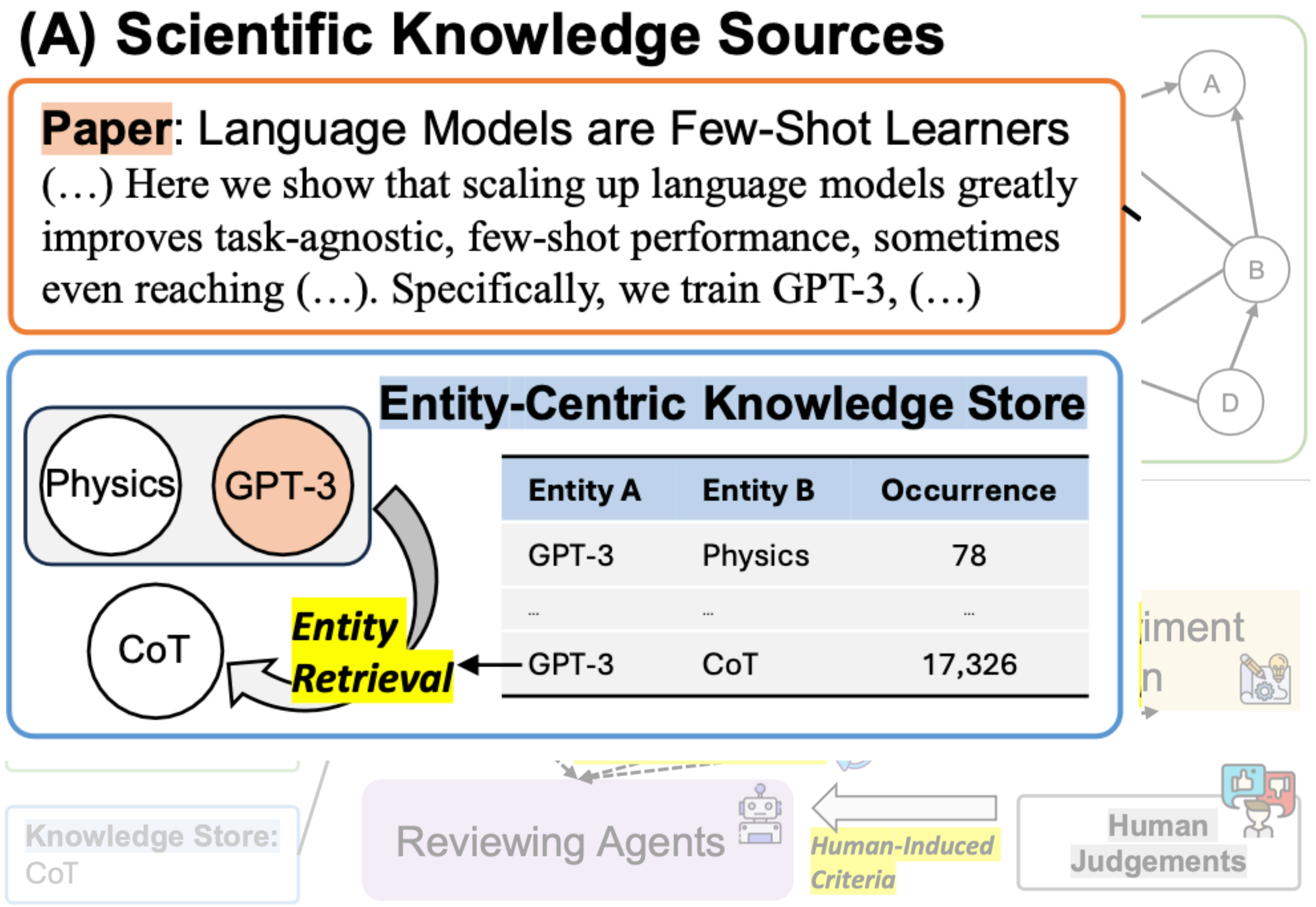
- ➡ Citation Graph-based Literature Survey



ResearchAgent

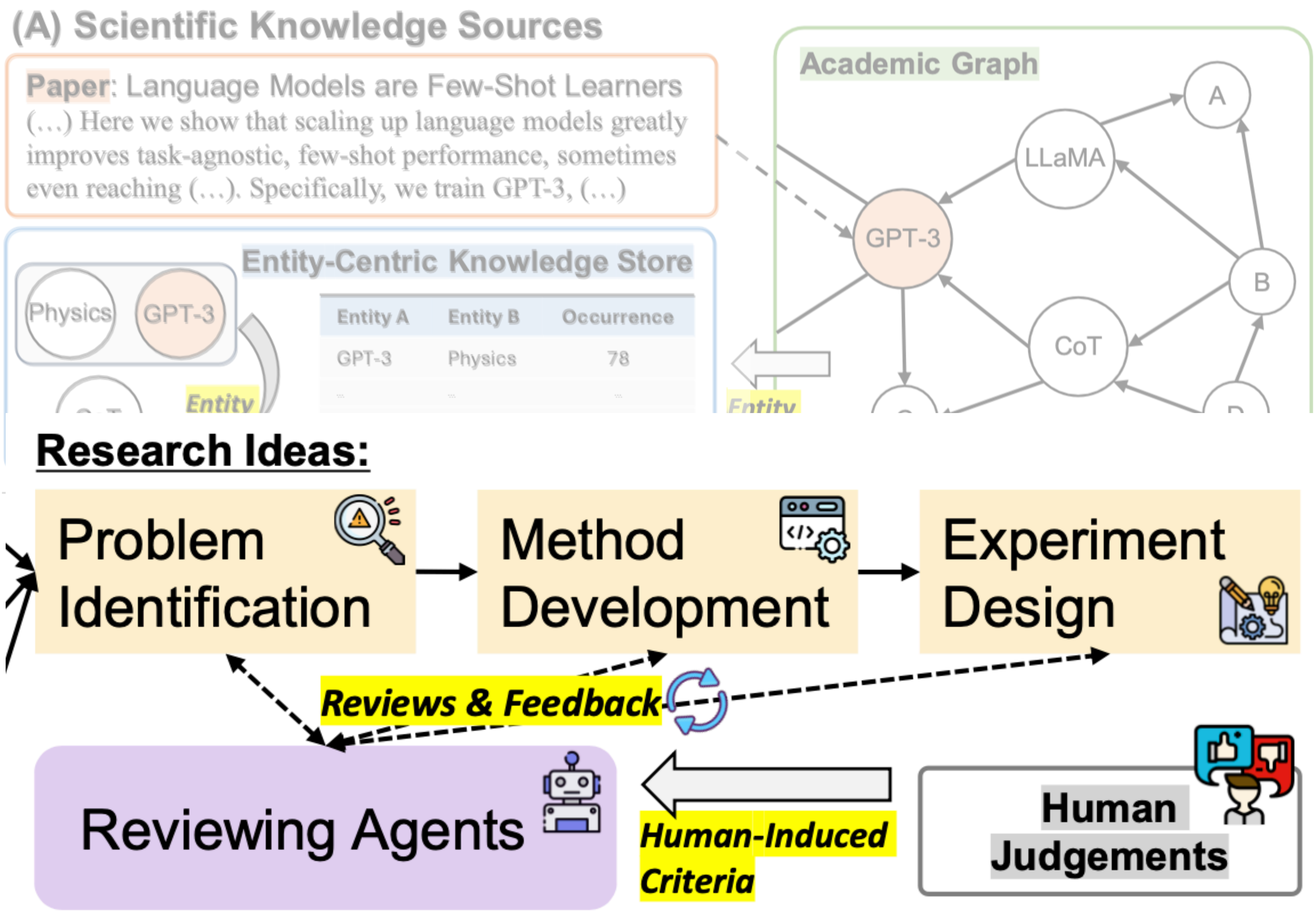
➡ Citation Graph-based Literature Survey

➡ Entity-Centric Knowledge Augmentation



ResearchAgent

- ➡ Citation Graph-based Literature Survey
- ➡ Entity-Centric Knowledge Augmentation
- ➡ Iterative Research Idea Refinements

