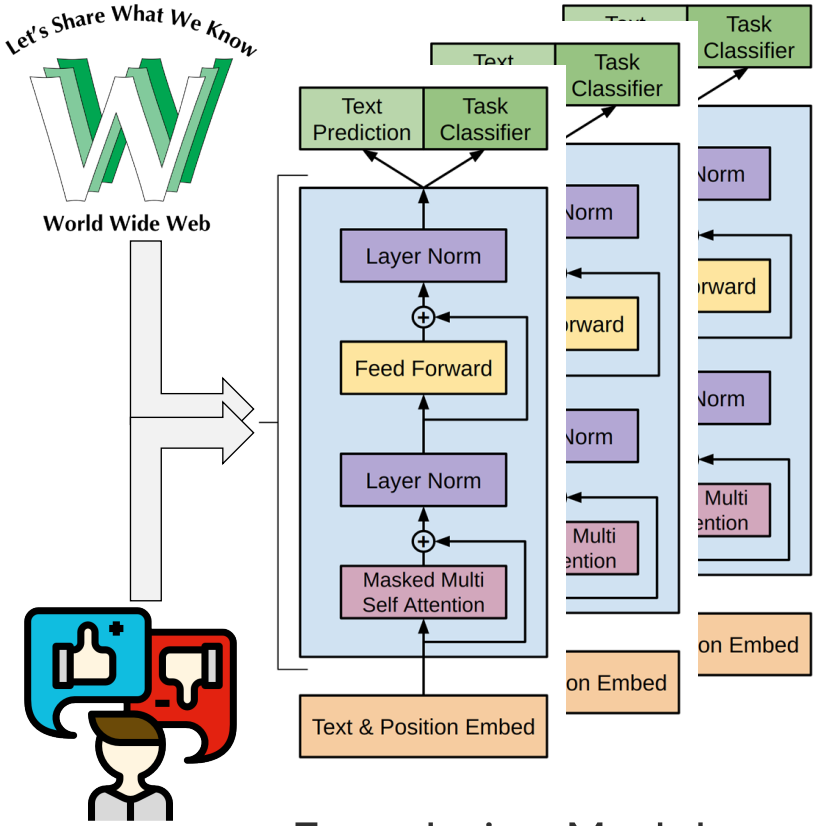
KAIST

Beyond Parameters:
Contextualizing Foundation Models with Universal Retrieval and Verification

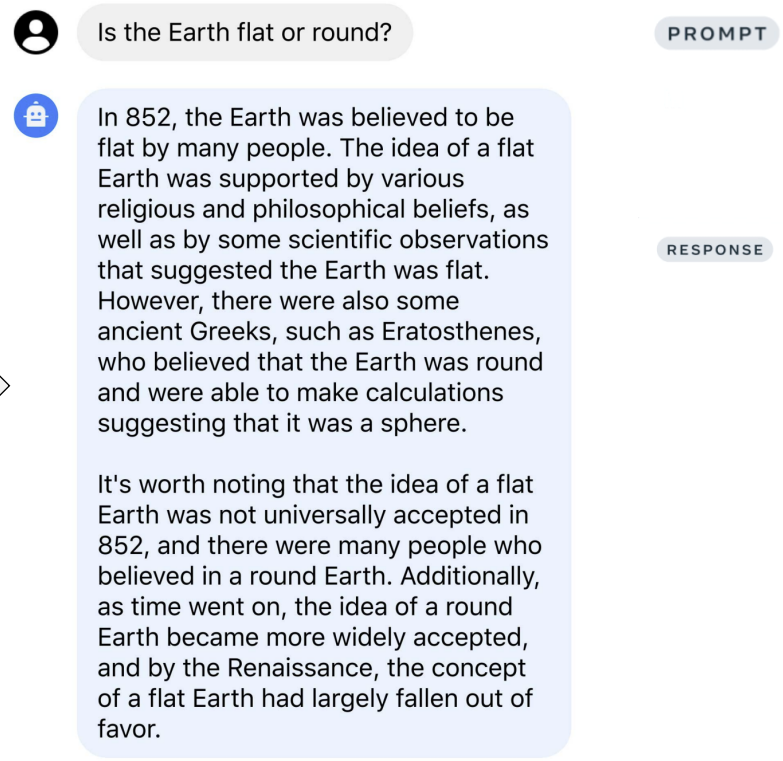
Jinheon Baek (jinheon.baek@kaist.ac.kr)
Ph.D. student in MLAI Lab at KAIST AI

Foundation Models

Foundation models, trained virtually on the entire web and further tuned to align with human values, have shown impressive capabilities across diverse tasks and domains.



Foundation Models

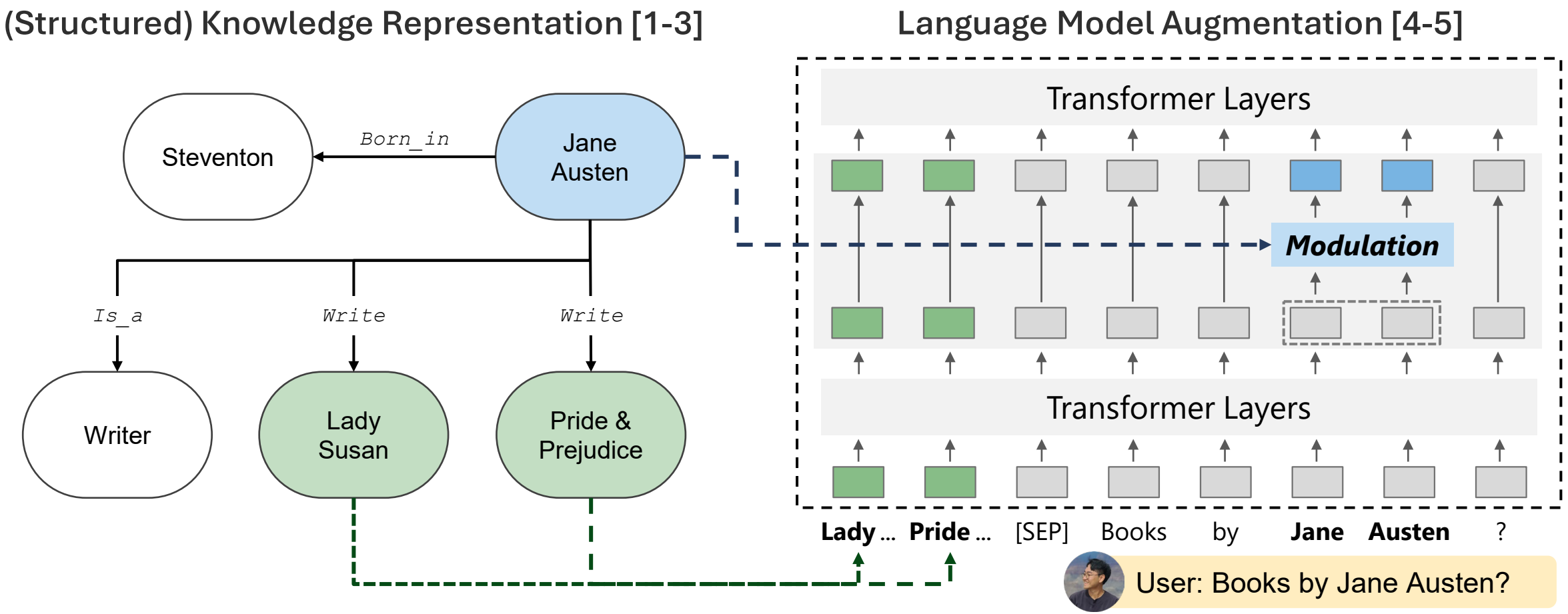


Chat Response

Code Writing

Prior Knowledge-Augmented Language Models

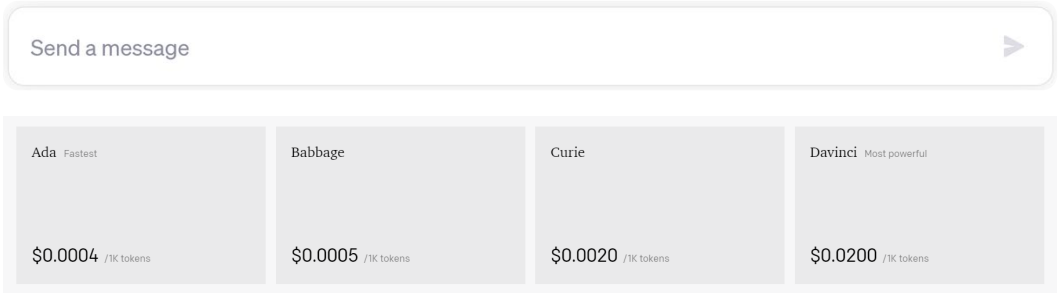
As knowledge internalized in models is inaccurate and insufficient, I proposed model augmentation and knowledge representation (for retrieval) methods in my master's.