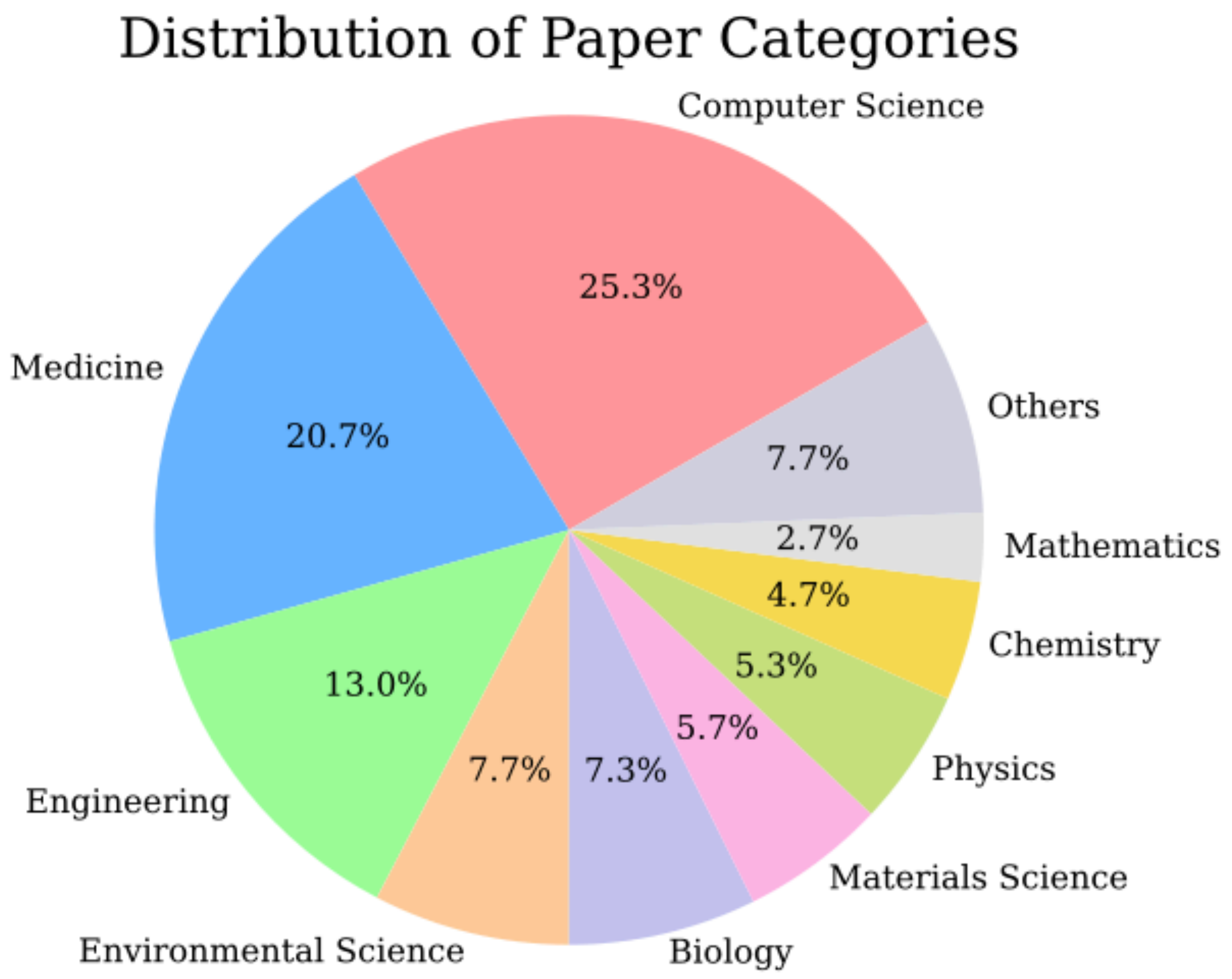


Data

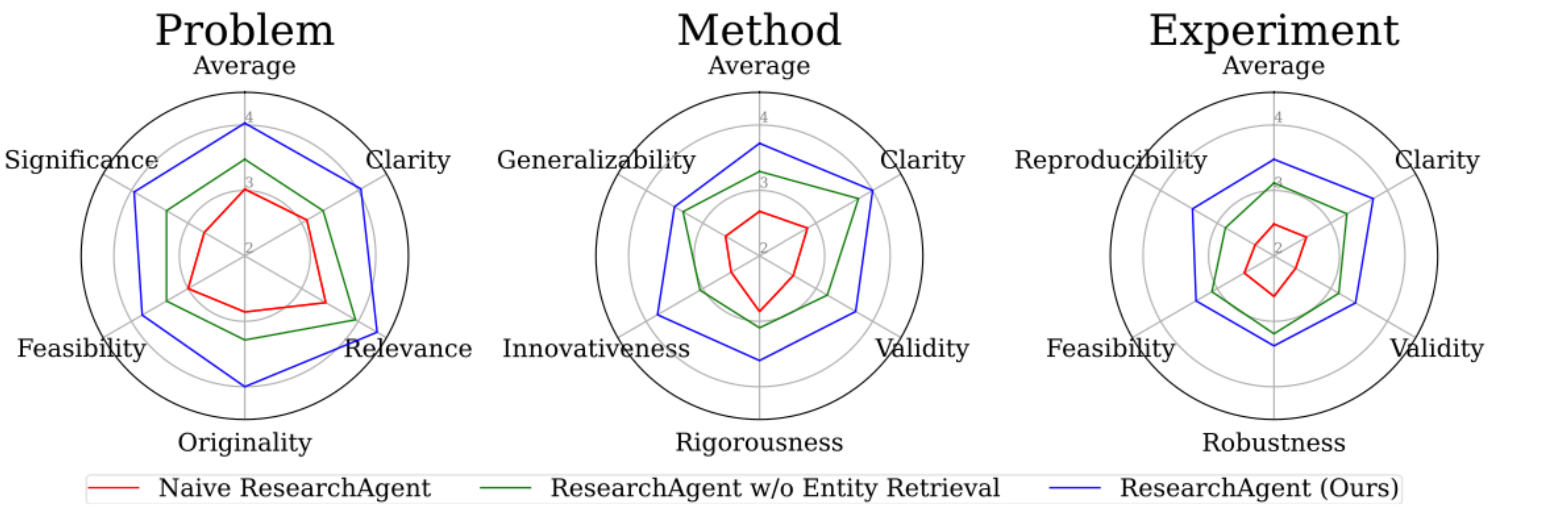
➡ Semantic Scholar Academic Graph API

➡ Papers appearing after May 01, 2023

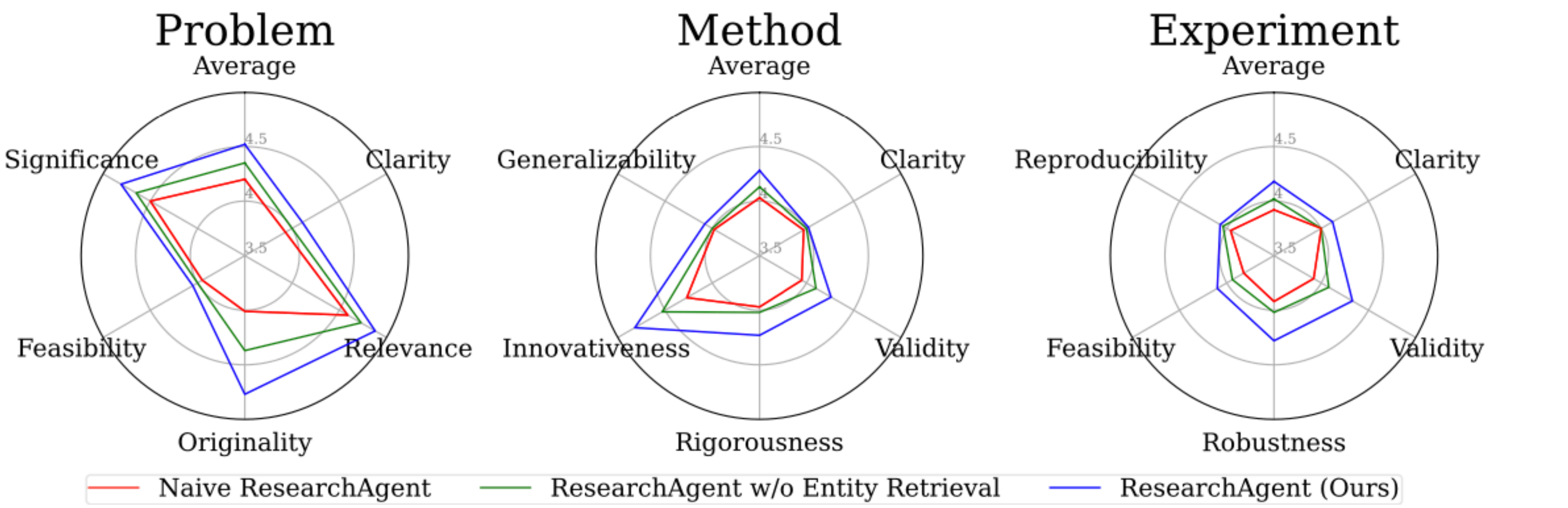
- Unavailable to GPT-4
- Select high-impact ones
- 87 references on average
- 2.17 entities on average
- Serve as core paper



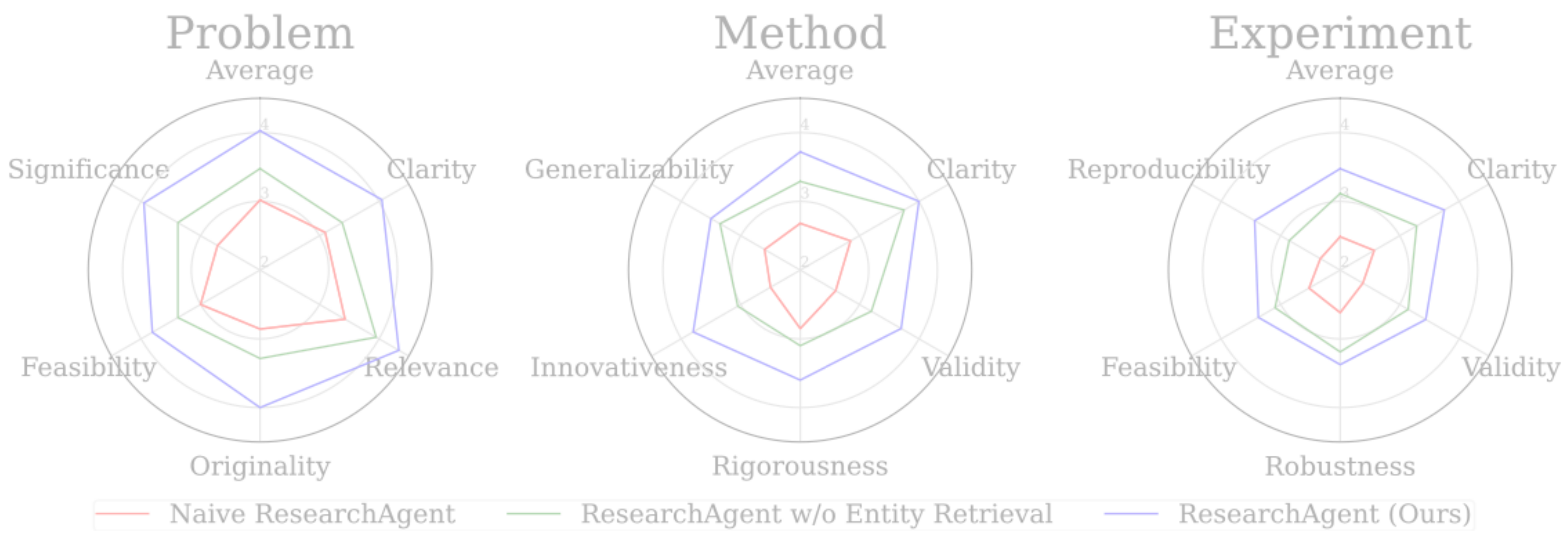
Evaluation



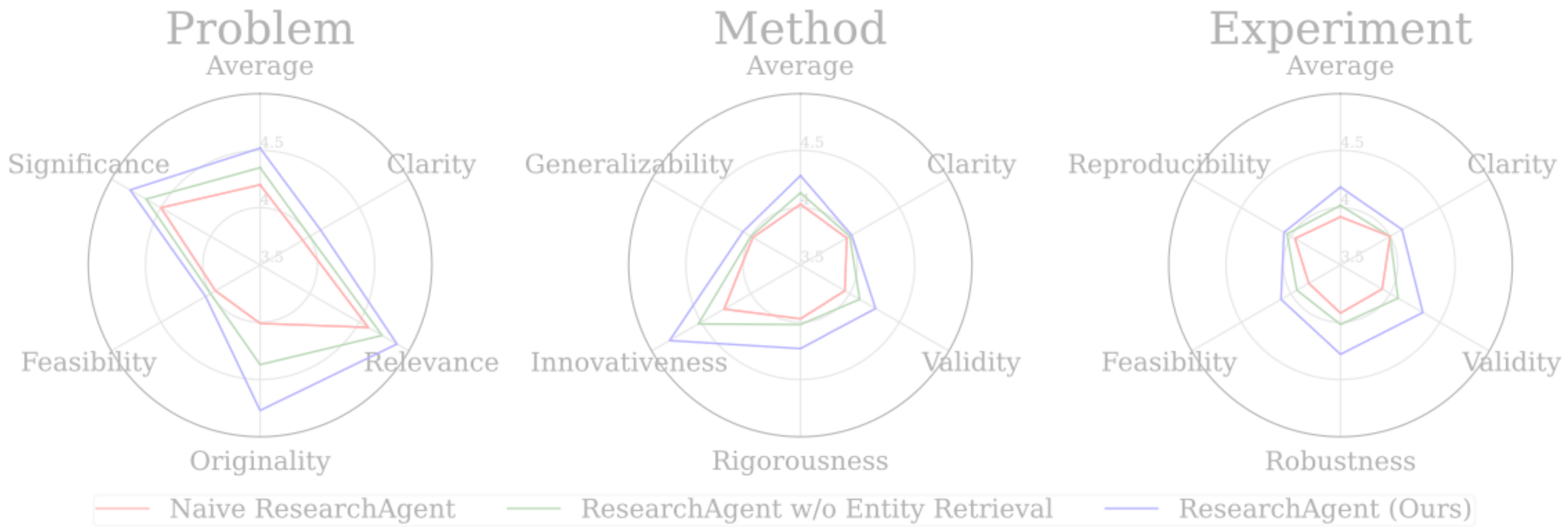
(a) Human Evaluation



Evaluation



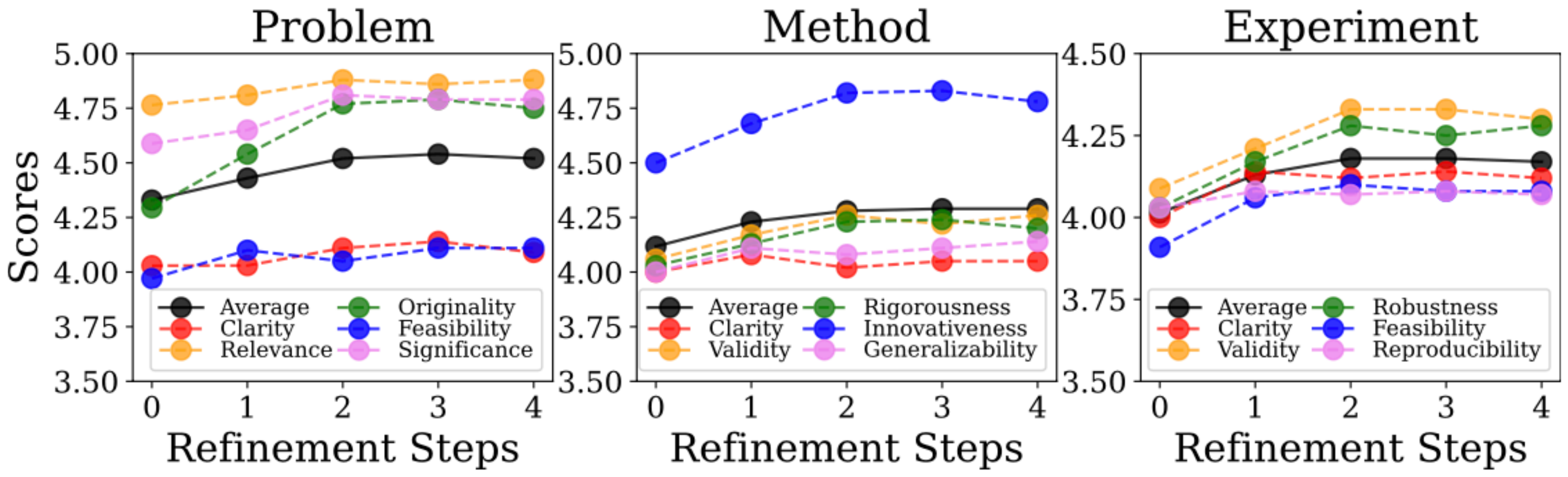
(a) Human Evaluation



Categories	Metrics	Problem	Method	Experiment
Human and Human	Scoring	0.83	0.76	0.67
	Pairwise	0.62	0.62	0.41
Human and Model	Scoring	0.64	0.58	0.49
	Pairwise	0.71	0.62	0.52

Evaluation

Categories	Metrics	Problem	Method	Experiment
Human and Human	Scoring	0.83	0.76	0.67
	Pairwise	0.62	0.62	0.41
Human and Model	Scoring	0.64	0.58	0.49
	Pairwise	0.71	0.62	0.52