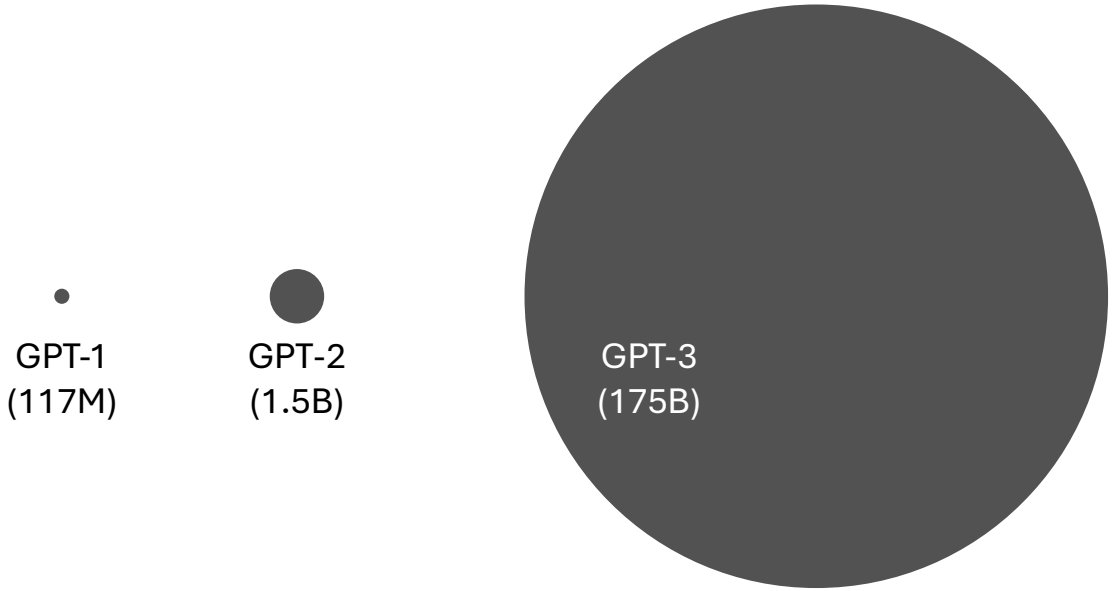
[1] Baek et al., NeurIPS 2020; [2] Baek*, Kang* et al., ICLR 2021; [3] Jo*, Baek* et al., NeurIPS 2022; [4] Kang*, Baek* et al., NAACL 2022; [5] Kang*, Kwak*, Baek* et al., KRLM @ ICML 2022

Motivation for Beyond Parameters

When starting PhD, frontier models (GPT-3 Davinci and ChatGPT) become released as services, with parameters closed and training very expensive even when available.



Unavailable



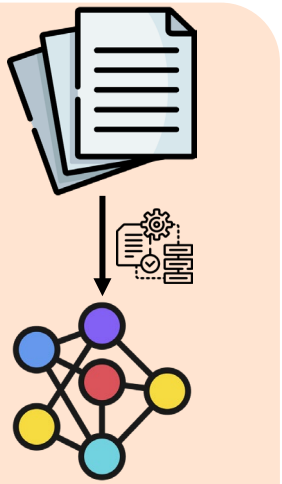
Too costly to train:

- Training: 48 x H100 (80GB)
- Inference: 8 x H100 (80GB)

Contextualizing Foundation Models with Knowledge

My thesis consists of three parts: retrieval, contextualization, and application.

Part 1. Advancing Foundation Models with Contextualization



Contextualizing Foundation Models with Knowledge

My thesis consists of three parts: retrieval, contextualization, and application.

Part 1. Advancing Foundation Models with Contextualization

Part 2. Representing and Retrieving Knowledge for Contextualization

Contextualizing Foundation Models with Knowledge

My thesis consists of three parts: retrieval, contextualization, and application.

Part 1. Advancing Foundation Models with Contextualization

Part 2. Representing and Retrieving Knowledge for Contextualization

Part 3. Expanding Contextualization to Real-World Applications

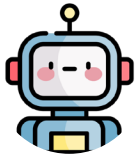
Contextualizing Models without Parameter Updates

I first investigate whether larger and more capable foundation models, without training, can leverage (structured) knowledge when incorporated into their context.

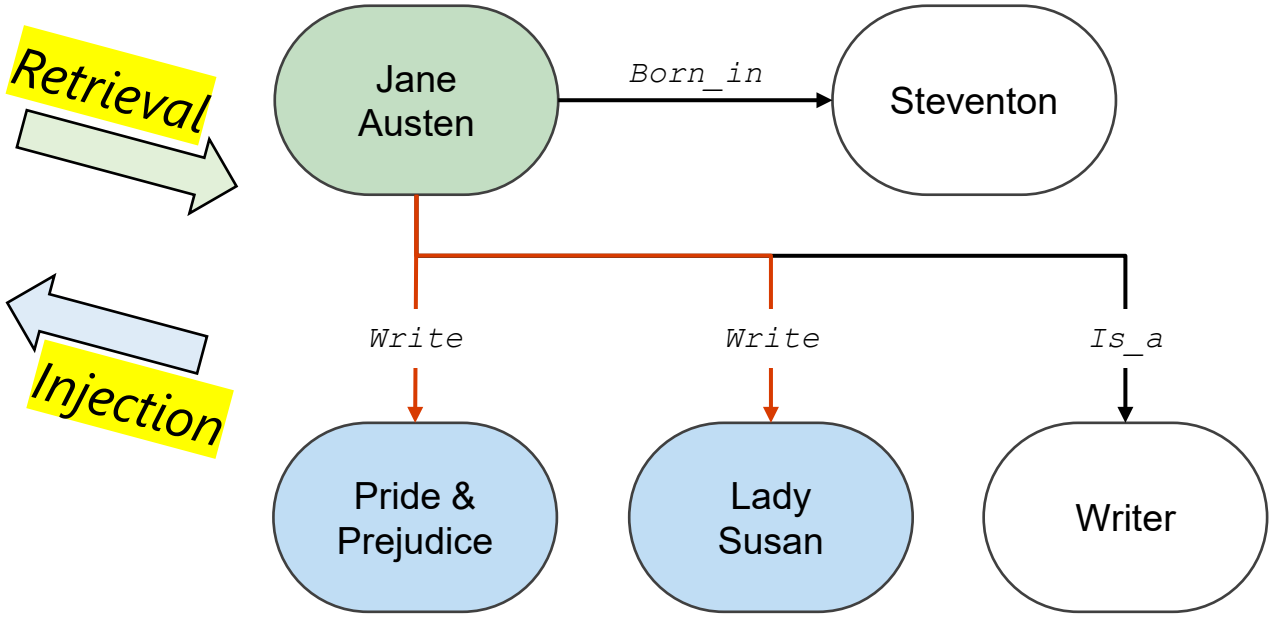
Jinheon: Could you recommend any books written by Jane Austen?

Here are some facts that might be relevant to answering the question:
Jane Austen wrote Lady Susan
Jane Austen wrote Pride & Prejudice

Question: Could you recommend any …
Answer:



Sure! Her works include Lady Susan and Pride & Prejudice.



Baek et al., Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering, NLRSE @ ACL 2023 (Best Paper Award, > 300 Citations)

Contextualizing Models without Parameter Updates

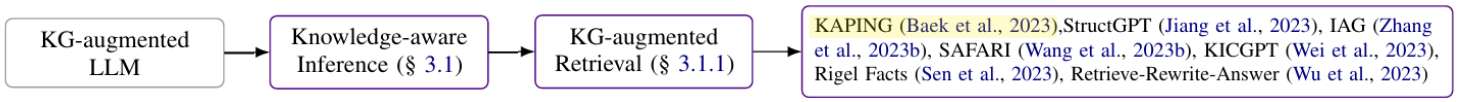
The experimental results show that the proposed Knowledge-Augmented language model PromptING (KAPING) method outperforms baselines on knowledge graph QA.

Datasets	Methods	T5 (0.8B)	T5 (3B)	T5 (11B)	OPT (2.7B)	OPT (6.7B)	OPT (13B)	T0 (3B)	T0 (11B)	GPT-3 (6.7B)	GPT-3 (175B)	AlexaTM (20B)	Average
WebQSP w/ Freebase	No Knowledge	6.95	13.40	9.48	19.85	29.77	28.38	21.43	40.77	44.63	63.59	46.79	29.55
	Random Knowledge	21.55	19.15	17.57	28.07	31.73	33.31	32.62	51.20	51.01	65.87	57.37	37.22
	Popular Knowledge	15.30	16.88	18.39	28.32	28.13	24.21	27.05	47.22	45.58	62.26	54.91	33.48
	Generated Knowledge	6.19	7.84	6.76	7.46	11.50	8.22	19.41	38.81	45.89	62.14	35.13	22.67
	KAPING (Ours)	**34.70**	**25.41**	**24.91**	**41.09**	**43.93**	**40.20**	**52.28**	**62.85**	**60.37**	**73.89**	**67.67**	**47.94**
WebQSP w/ Wikidata	No Knowledge	10.30	18.42	15.21	23.94	33.77	32.40	24.56	44.20	48.50	67.60	42.41	32.85
	Random Knowledge	17.94	22.78	24.28	37.24	35.61	38.27	28.85	47.68	52.05	60.64	55.63	38.27
	Popular Knowledge	15.35	20.80	20.74	30.83	30.01	27.83	24.83	48.02	47.41	63.37	53.92	34.83
	Generated Knowledge	11.94	13.30	12.28	11.26	17.53	14.19	22.92	41.34	48.77	65.89	31.16	26.42
	KAPING (Ours)	**23.67**	**40.38**	**35.47**	**49.52**	**53.34**	**51.57**	**49.86**	**58.73**	**60.44**	**69.58**	**65.04**	**50.69**
Mintaka w/ Wikidata	No Knowledge	11.23	14.25	17.06	19.76	27.19	26.83	14.75	23.74	34.65	56.33	41.97	26.16
	Random Knowledge	17.59	18.19	18.83	28.11	26.58	28.36	16.10	26.15	32.98	51.56	46.02	28.22
	Popular Knowledge	17.56	18.09	18.73	26.97	27.08	23.10	16.74	27.15	32.48	53.16	46.41	27.95
	Generated Knowledge	13.61	14.61	14.29	11.87	14.96	16.24	14.46	23.13	33.12	55.65	34.58	22.41
	KAPING (Ours)	**19.72**	**22.00**	**22.85**	**32.94**	**32.37**	**33.37**	**20.68**	**29.50**	**35.61**	**56.86**	**49.08**	**32.27**

Baek et al., Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering, NLRSE @ ACL 2023 (Best Paper Award, > 300 Citations)

Subsequent Works in Model Contextualization

New Paradigm for Model Contextualization

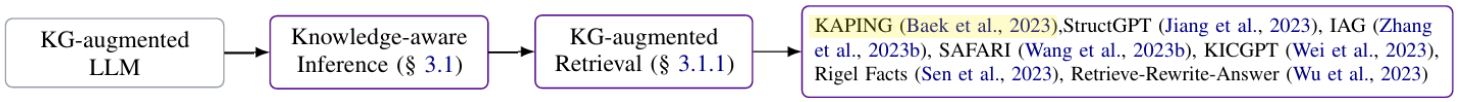


Agrawal et al., Can Knowledge Graphs Reduce Hallucinations in LLMs: A Survey, NAACL 2024

GNN-RAG	(Mavromatis et al., 2024)
G-Retriever	(He et al., 2024)
Give us the Facts	(Yang et al., 2024)
GraphRAG	(Han et al., 2025)

Subsequent Works in Model Contextualization

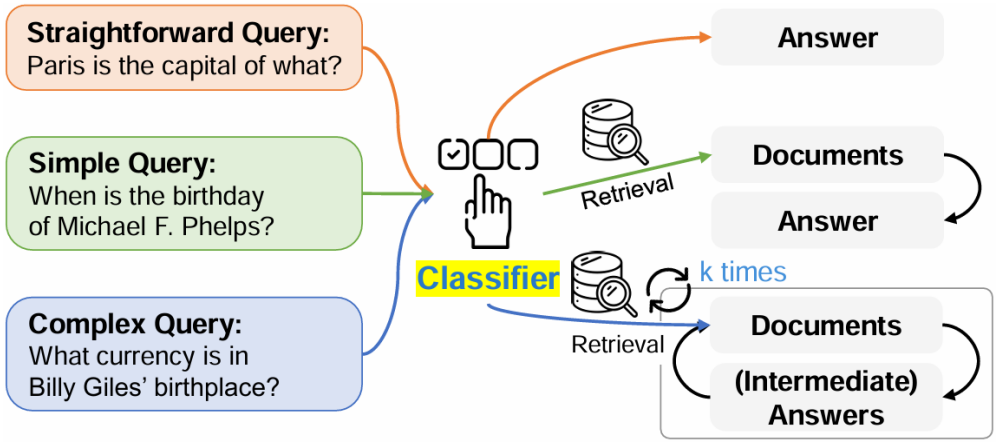
New Paradigm for Model Contextualization



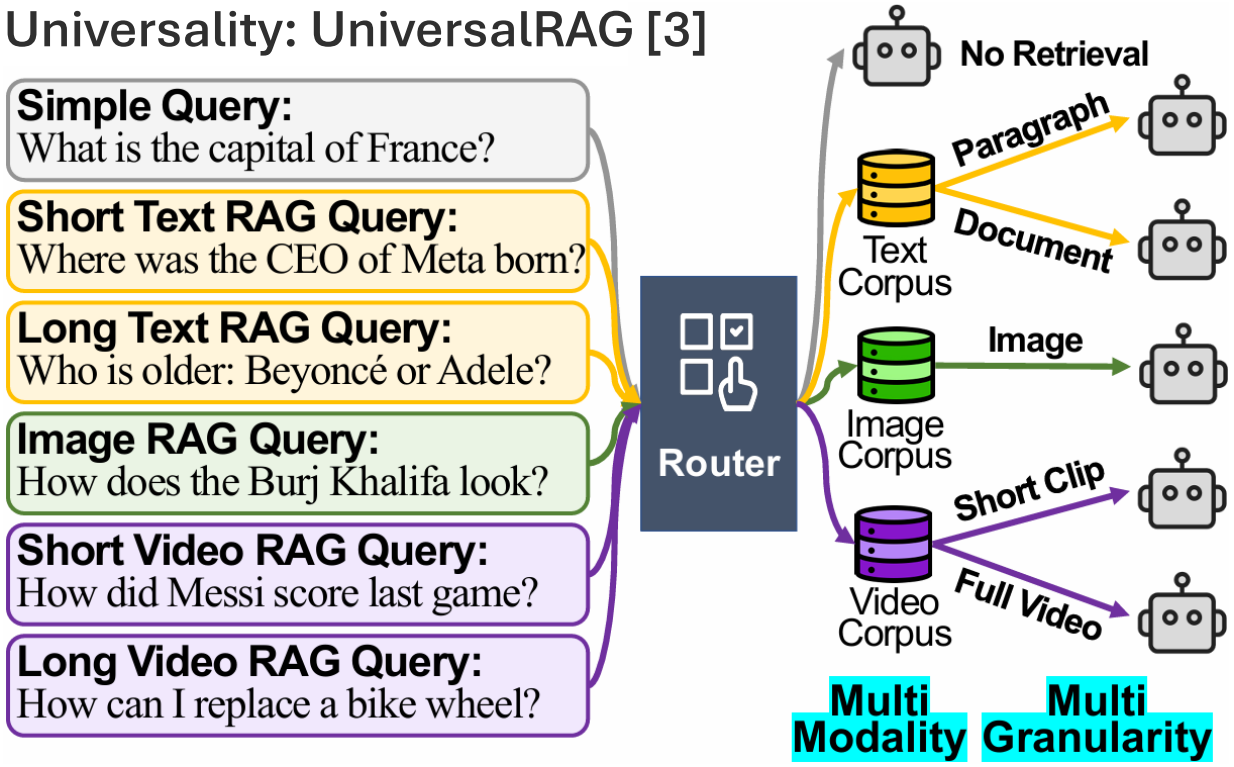
Agrawal et al., Can Knowledge Graphs Reduce Hallucinations in LLMs: A Survey, NAACL 2024