Ablation Study

➡ Naive ResearchAgent

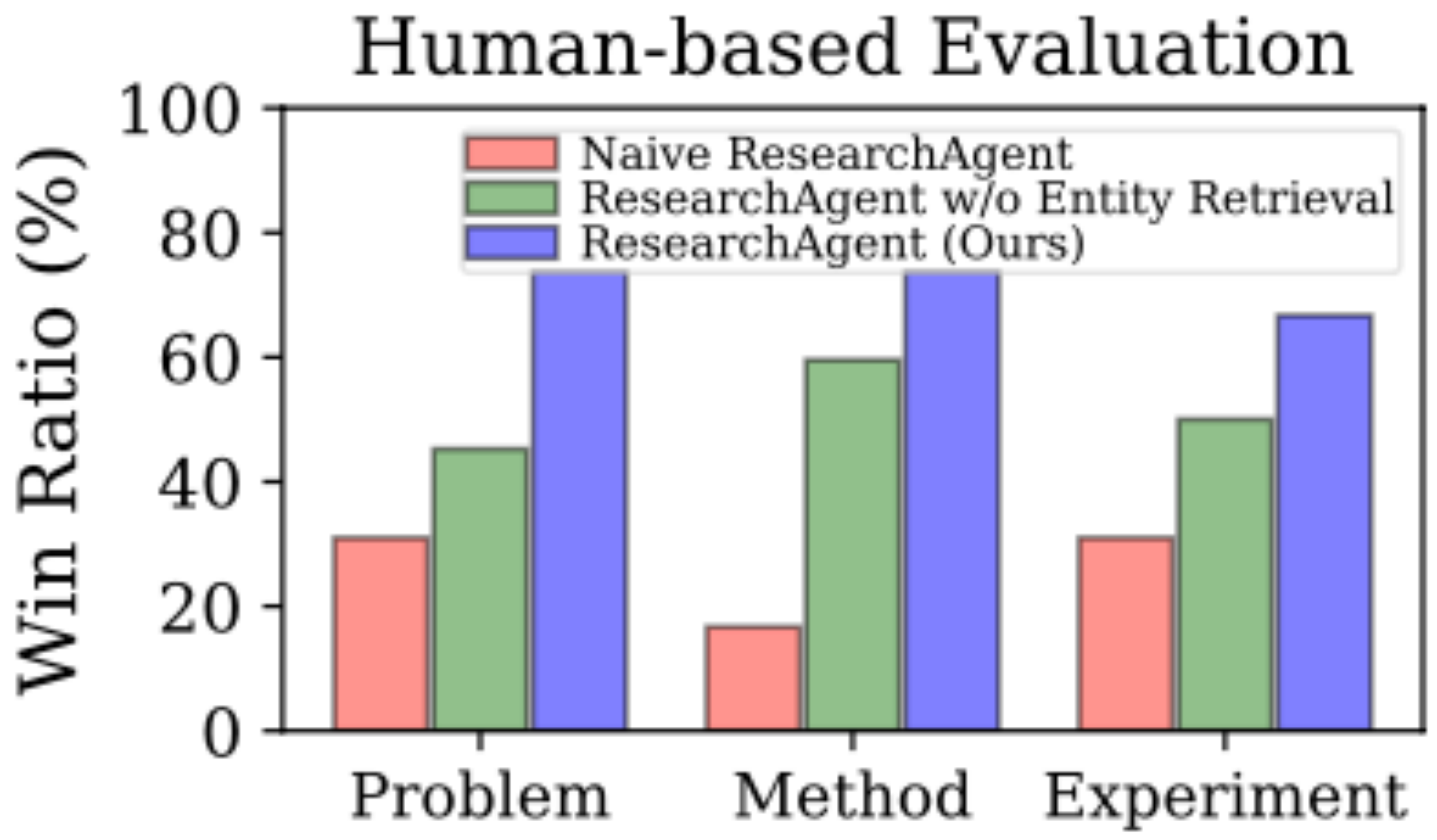
- Uses only a core paper to generate research ideas.

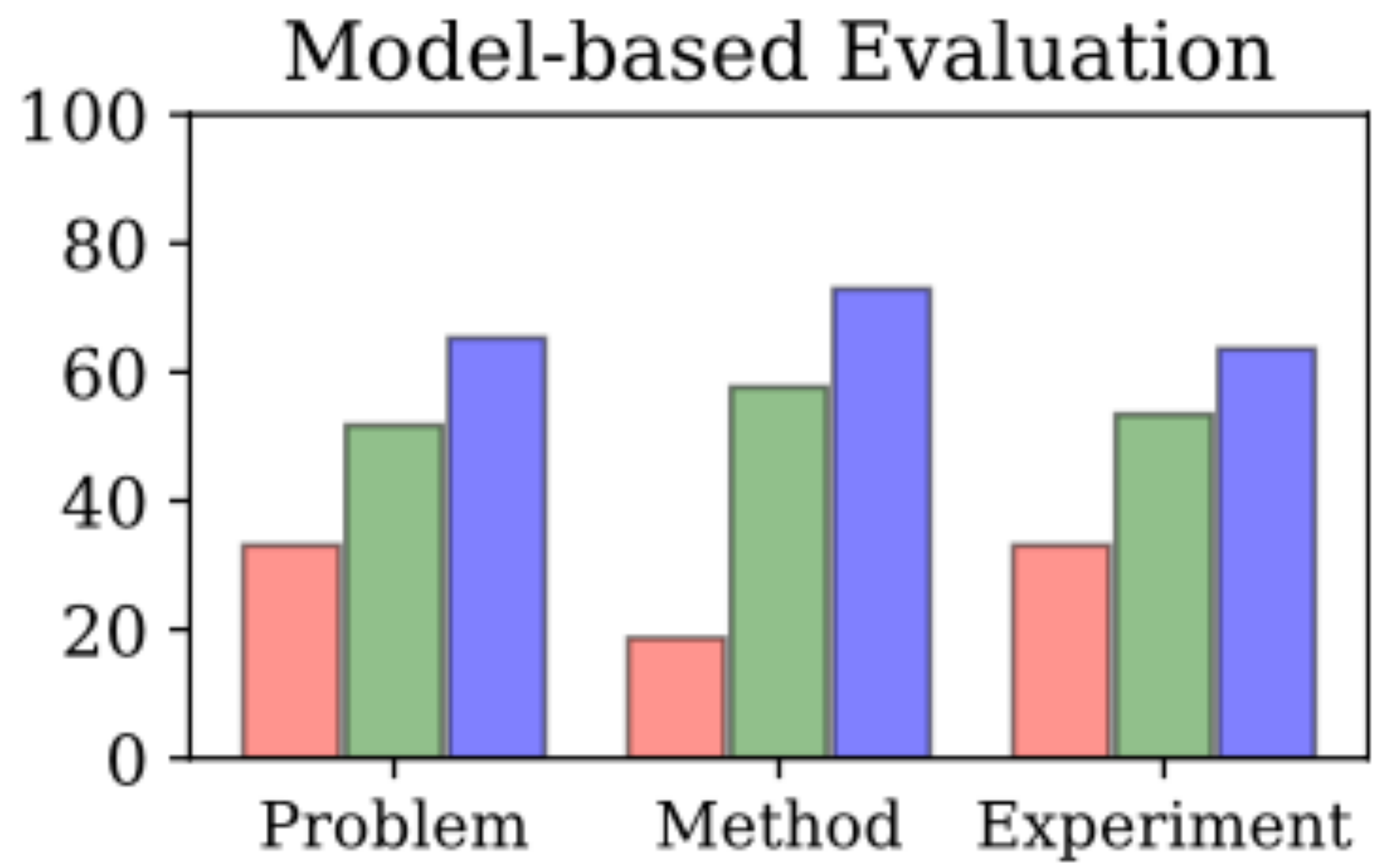
➡ ResearchAgent w/o Entity Retrieval

- Uses the core paper and its relevant references without considering entities.

➡ ResearchAgent

- Full model.





Ablation Study

➡ ResearchAgent

- Entities
- References
- Entities & References

Methods	Problem	Method	Experiment
ResearchAgent	**4.52**	**4.28**	**4.18**
- w/o Entities	4.35	4.13	4.02
- w/ Random Entities	4.41	4.19	4.13
- w/o References	4.26	4.08	3.97
- w/ Random References	4.35	4.16	4.02
- w/o Entities & References	4.20	4.03	3.92

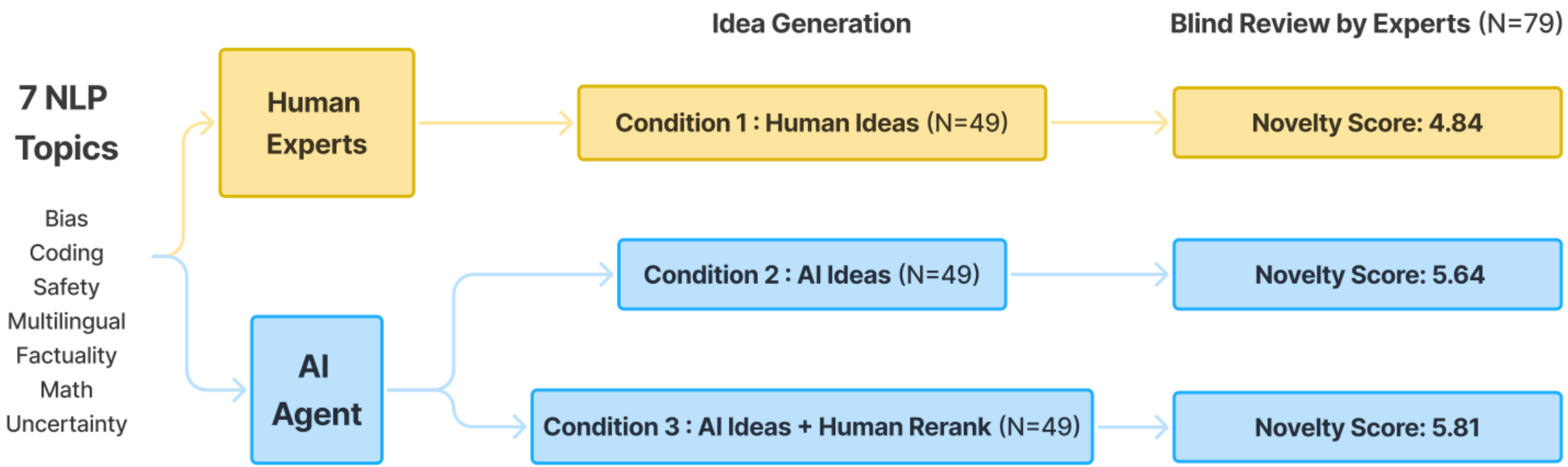
Agenda

- ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models [NAACL 2025]
- Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers [ICLR 2025]
- Nova: An Iterative Planning and Search Approach to Enhance Novelty and Diversity of LLM Generated Ideas [arXiv 2024]

Motivation

- ➡ LMs are increasingly used for scientific ideation
- ➡ But: Can they truly generate **expert-level novel** research ideas?
- ➡ Prior work lacked rigorous comparison against human experts

Overview


Idea Generation
Blind Review by Experts (N=79)
7 NLP Topics
Bias
Coding
Safety
Multilingual
Factuality
Math
Uncertainty
Human Experts
AI Agent
Condition 1 : Human Ideas (N=49)
Condition 2 : AI Ideas (N=49)
Condition 3 : AI Ideas + Human Rerank (N=49)
Novelty Score: 4.84
Novelty Score: 5.64
Novelty Score: 5.81

Idea Generation Agent

➡ Paper Retrieval for RAG

- Given a research topic, prompt an LLM to generate a sequence of function calls to the Semantic Scholar API.
- Action space: `{KeywordQuery(keywords), PaperQuery(paperId), GetReferences(paperId)}`
- Use the LLM to score (1 to 10) and rerank all retrieved papers.

Idea Generation Agent

➡ Paper Retrieval for RAG

➡ Idea Generation

- Prompt the LLM to generate 4000 seed ideas on each research topic.
- Manually summarized exemplar papers + Retrieved papers.
- Remove duplications (5% left).

Idea Generation Agent

➡ Paper Retrieval for RAG

➡ Idea Generation

➡ Idea Ranking

- Choose Claude-3.5-Sonnet as the zero-shot ranker.

Idea Generation Agent

➡ Paper Retrieval for RAG

➡ Idea Generation

➡ Idea Ranking

- Choose Claude-3.5-Sonnet as the zero-shot ranker.
- Swiss System Tournament

The **Swiss System Tournament** is an iterative ranking method where items (e.g., research ideas) are **paired against others with similar scores**, and each "win" increases their score. After several rounds, the most consistently high-performing items rise to the top. It's efficient and fair for large sets.

Idea Generation Agent

➡ Paper Retrieval for RAG

➡ Idea Generation

➡ Idea Ranking

- Choose Claude-3.5-Sonnet as the zero-shot ranker.
- Swiss System Tournament.
- Another condition: Human Rerank.