GNN-RAG	(Mavromatis et al., 2024)
G-Retriever	(He et al., 2024)
Give us the Facts	(Yang et al., 2024)
GraphRAG	(Han et al., 2025)

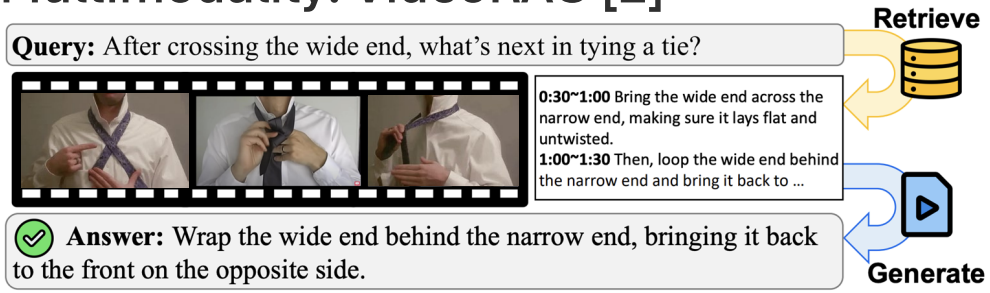
Adaptivity: Adaptive-RAG [1]



Universality: UniversalRAG [3]



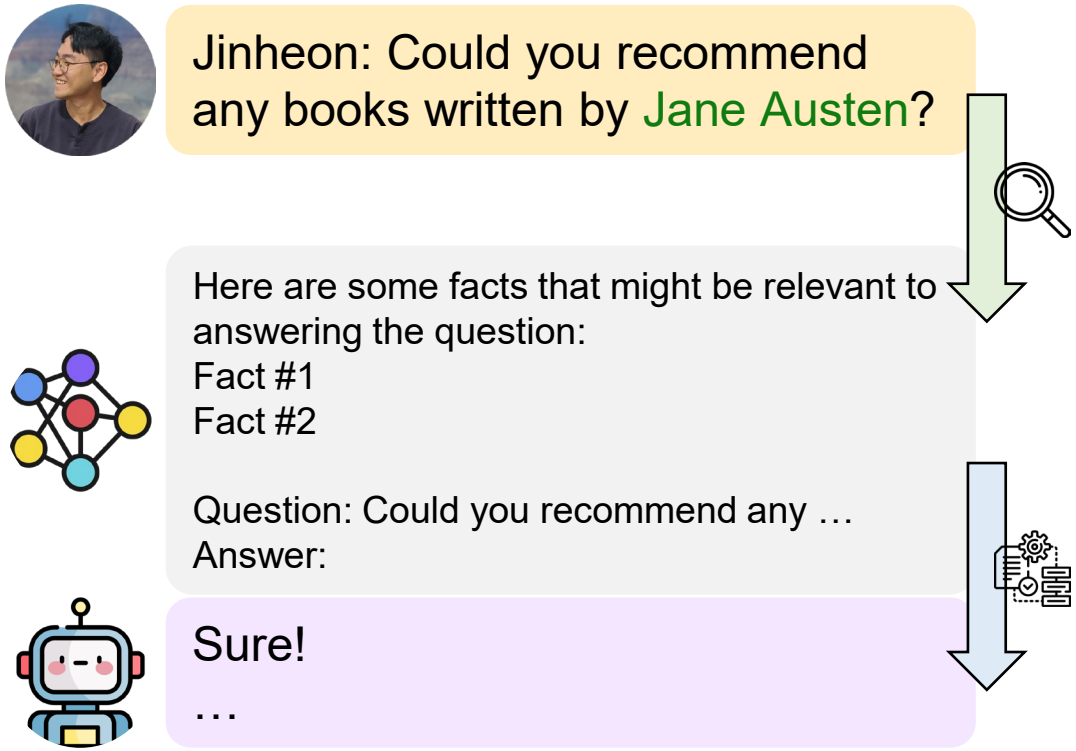
Multimodality: VideoRAG [2]



[1] Jeong, Baek et al., NAACL 2024 (Oral Presentation, #7 Most Influential Paper); [2] Jeong*, Kim*, Baek* et al., ACL Findings 2025; [3] Yeo*, Kim*, Jeong, Baek et al., arXiv 2025

What Breaks in Model Contextualization?

The contextualization approaches (that augment models with the knowledge retrieved from the data store) still fall short of guaranteeing fully accurate answers.

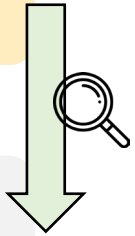


Baek et al., Knowledge-Augmented Language Model Verification, EMNLP 2023

What Breaks in Model Contextualization?

The contextualization approaches (that augment models with the knowledge retrieved from the data store) still fall short of guaranteeing fully accurate answers.

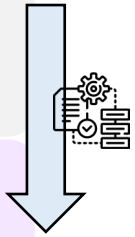
Jinheon: Could you recommend any books written by Jane Austen?

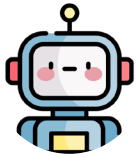


Retrieval fails

Here are some facts that might be relevant to answering the question:
Jane Austen was a Novelist
Jane Austen was born in Steventon

Question: Could you recommend any …
Answer:





Sure! Jane Austen, born in Steventon, was a novelist.

Baek et al., Knowledge-Augmented Language Model Verification, EMNLP 2023

What Breaks in Model Contextualization?

The contextualization approaches (that augment models with the knowledge retrieved from the data store) still fall short of guaranteeing fully accurate answers.

Jinheon: Could you recommend any books written by Jane Austen?

Here are some facts that might be relevant to answering the question:
Jane Austen wrote Lady Susan
Jane Austen wrote Pride & Prejudice

Question: Could you recommend any …
Answer:

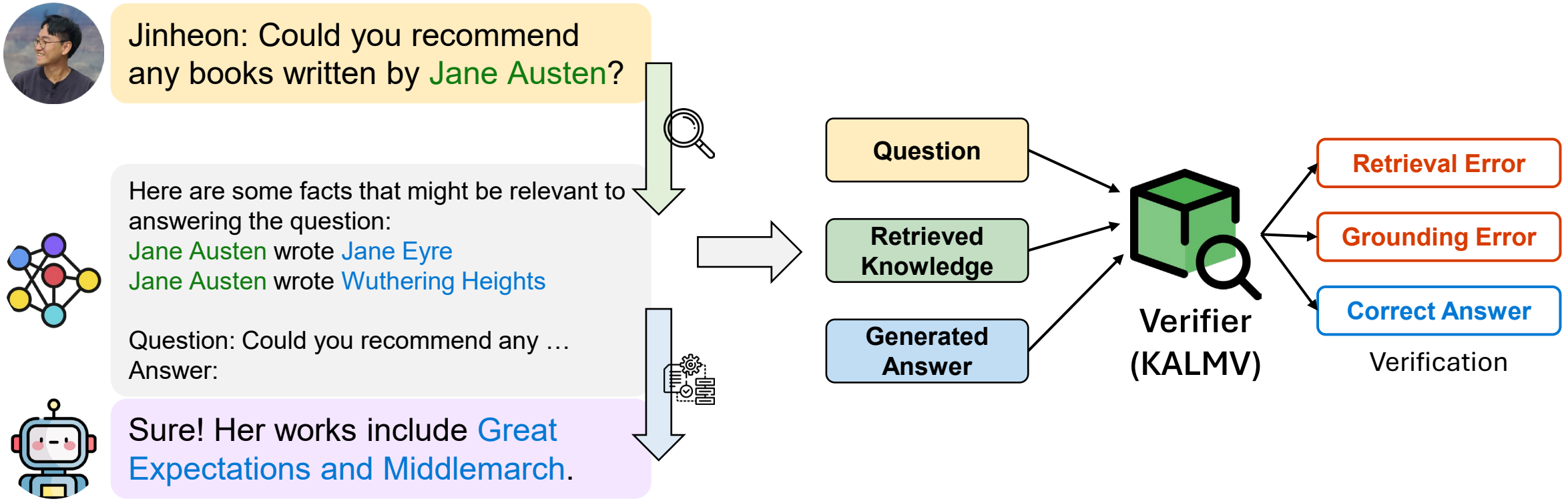
Sure! Her works include Great Expectations and Middlemarch.

Grounding fails

Baek et al., Knowledge-Augmented Language Model Verification, EMNLP 2023

Verification of Retrieval and Grounding Errors

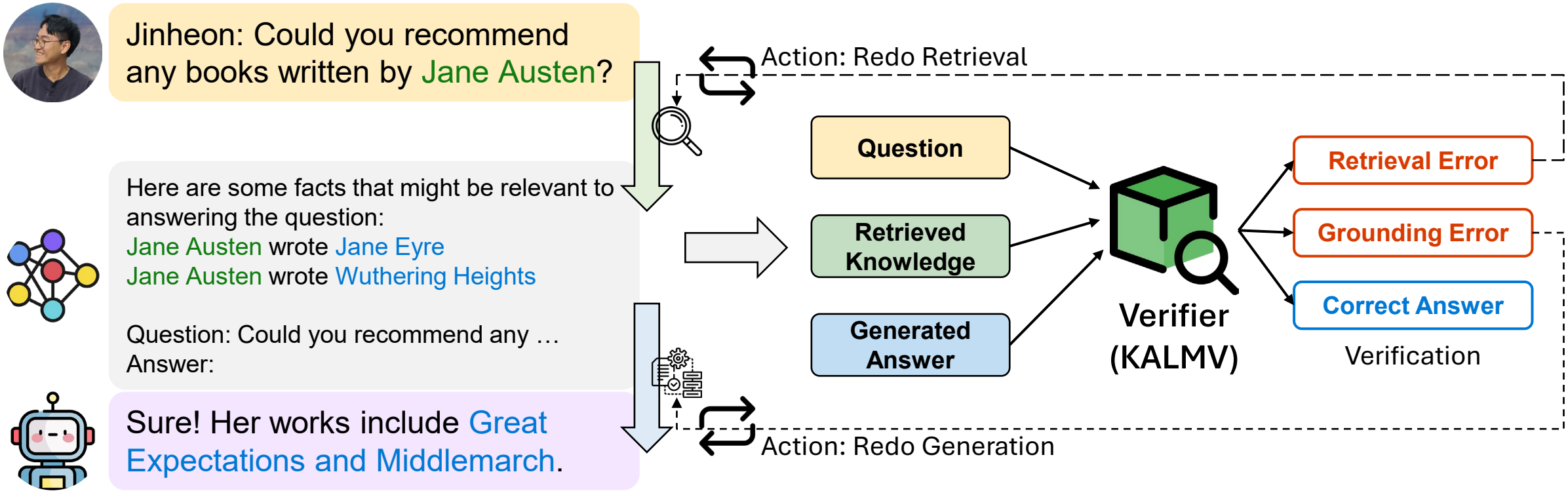
I propose a Knowledge-Augmented Language Model Verification (KALMV) method, which detects retrieval and grounding errors.



Baek et al., Knowledge-Augmented Language Model Verification, EMNLP 2023

Verification of Retrieval and Grounding Errors

I propose a Knowledge-Augmented Language Model Verification (KALMV) method, which detects retrieval and grounding errors but also further rectifies them, iteratively.



Baek et al., Knowledge-Augmented Language Model Verification, EMNLP 2023

Verification & Refinement Greatly Reduce Errors

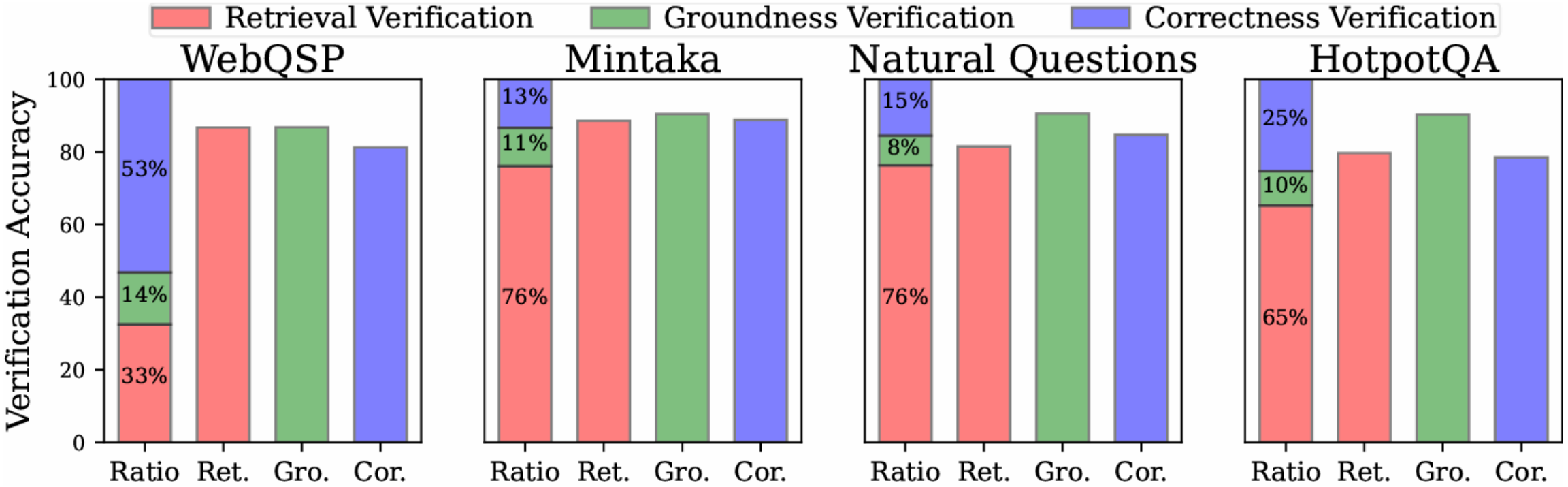
We validate KALMV on question answering tasks with both the structured and unstructured knowledge sources, demonstrating its effectiveness over baselines.

Datasets	Methods	Base (250M)			Large (780M)			XL (3B)		
		F1	EM	Acc	F1	EM	Acc	F1	EM	Acc
Natural Questions w/ Wikipedia	Naive Language Models	7.53	3.24	4.57	11.09	6.29	7.81	16.89	11.16	12.94
	Knowledge-Augmented LMs	18.06	12.30	15.26	18.61	13.74	16.40	19.03	14.13	16.90
	Adaptive Retrieval w/ Confidence	16.70	11.02	14.07	18.16	13.07	15.60	20.89	15.76	18.28
	LLM-Augmenter w/ Knowledge F1	19.58	13.56	16.81	28.53	21.22	25.32	31.00	23.06	27.59
	LLM-Augmenter w/ Confidence	19.91	14.14	17.19	20.19	14.97	18.29	22.88	17.17	20.49
	KALMV (Ours)	**52.98**	**42.36**	**50.43**	**56.80**	**46.13**	**53.57**	**67.43**	**58.06**	**63.17**
HotpotQA w/ Wikipedia	Naive Language Models	14.25	9.68	10.36	16.80	11.78	12.41	21.97	15.06	16.22
	Knowledge-Augmented LMs	31.20	22.77	25.13	33.46	25.29	27.37	35.47	27.08	29.14
	Adaptive Retrieval w/ Confidence	26.82	19.10	21.11	26.80	19.65	21.23	29.41	21.55	23.54
	LLM-Augmenter w/ Knowledge F1	32.89	23.24	26.12	39.40	28.55	31.60	46.97	34.54	37.72
	LLM-Augmenter w/ Confidence	34.75	25.67	28.20	35.78	27.29	29.38	40.57	31.35	33.71
	KALMV (Ours)	**64.06**	**52.31**	**55.84**	**63.74**	**52.39**	**55.98**	**67.21**	**54.99**	**58.07**
WebQSP w/ Wikidata	Naive Language Models	32.53	21.35	25.78	40.33	30.08	32.74	46.20	36.43	40.11
	Knowledge-Augmented LMs	53.57	43.25	53.68	42.37	26.13	62.28	49.45	36.02	59.28
	Adaptive Retrieval w/ Entity	49.13	37.79	46.32	47.81	35.68	49.32	51.99	41.54	51.16
	Adaptive Retrieval w/ Confidence	46.76	36.49	43.66	48.32	36.56	51.98	53.17	43.32	53.89
	LLM-Augmenter w/ Knowledge F1	56.42	45.95	56.26	44.41	27.79	64.56	51.95	38.12	61.96
	LLM-Augmenter w/ Confidence	56.62	47.33	56.36	44.35	28.79	64.47	50.63	36.62	60.67
	KALMV (Ours)	**74.31**	**63.92**	**77.78**	**54.79**	**45.46**	**82.71**	**67.10**	**50.81**	**83.21**