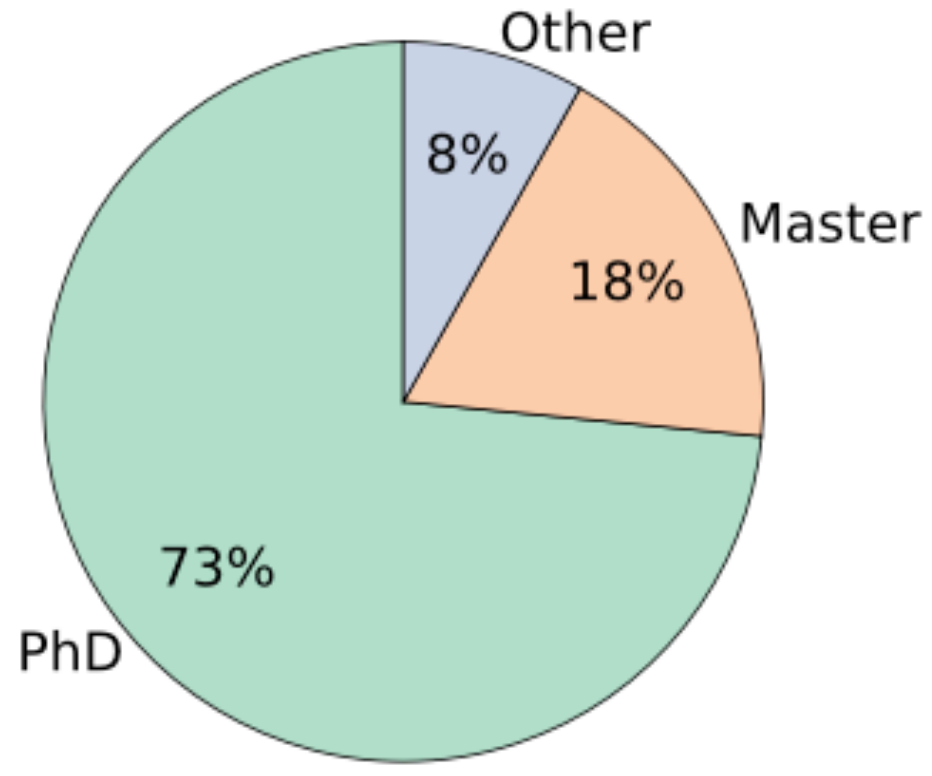
Expert Idea Writing

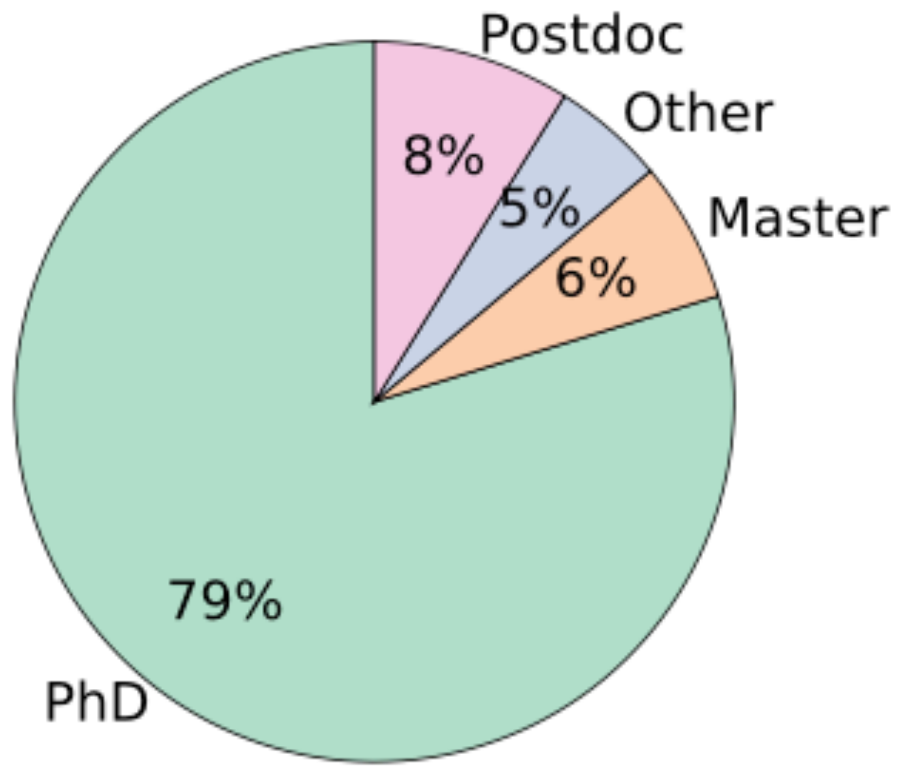
➡ Expert Recruitment

- N = 49 for writing ideas.
- N = 79 for reviewing ideas.
- 24 overlaps, N = 104 in total.

Expert Idea Writing

- ➡ Expert Recruitment
- ➡ Expert Qualifications





	Idea Writing Participants (N=49)					Idea Reviewing Participants (N=79)				
Metric	Mean	Median	Min	Max	SD	Mean	Median	Min	Max	SD
papers	12	10	2	52	9	15	13	2	52	10
citations	477	125	2	4553	861	635	327	0	7276	989
h-index	5	4	1	21	4	7	7	0	21	4
i10-index	5	4	0	32	6	7	5	0	32	6

Expert Idea Writing

- ➡ Expert Recruitment
- ➡ Expert Qualifications
- ➡ Idea Writing

Metric	Mean	Median	Min	Max	SD
Human Ideas					
Familiarity (1-5)	3.7	4.0	1.0	5.0	1.0
Difficulty (1-5)	3.0	3.0	1.0	5.0	0.7
Time (Hours)	5.5	5.0	2.0	15.0	2.7
Length (Words)	901.7	876.0	444.0	1704.0	253.5
AI Ideas					
Length (Words)	1186.3	1158.0	706.0	1745.0	233.7
AI + Human Rerank Ideas					
Length (Words)	1174.0	1166.0	706.0	1708.0	211.0

Topic	Count
Bias	4
Coding	9
Safety	5
Multilingual	10
Factuality	11
Math	4
Uncertainty	6
Total	49

Expert Idea Writing

- ➡ Expert Recruitment
- ➡ Expert Qualifications
- ➡ Idea Writing
- ➡ Idea Reviewing

Metric	Mean	Median	Min	Max	SD
Ours					
Familiarity (1-5)	3.7	3.0	1.0	5.0	0.9
Confidence (1-5)	3.7	4.0	1.0	5.0	0.7
Time (Minutes)	31.7	30.0	5.0	120.0	16.8
Length (Word)	231.9	208.0	41.0	771.0	112.1
ICLR 2024					
Confidence (1-5)	3.7	4.0	1.0	5.0	0.8
Length (Word)	421.5	360.0	14.0	2426.0	236.4
Length (Word; Strengths & Weaknesses)	247.4	207.0	2.0	2010.0	176.4

Metric	Mean	Min	Max	SD
# Reviews	3.8	2.0	7.0	1.3
# Conditions	2.5	2.0	3.0	0.5
# Topics	1.5	1.0	3.0	0.6

Evaluation

➡ Treating Each Review as an Independent Datapoint

Condition	Size	Mean	Median	SD	SE	Min	Max	p-value
Novelty Score								
Human Ideas	119	4.84	5	1.79	0.16	1	8	–
AI Ideas	109	5.64	6	1.76	0.17	1	10	**0.00****
AI Ideas + Human Rerank	109	5.81	6	1.66	0.16	2	10	**0.00*****
Excitement Score								
Human Ideas	119	4.55	5	1.89	0.17	1	8	–
AI Ideas	109	5.19	6	1.73	0.17	1	9	**0.04***
AI Ideas + Human Rerank	109	5.46	6	1.82	0.17	1	9	**0.00****
Feasibility Score								
Human Ideas	119	6.61	7	1.99	0.18	1	10	–
AI Ideas	109	6.34	6	1.88	0.18	2	10	1.00
AI Ideas + Human Rerank	109	6.44	6	1.63	0.16	1	10	1.00
Expected Effectiveness Score								
Human Ideas	119	5.13	5	1.76	0.16	1	8	–
AI Ideas	109	5.47	6	1.58	0.15	1	10	0.67
AI Ideas + Human Rerank	109	5.55	6	1.52	0.15	1	9	0.29
Overall Score								
Human Ideas	119	4.68	5	1.90	0.17	1	9	–
AI Ideas	109	4.85	5	1.70	0.16	1	9	1.00
AI Ideas + Human Rerank	109	5.34	6	1.79	0.17	1	9	**0.04***

Evaluation

➡ Treating Each Reviewer as an Independent Datapoint

	N	Mean Diff	p-value
Novelty Score			
`AI Ideas` vs `Human Ideas`	70	0.94	**0.00****
`AI Ideas + Human Rerank` vs `Human Ideas`	65	0.86	**0.00****
Excitement Score			
`AI Ideas` vs `Human Ideas`	70	0.73	**0.01***
`AI Ideas + Human Rerank` vs `Human Ideas`	65	0.87	**0.00****
Feasibility Score			
`AI Ideas` vs `Human Ideas`	70	-0.29	0.36
`AI Ideas + Human Rerank` vs `Human Ideas`	65	-0.08	0.74
Effectiveness Score			
`AI Ideas` vs `Human Ideas`	70	0.42	0.16
`AI Ideas + Human Rerank` vs `Human Ideas`	65	0.39	0.16
Overall Score			
`AI Ideas` vs `Human Ideas`	70	0.24	0.36
`AI Ideas + Human Rerank` vs `Human Ideas`	65	0.66	**0.01***