Baek et al., Knowledge-Augmented Language Model Verification, EMNLP 2023

Verification Is Easier than Generation

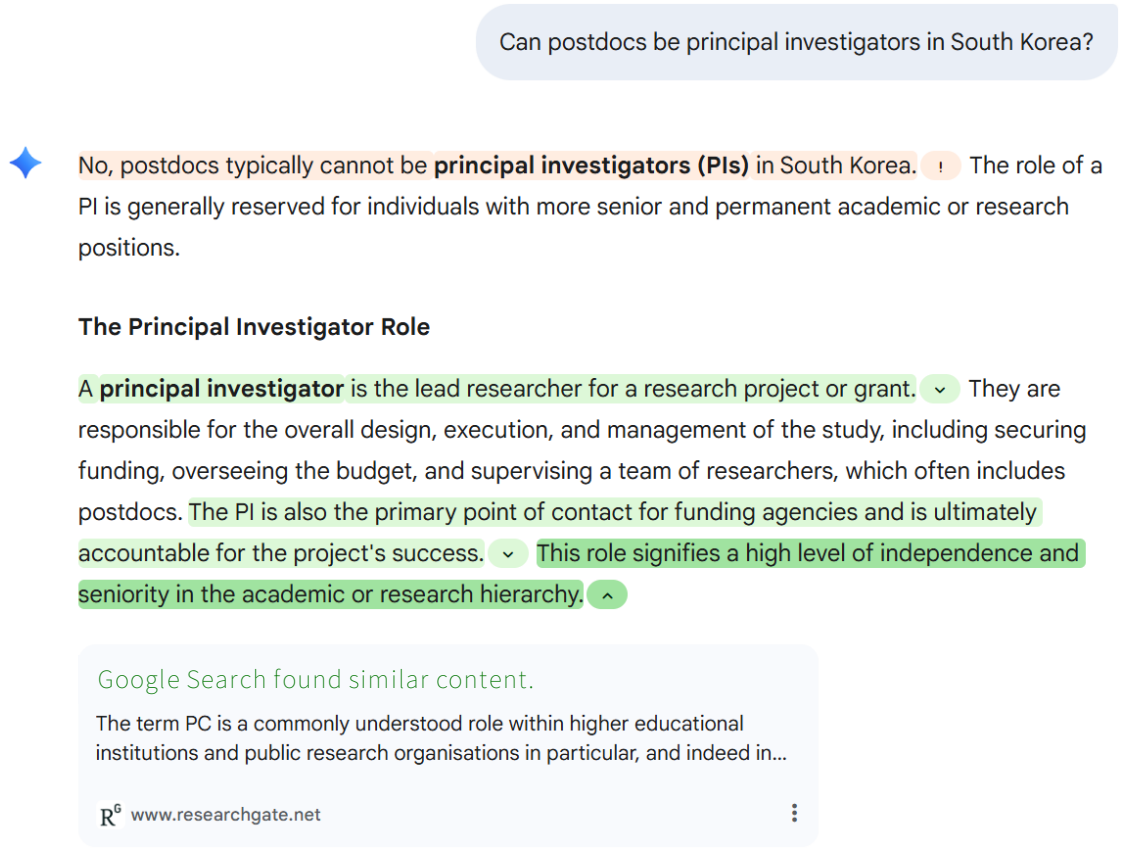
The first bar in each subplot shows the distribution of error types, while the remaining bars report the verifier accuracy, and we observe that accuracy reaches around 80%.

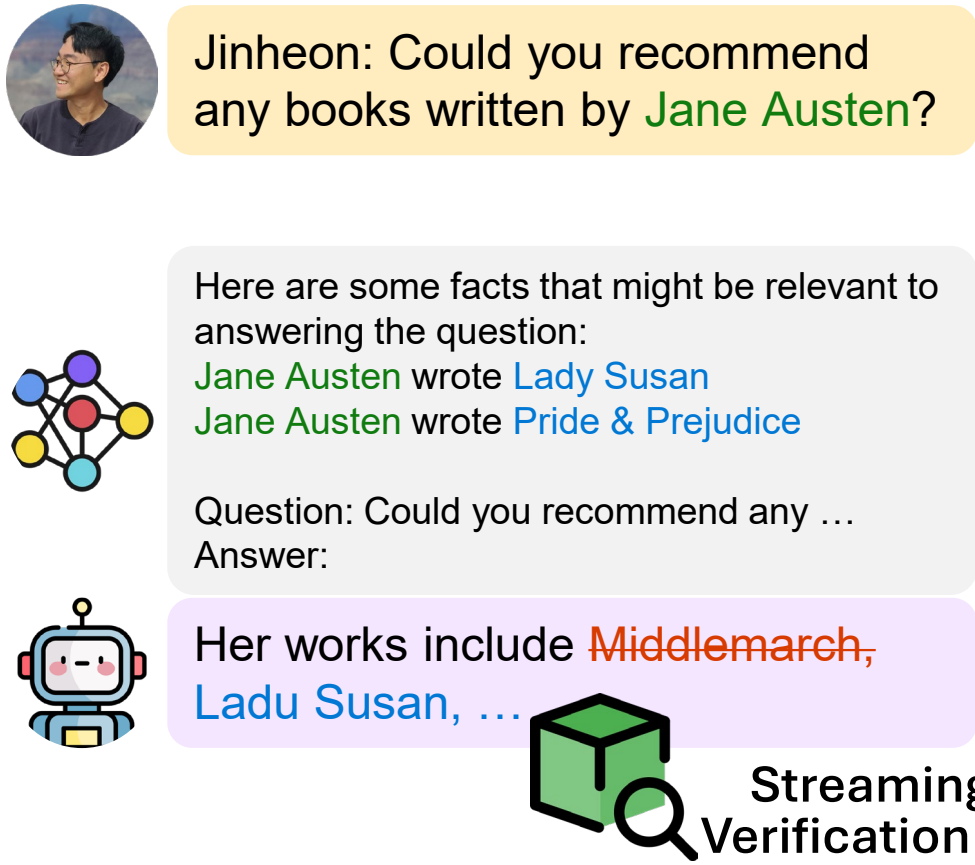


Baek et al., Knowledge-Augmented Language Model Verification, EMNLP 2023

Verification of (Contextualized) Models in the Wild

A similar idea of verifying the generated response from models is deployed in Gemini. Also, I make the verification and refinement process operate in a streaming manner.





Ko, Baek et al., Real-time Verification and Refinement of Language Model Text Generation, EMNLP 2025

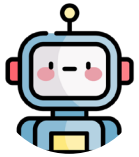
What Powers Contextualization: Relevant Knowledge

Effective contextualization requires relevant knowledge.

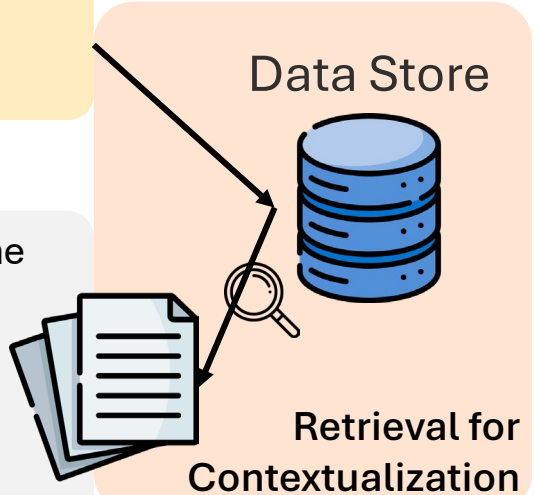
Recent approaches to reduce transformer inference latency?

Answer the user question by leveraging the retrieved knowledge from recent papers:
1. FlashAttention (…)
2. SpecInfer (…)
3. PagedAttention (…)

Question: Recent approaches to reduce …



Recent methods to reduce latency include FlashAttention for …



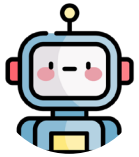
What Powers Contextualization: Relevant Knowledge

Effective contextualization requires relevant knowledge.
However, retrieval is the most challenging part in model contextualization.

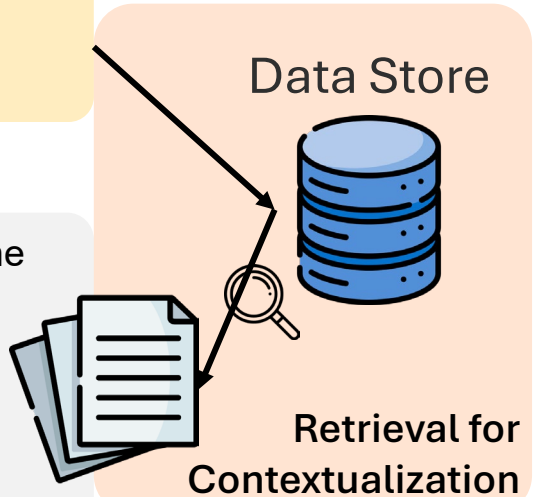
Recent approaches to reduce transformer inference latency?

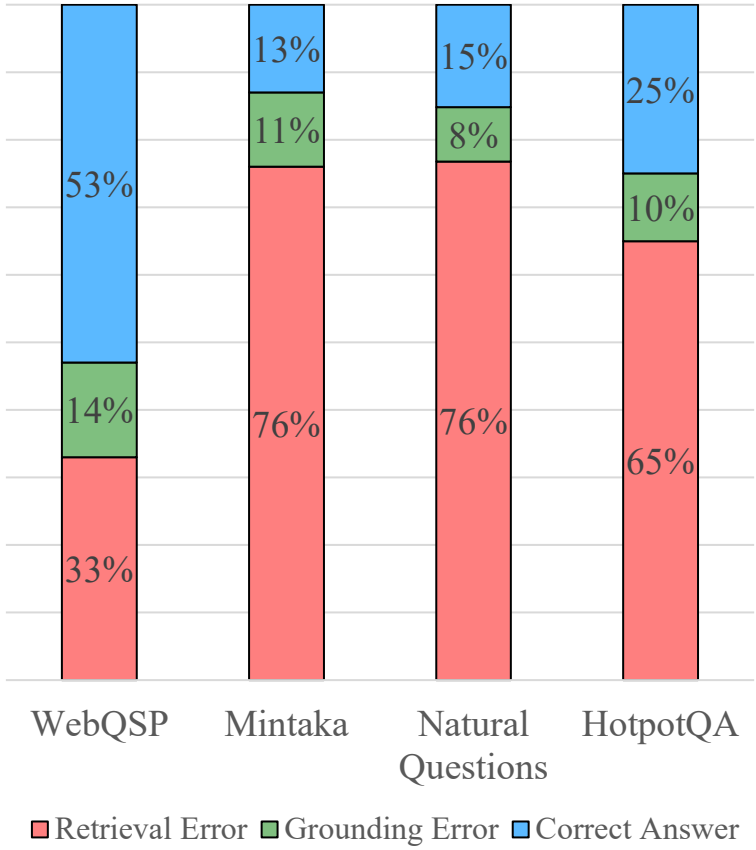
Answer the user question by leveraging the retrieved knowledge from recent papers:
1. FlashAttention (…)
2. SpecInfer (…)
3. PagedAttention (…)

Question: Recent approaches to reduce …



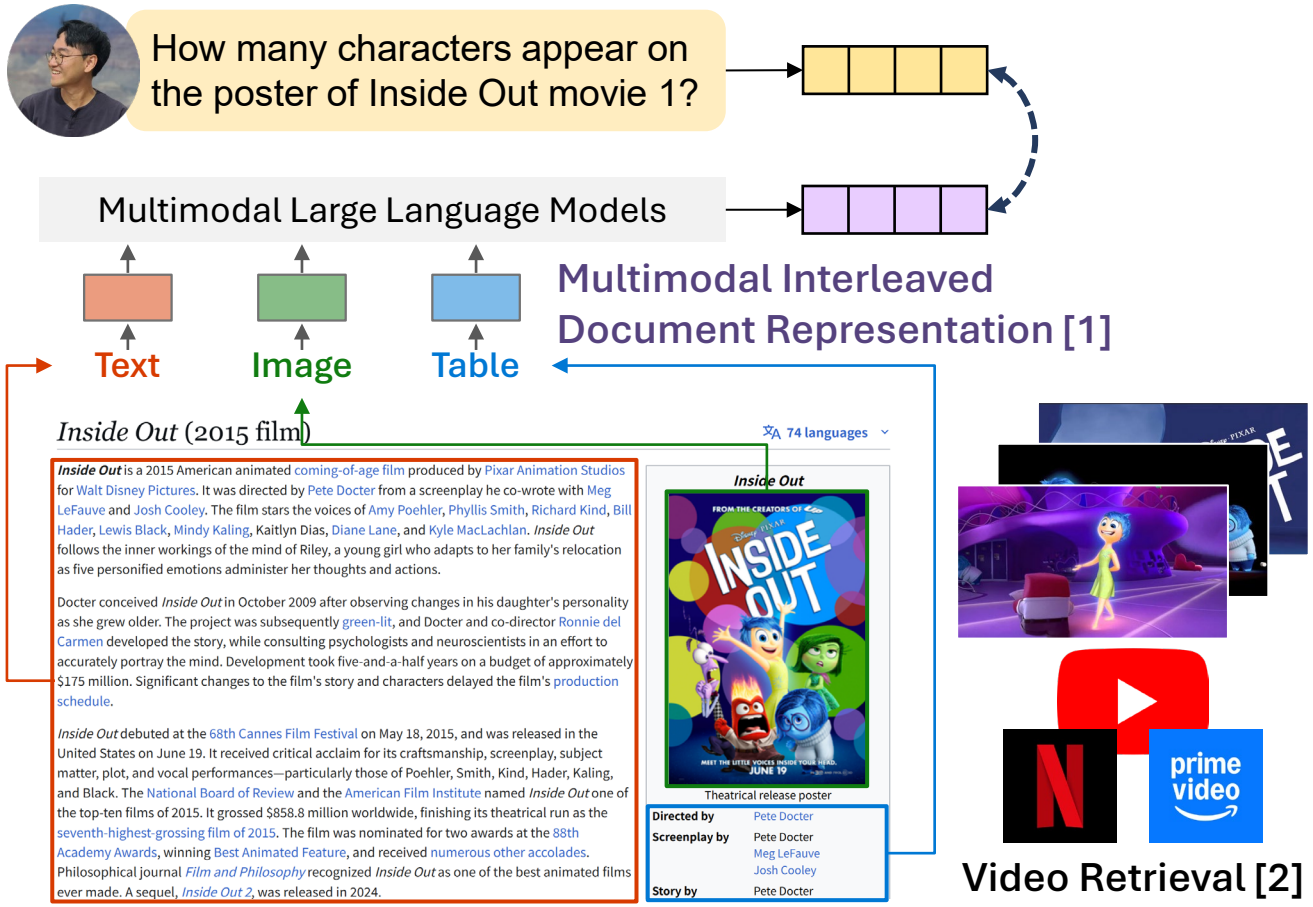
Recent methods to reduce latency include FlashAttention for …





Retrieving Richer Knowledge Beyond Textual Corpora

To advance retrieval, I have gone beyond textual corpora and proposed approaches for retrieving richer knowledge sources (such as multimodal documents and videos).



[1] Lee*, Ko*, Baek* et al., VecDB @ ICML 2025; [2] Jeong*, Kim*, Baek* et al., ACL Findings 2025

Different Interfaces for Different Knowledge Sources

Beyond unstructured corpora (such as text), another knowledge source, knowledge graphs organized through schemas of nodes and edges, requires a different retriever.

How many characters appear on the poster of Inside Out movie 1?