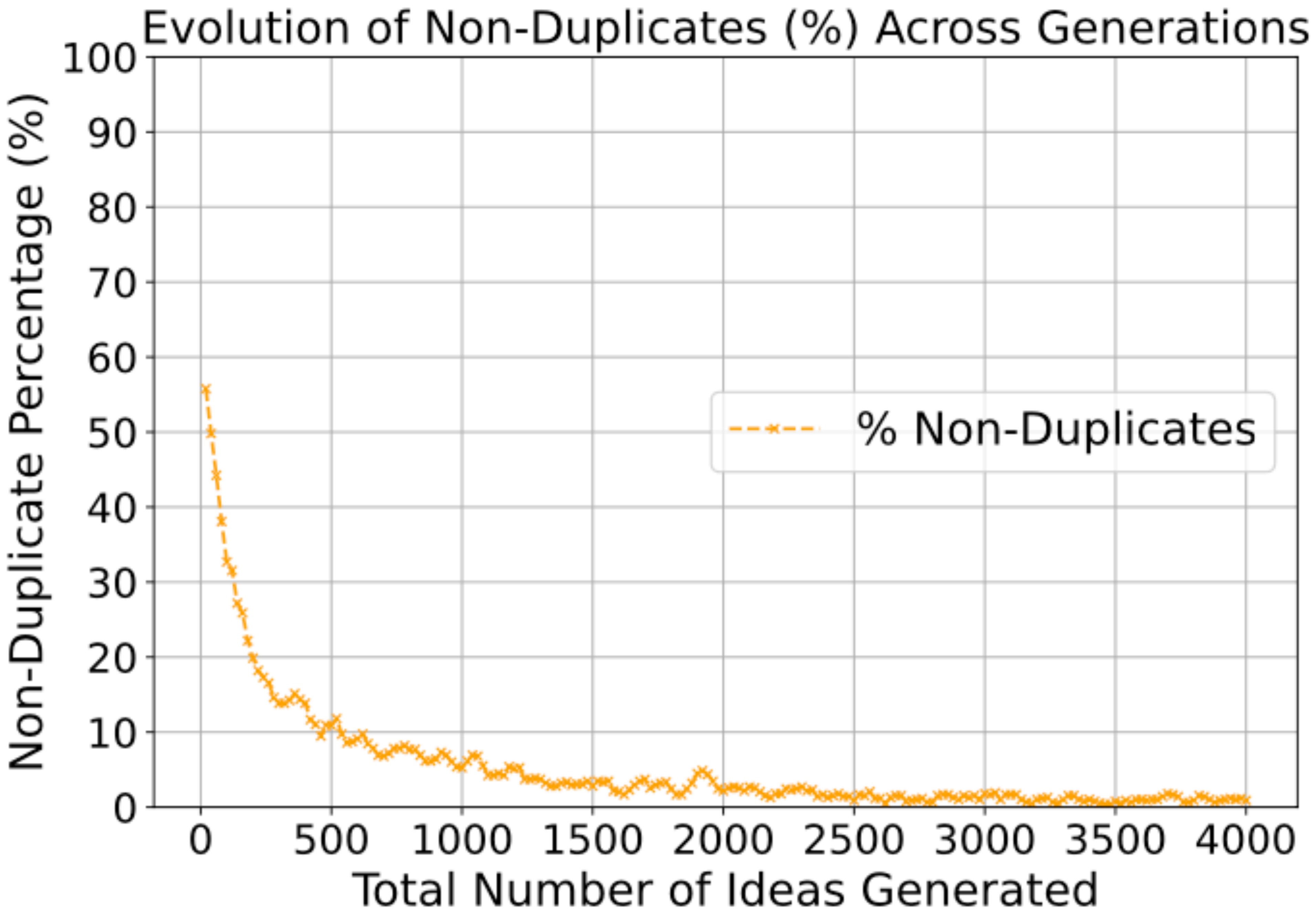
Key Findings

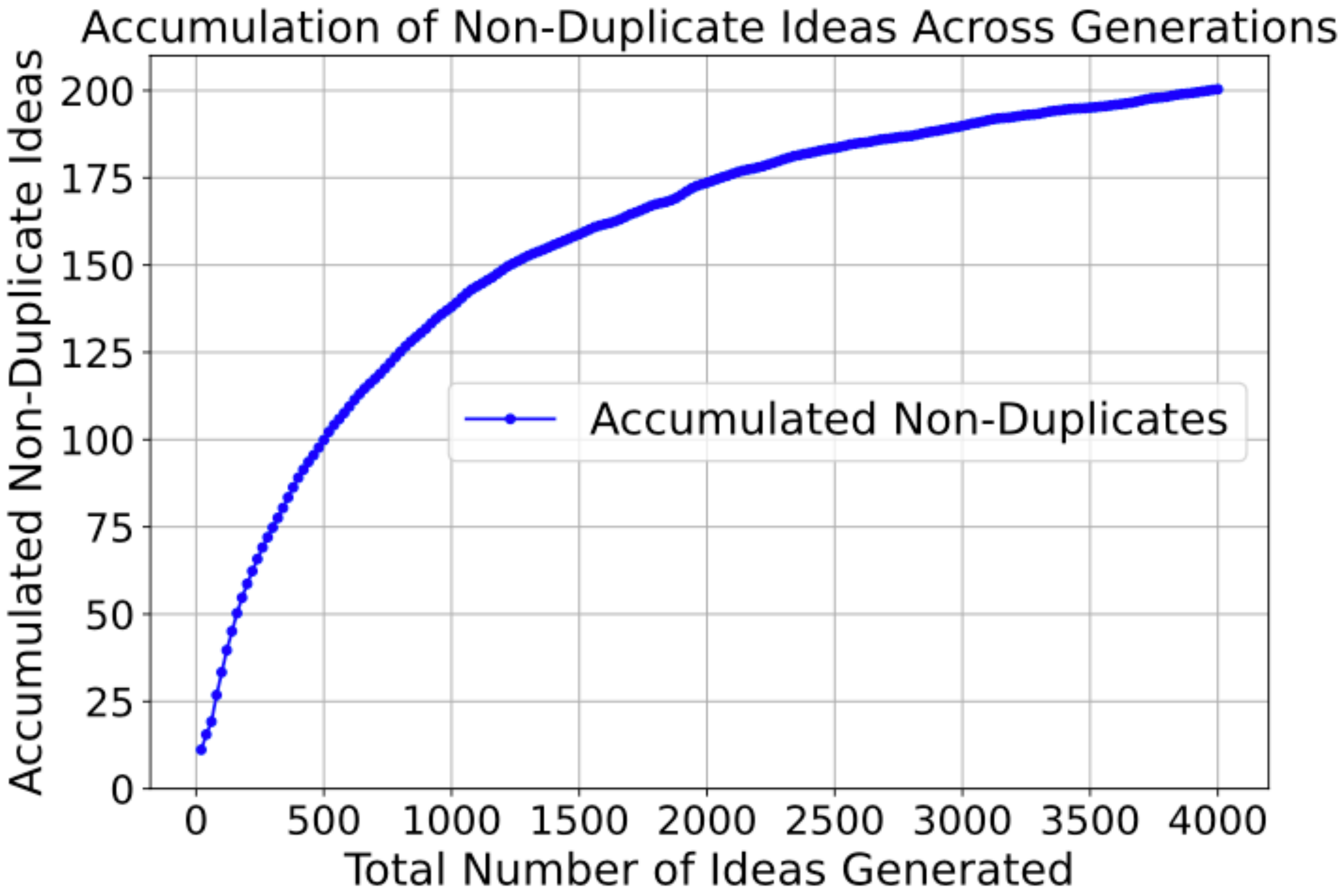
➡ LLM-generated ideas are judged as **more novel** ($p < 0.05$) than human expert ideas while being judged slightly weaker on feasibility.

Key Findings

➡ LLM-generated ideas are judged as **more novel** ($p < 0.05$) than human expert ideas while being judged slightly weaker on feasibility.

➡ LLMs lack diversity in idea generation.





Key Findings

➡ LLM-generated ideas are judged as **more novel** ($p < 0.05$) than human expert ideas while being judged slightly weaker on feasibility.

➡ LLMs lack diversity in idea generation.

➡ LLMs cannot evaluate ideas reliably.

Topic	Overlap	New
Bias	2	2
Coding	4	5
Safety	2	3
Multilingual	5	5
Factuality	2	9
Math	2	2
Uncertainty	1	5
Total	18	31

	Consistency
Random	50.0
NeurIPS'21	66.0
ICLR'24	71.9
Ours	56.1
GPT-4o Direct	50.0
GPT-4o Pairwise	45.0
Claude-3.5 Direct	51.7
Claude-3.5 Pairwise	53.3
"AI Scientist" Reviewer	43.3

Agenda

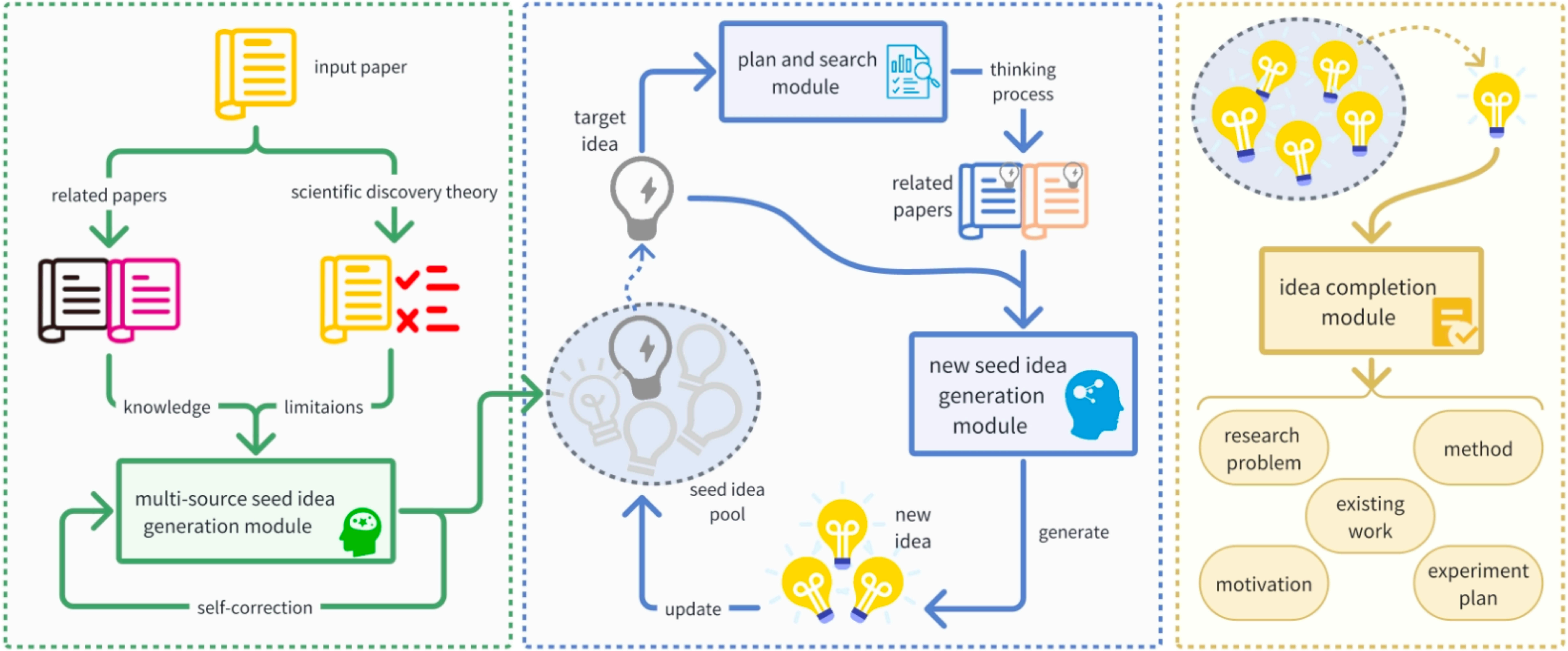
- ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models [NAACL 2025]
- Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers [ICLR 2025]
- Nova: An Iterative Planning and Search Approach to Enhance Novelty and Diversity of LLM Generated Ideas [arXiv 2024]

Motivation

➡ LLMs lack diversity in idea generation.

- Constrained scope.
- Lack of direction in knowledge acquisition.

Nova Pipeline

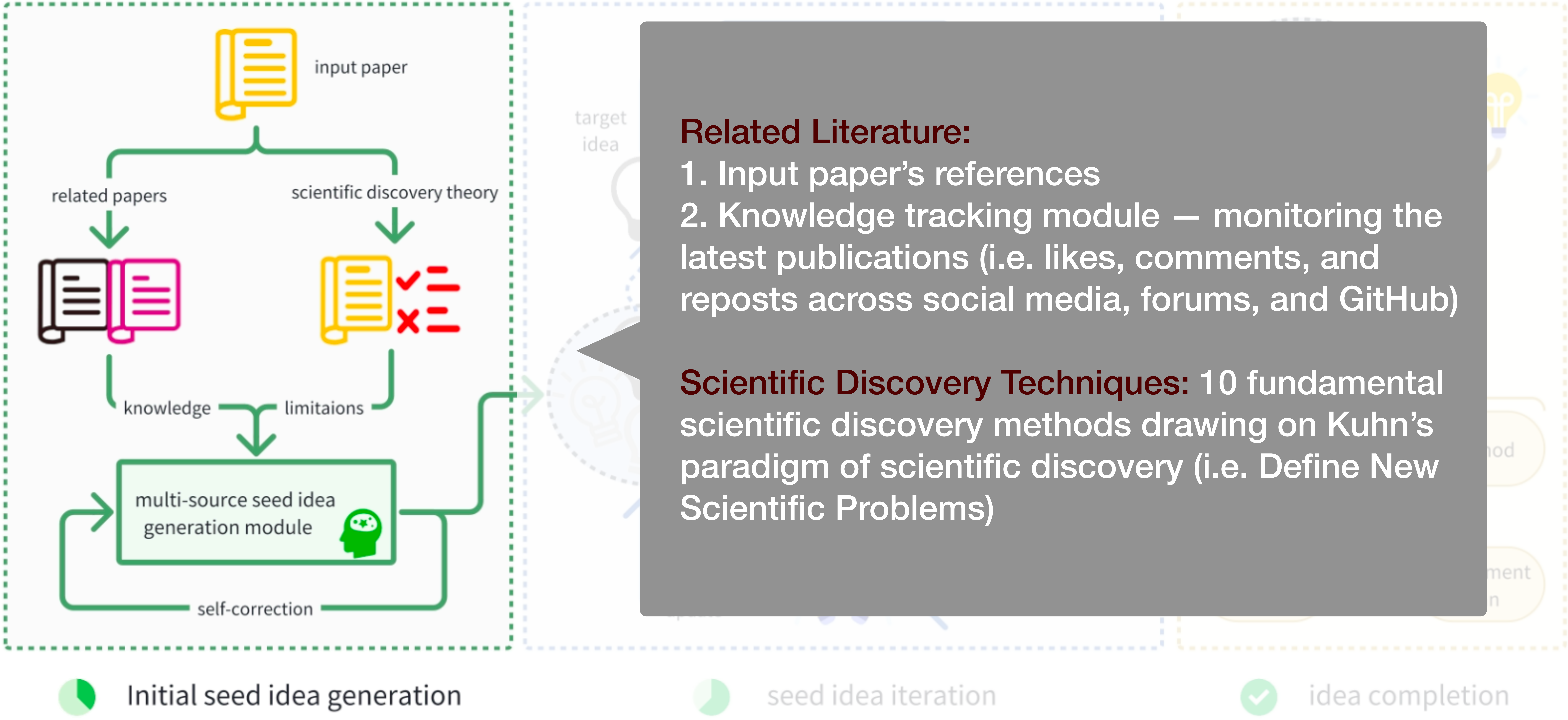


Initial seed idea generation

seed idea iteration

idea completion

Nova Pipeline

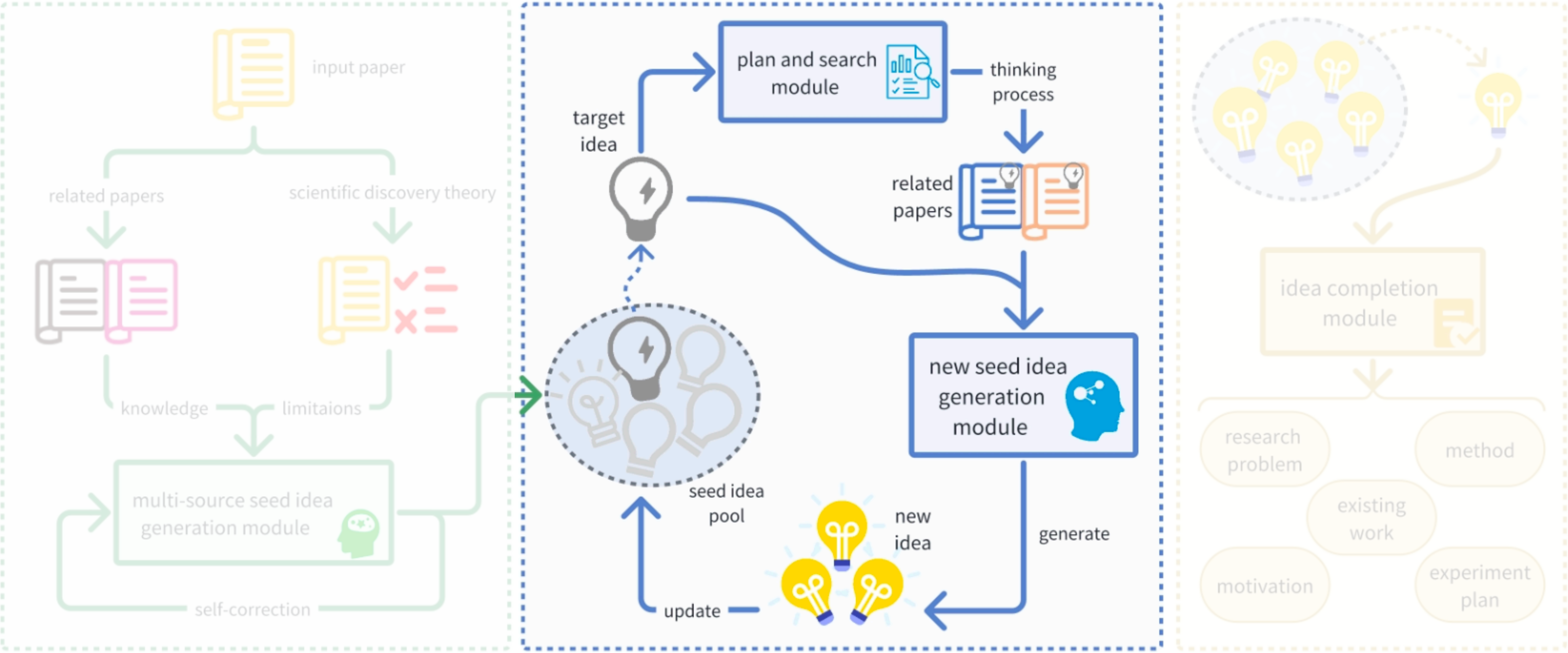


Related Literature:

1. Input paper's references
2. Knowledge tracking module — monitoring the latest publications (i.e. likes, comments, and reposts across social media, forums, and GitHub)

Scientific Discovery Techniques: 10 fundamental scientific discovery methods drawing on Kuhn's paradigm of scientific discovery (i.e. Define New Scientific Problems)

Nova Pipeline

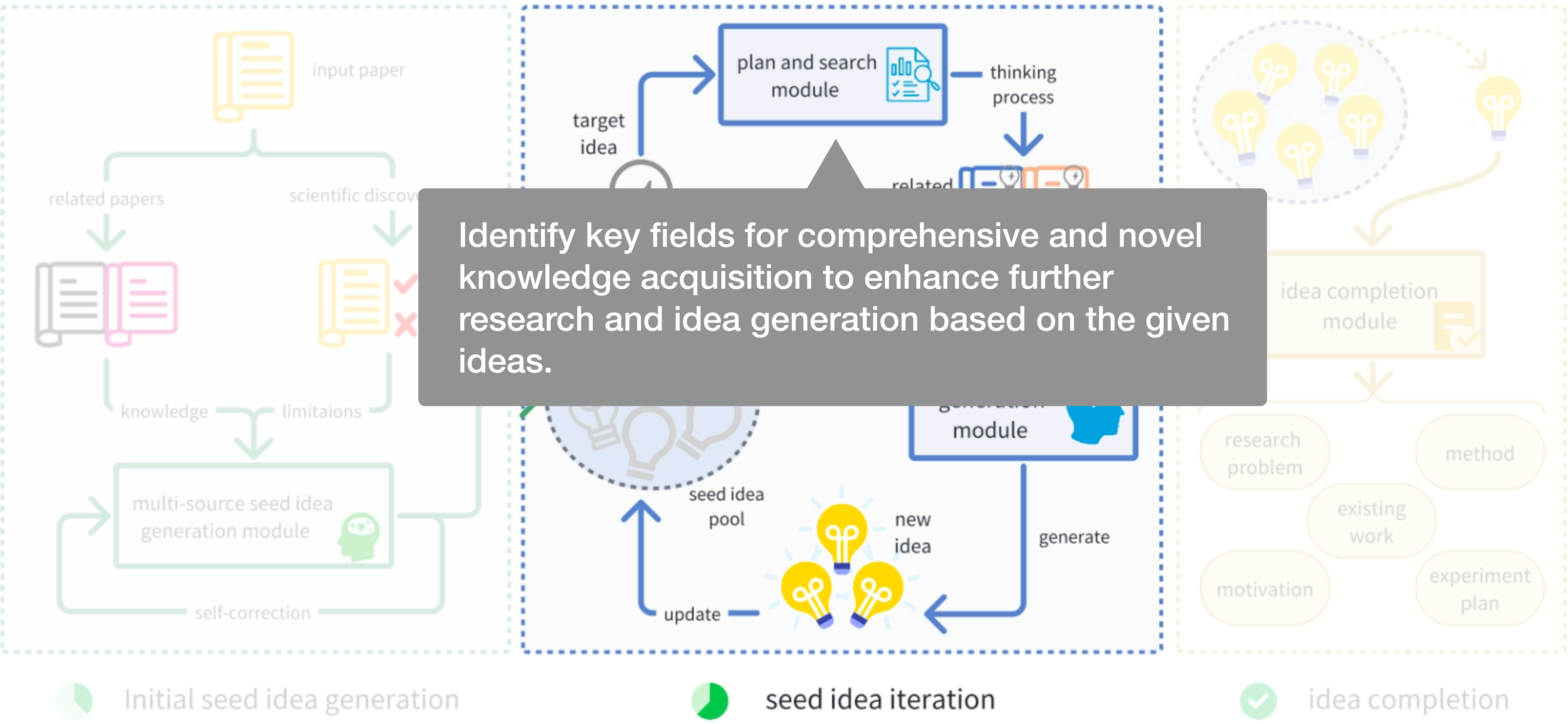


Initial seed idea generation

seed idea iteration

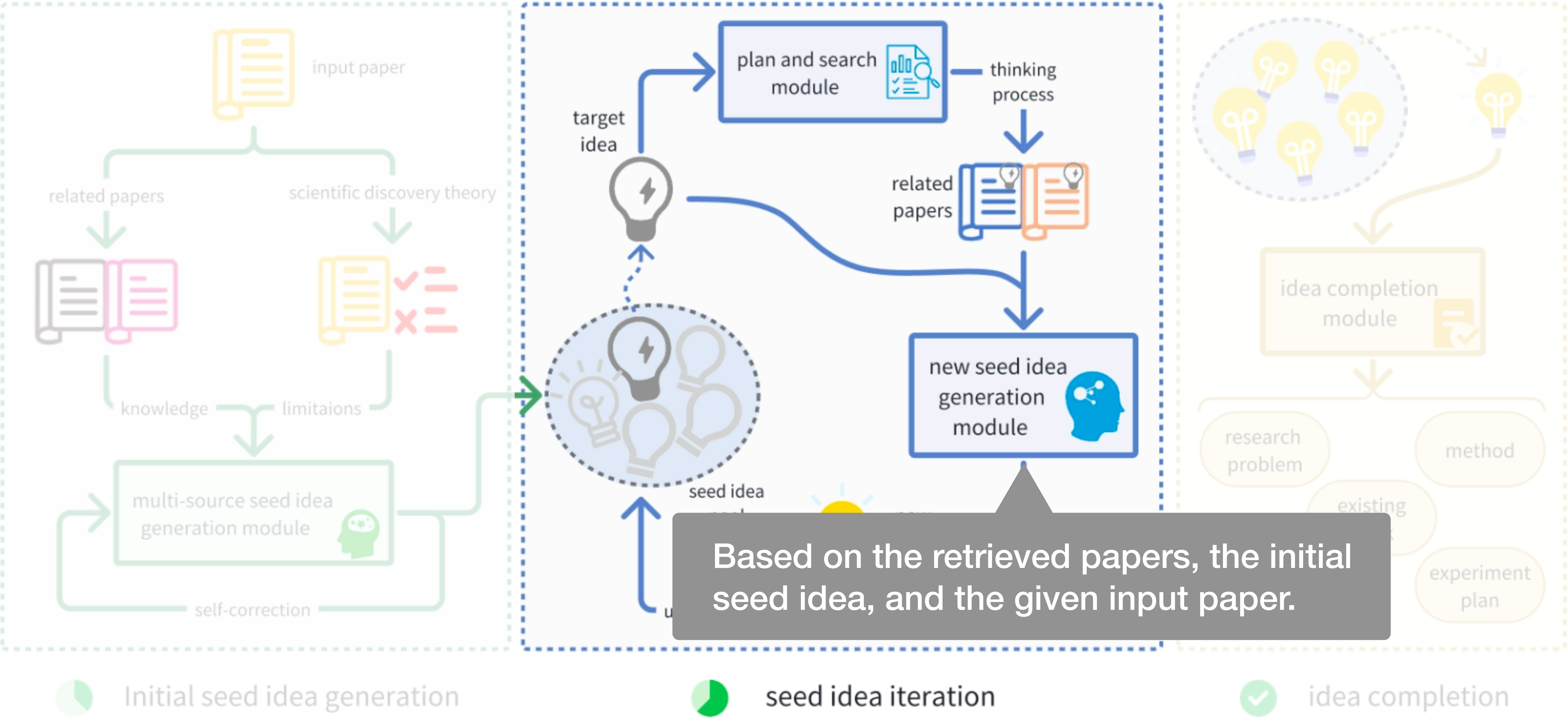
idea completion

Nova Pipeline



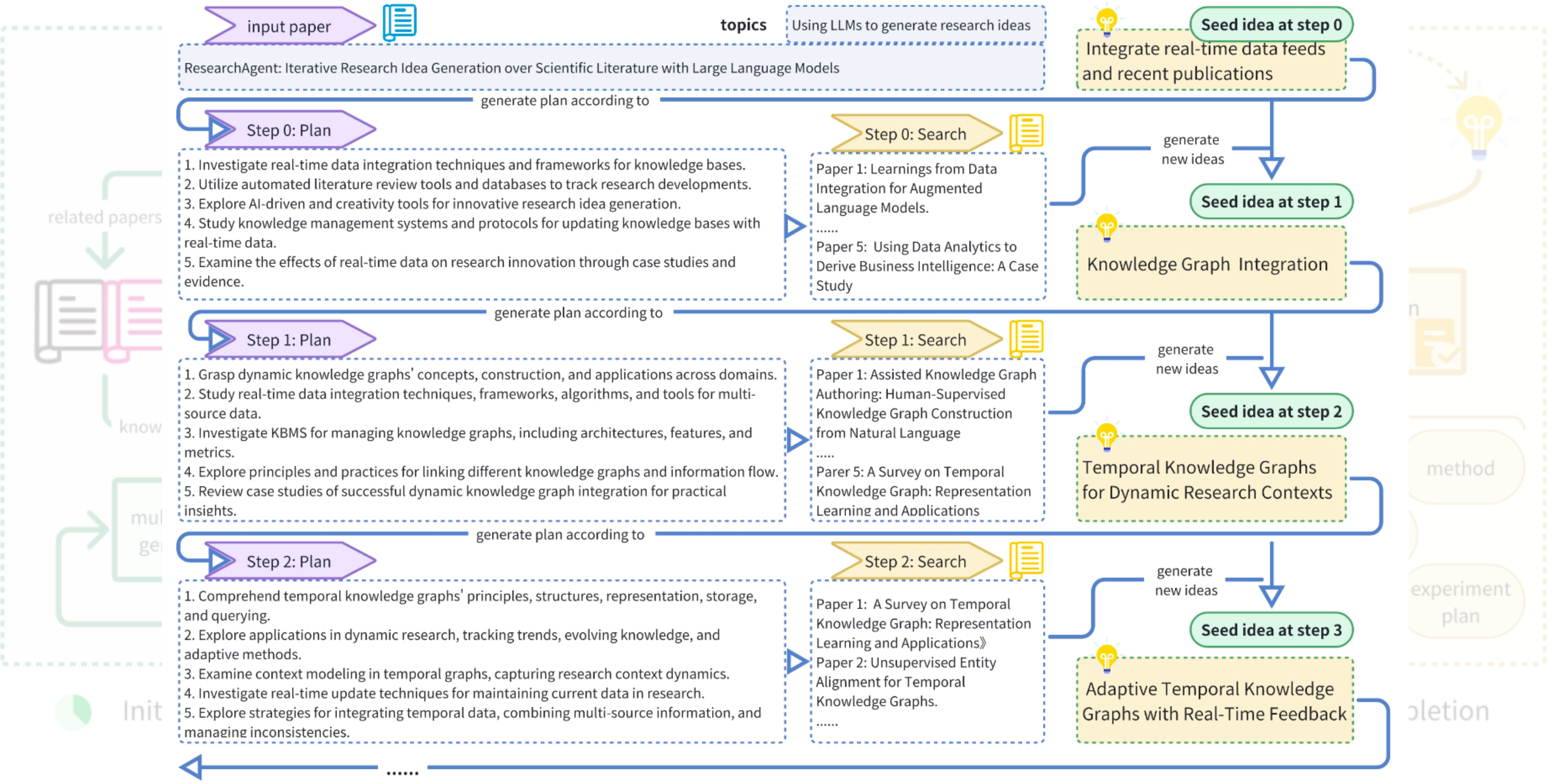
Identify key fields for comprehensive and novel knowledge acquisition to enhance further research and idea generation based on the given ideas.

Nova Pipeline

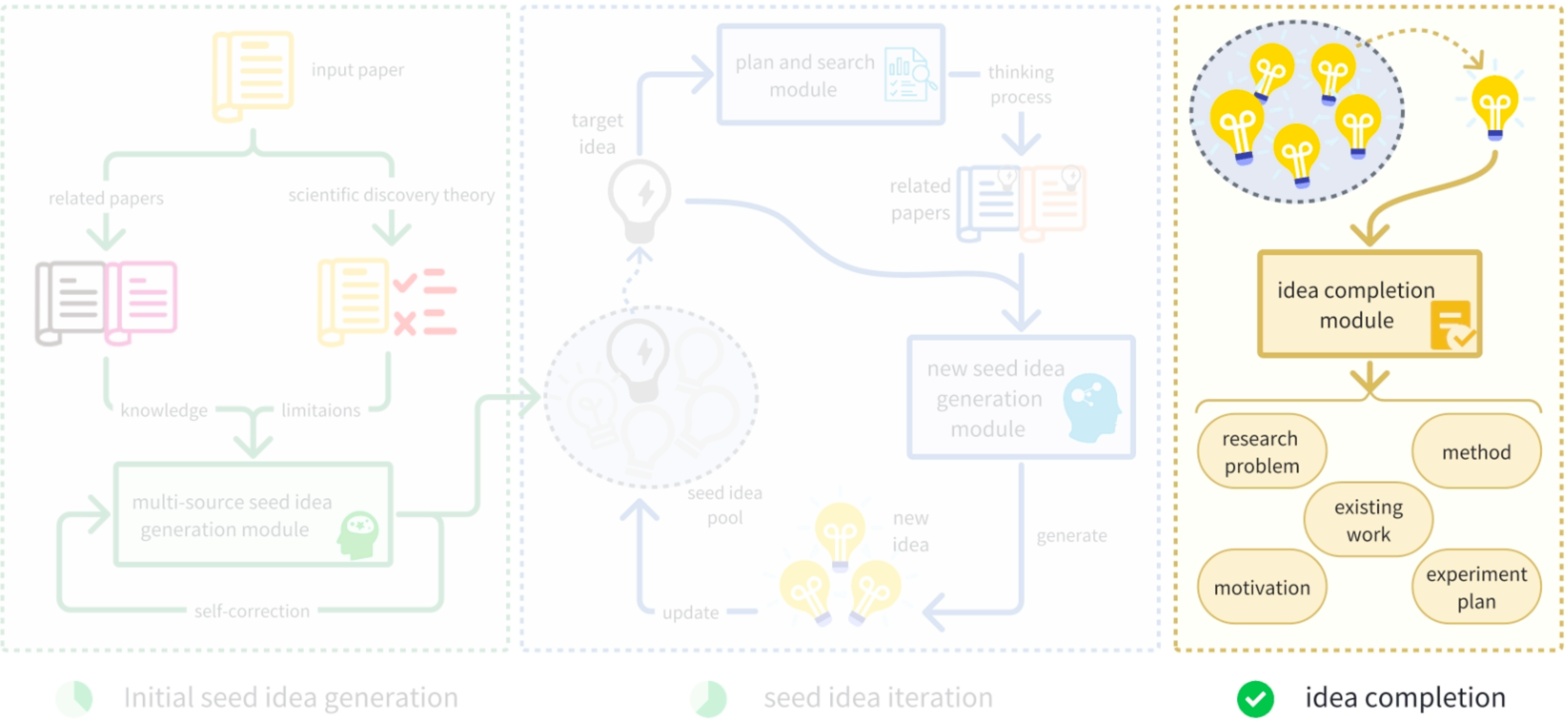


Based on the retrieved papers, the initial seed idea, and the given input paper.

Nova Pipeline



Nova Pipeline

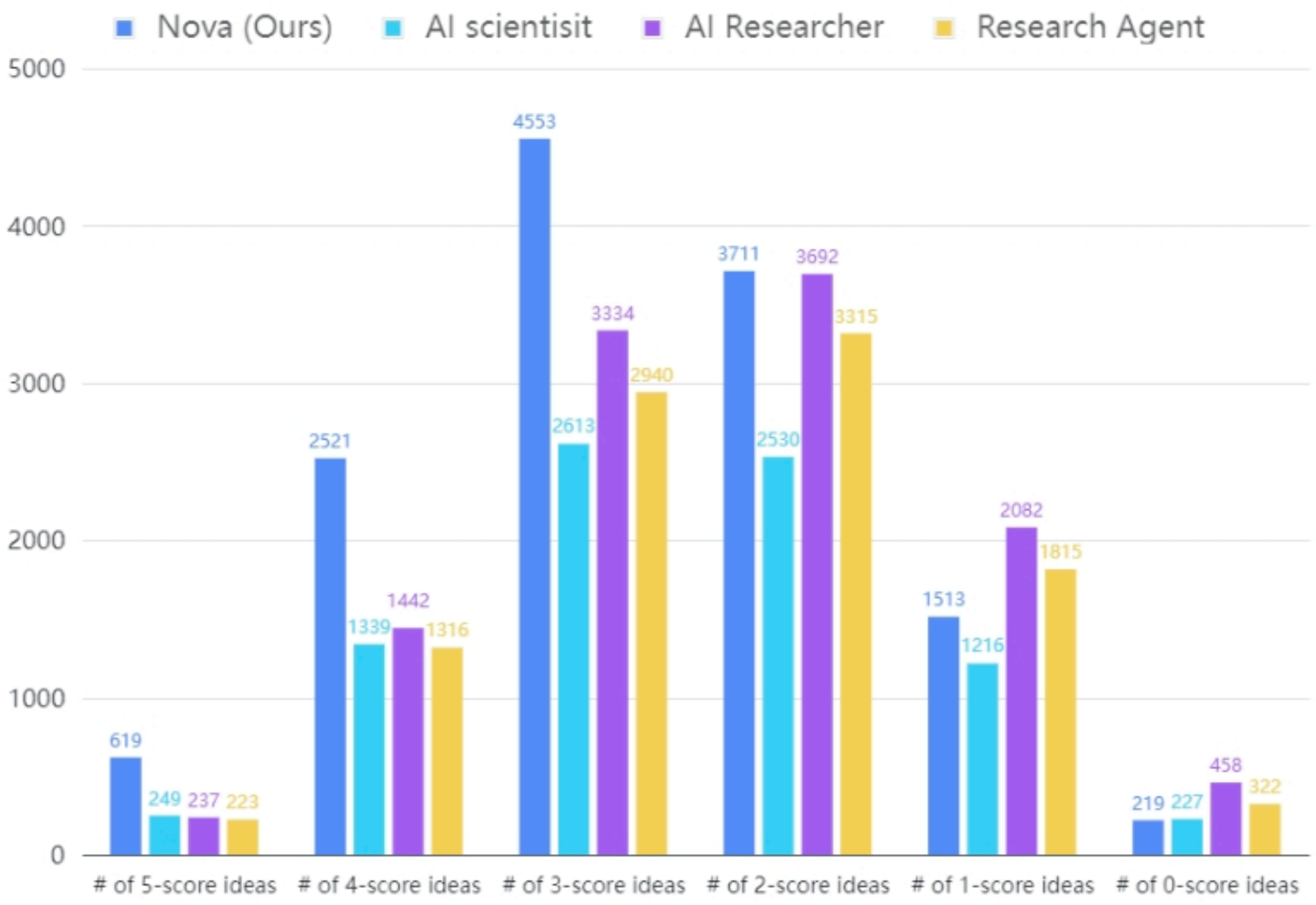


Data

- ➡ Papers from CVPR 2024, ACL 2024, ICLR 2024, and Hugging Face Daily Papers.
- ➡ With keywords related to “LLM”.

Evaluation

➡ Automatic Evaluation

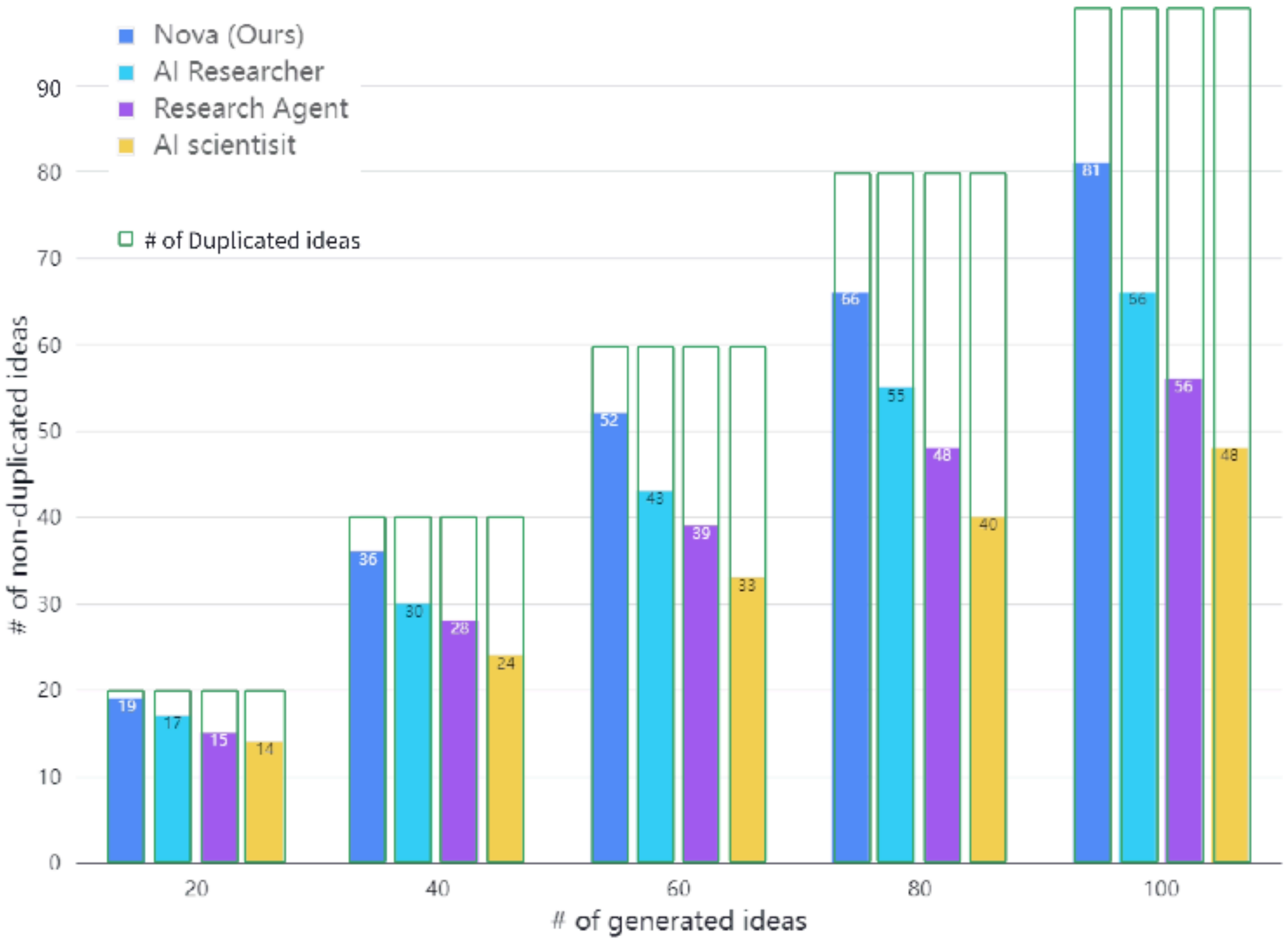


Swiss Tournament score
(quality evaluation)

Evaluation

➡ Automatic Evaluation

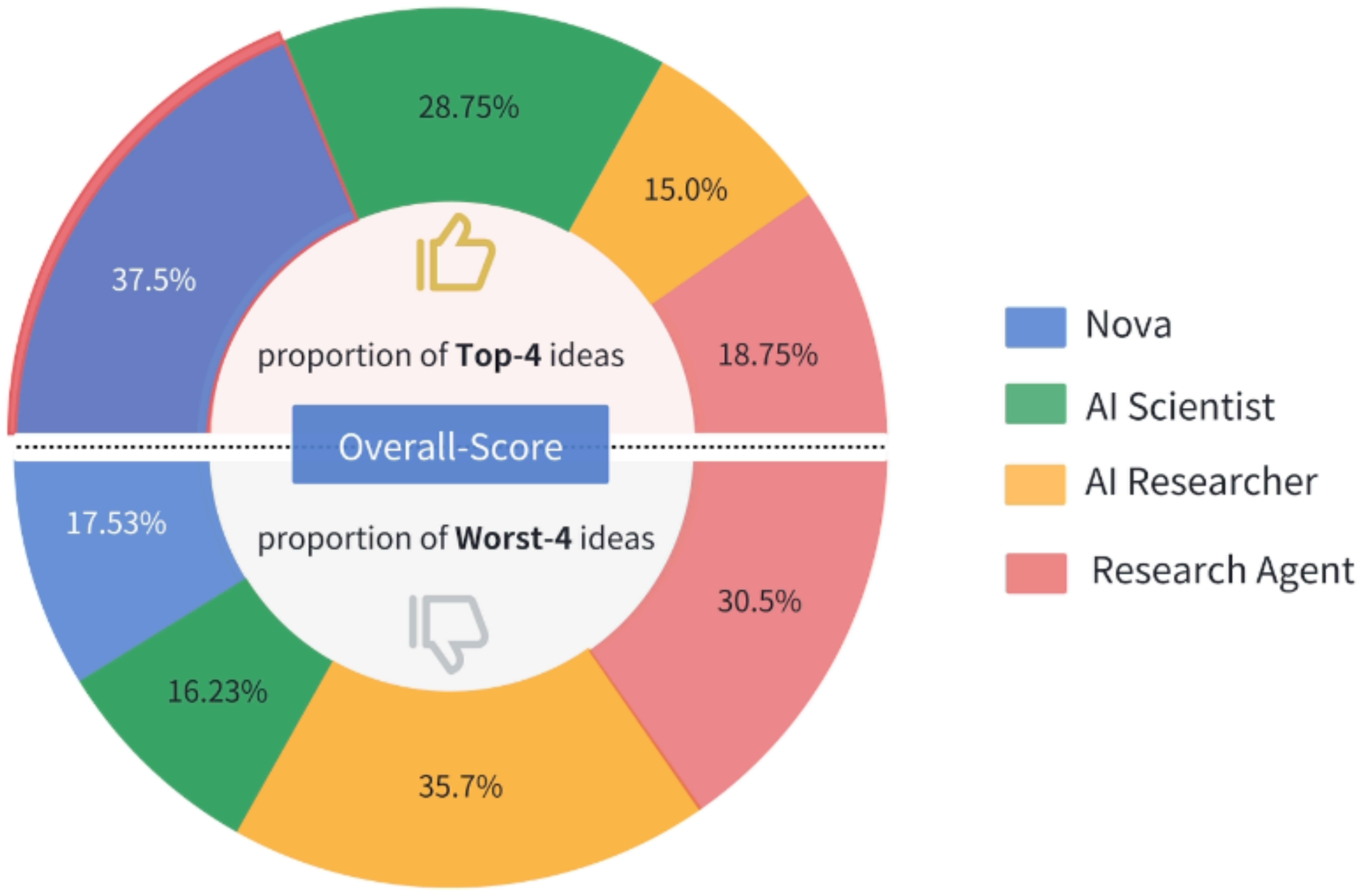
Novelty & Diversity comparison

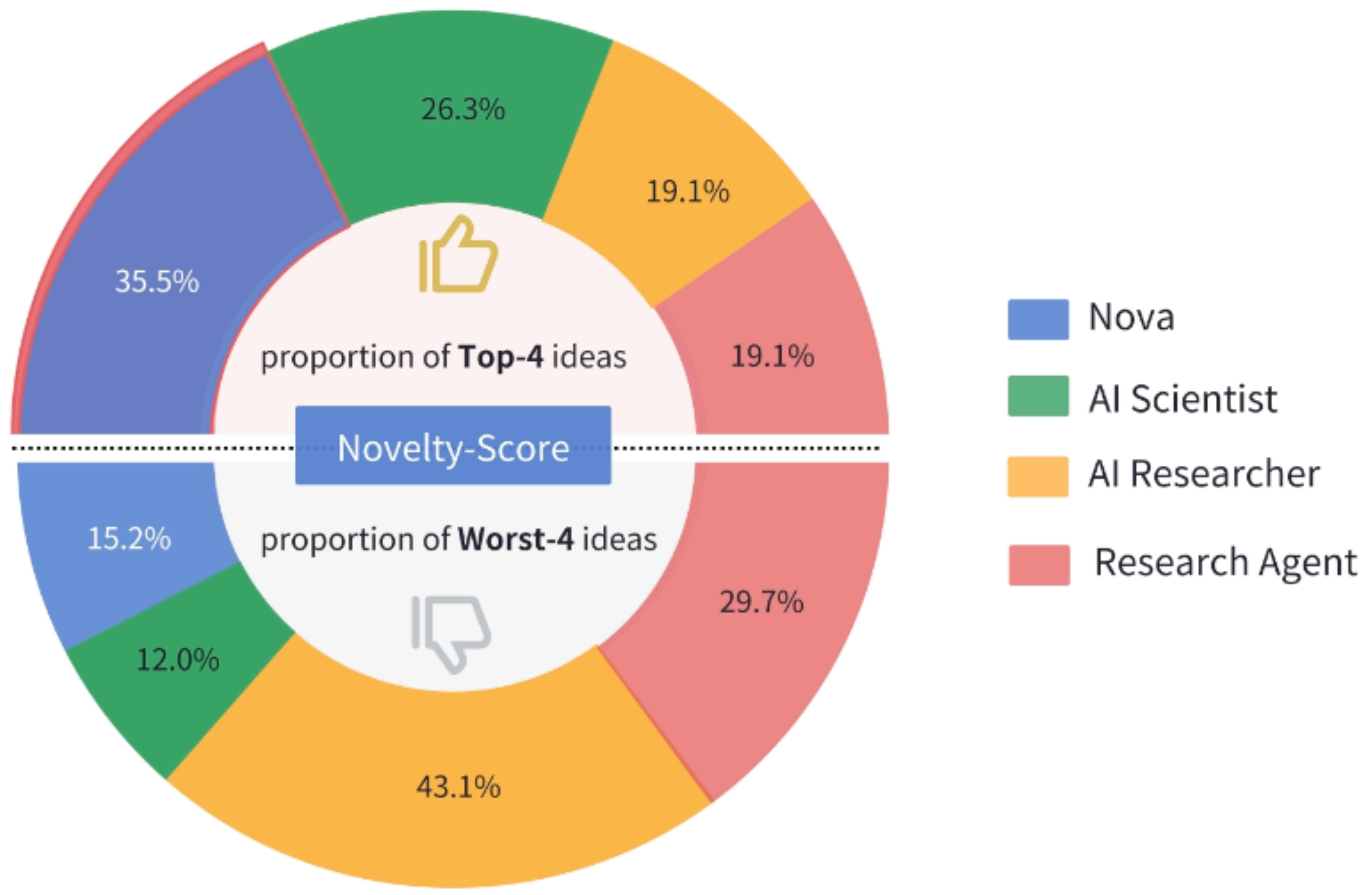


Evaluation

➡ Automatic Evaluation

➡ Human Evaluation



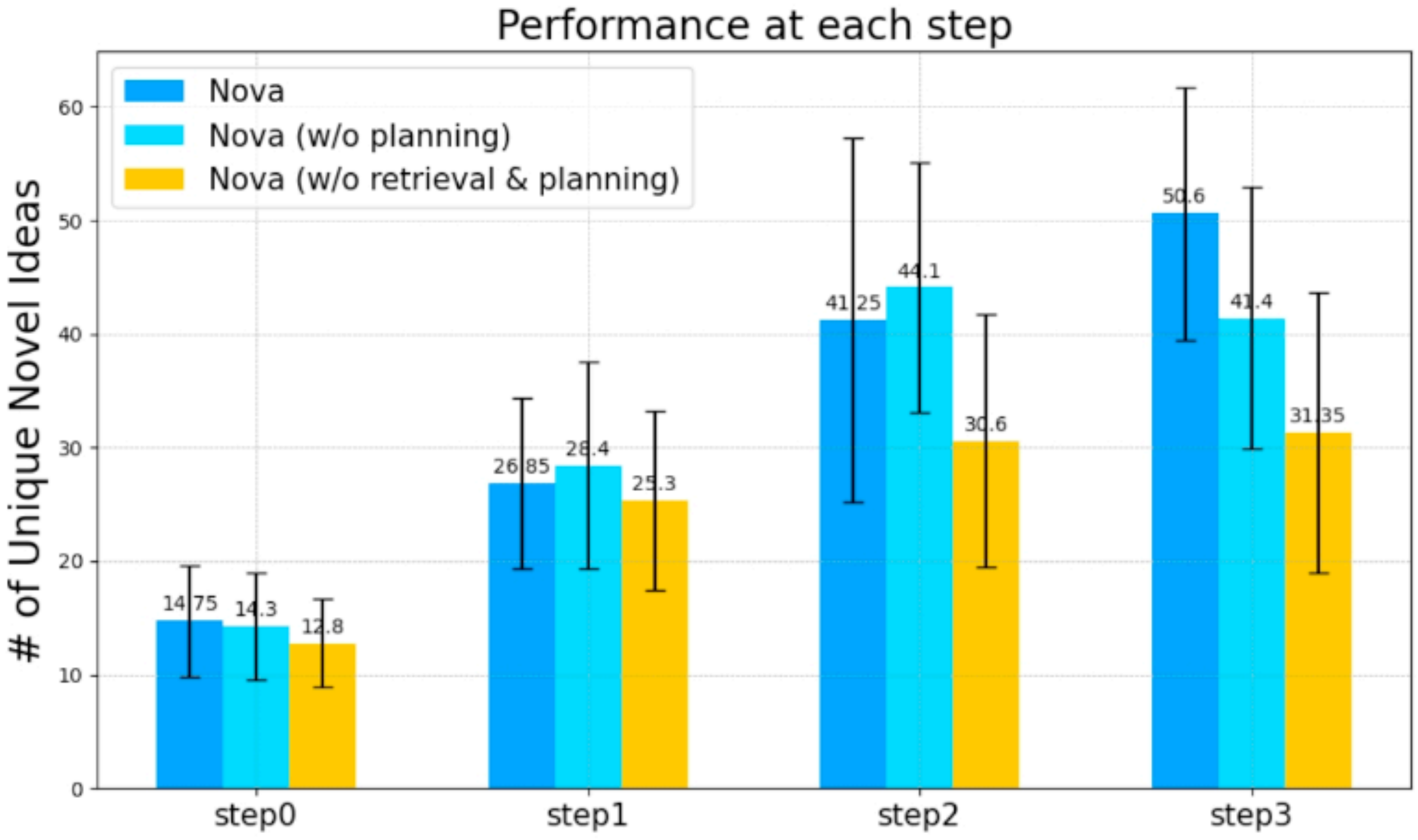


Evaluation

➡ Automatic Evaluation

➡ Human Evaluation

➡ Ablation Study



Takeaways

➡ Are these ideas truly novel and useful?

- Yes—carefully designed agent systems can be effective.
- Studies show LLMs can generate ideas rated **more novel** than human expert ideas.
- However, they often **lack feasibility**, detail, or realism.

➡ Can LLMs outperform human experts at ideation?

- In some settings, but humans still **excel in grounding ideas** with practical knowledge and detailed execution.

➡ How do we evaluate AI-generated ideas at scale?

- Methods like Swiss System Tournament.
- Blind human reviews.

Questions?

Thank You!