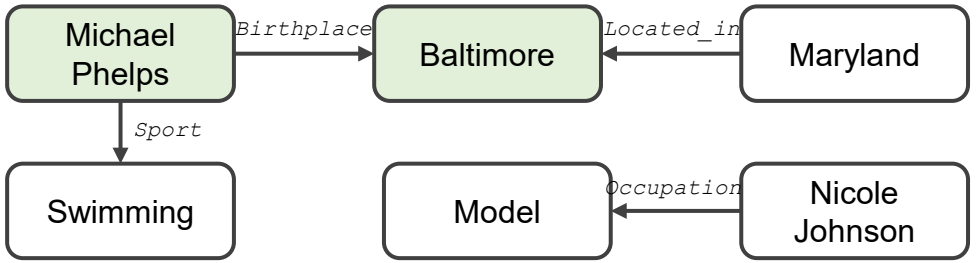
Multimodal Large Language Models

Text Image Table

Multimodal Interleaved Document Representation [1]

Video Retrieval [2]

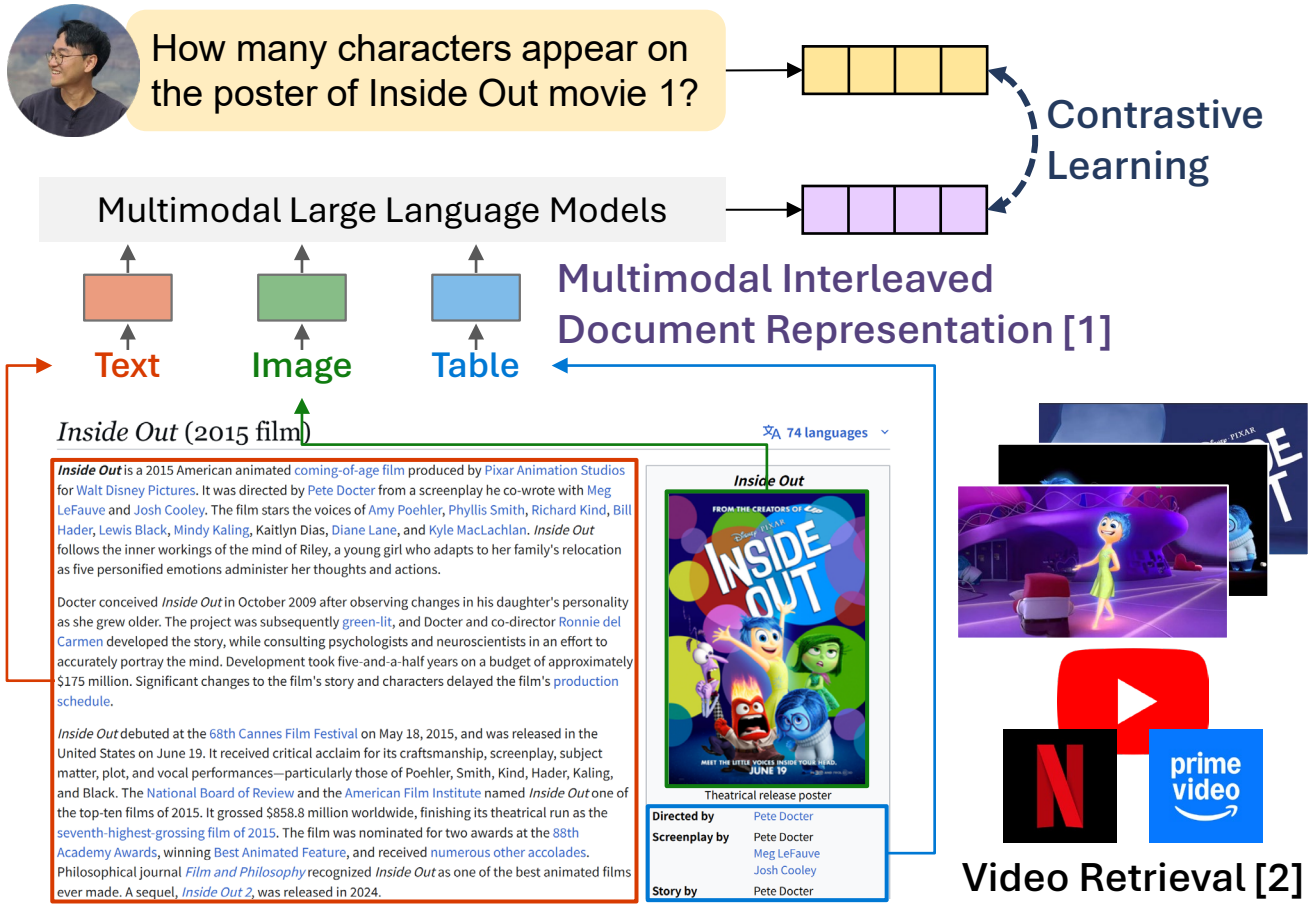
Structured Knowledge Graph

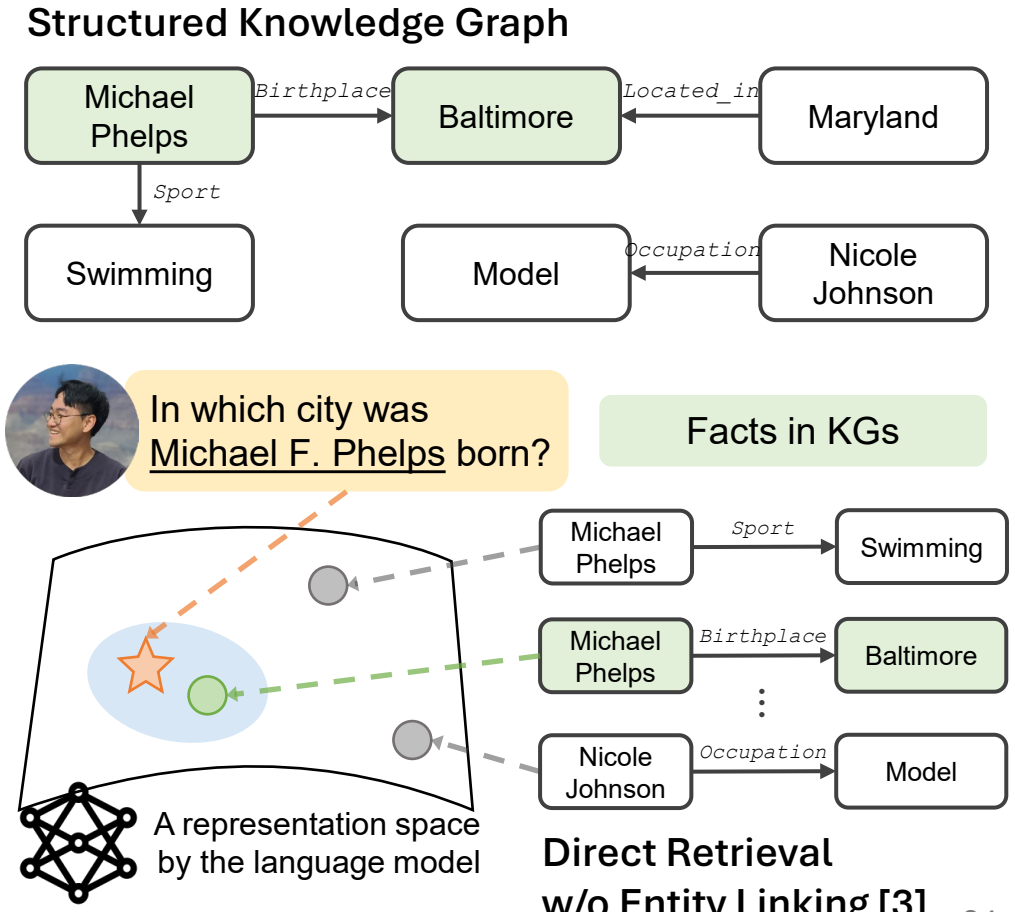


[1] Lee*, Ko*, Baek* et al., VecDB @ ICML 2025; [2] Jeong*, Kim*, Baek* et al., ACL Findings 2025; [3] Baek et al., ACL 2023

Different Interfaces for Different Knowledge Sources

To make structured knowledge graphs retrievable as unstructured sources, I propose linearizing facts from structured knowledge graphs and encoding them into vectors.





[1] Lee*, Ko*, Baek* et al., VecDB @ ICML 2025; [2] Jeong*, Kim*, Baek* et al., ACL Findings 2025; [3] Baek et al., ACL 2023

Different Interfaces for Different Knowledge Sources

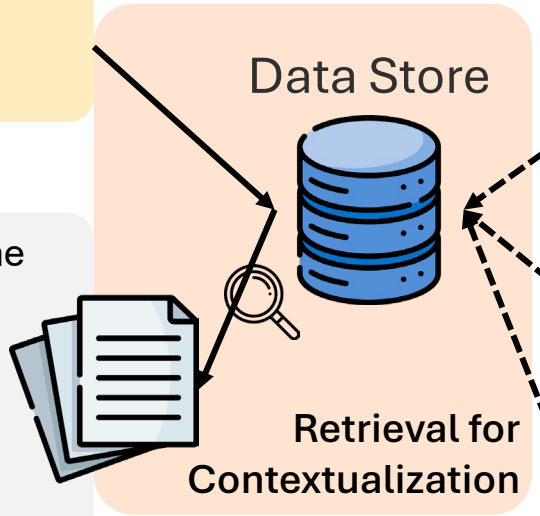
While my approaches handle unstructured corpora and structured graphs, real-world knowledge comes in many more forms, each with its own distinct retrieval strategies.

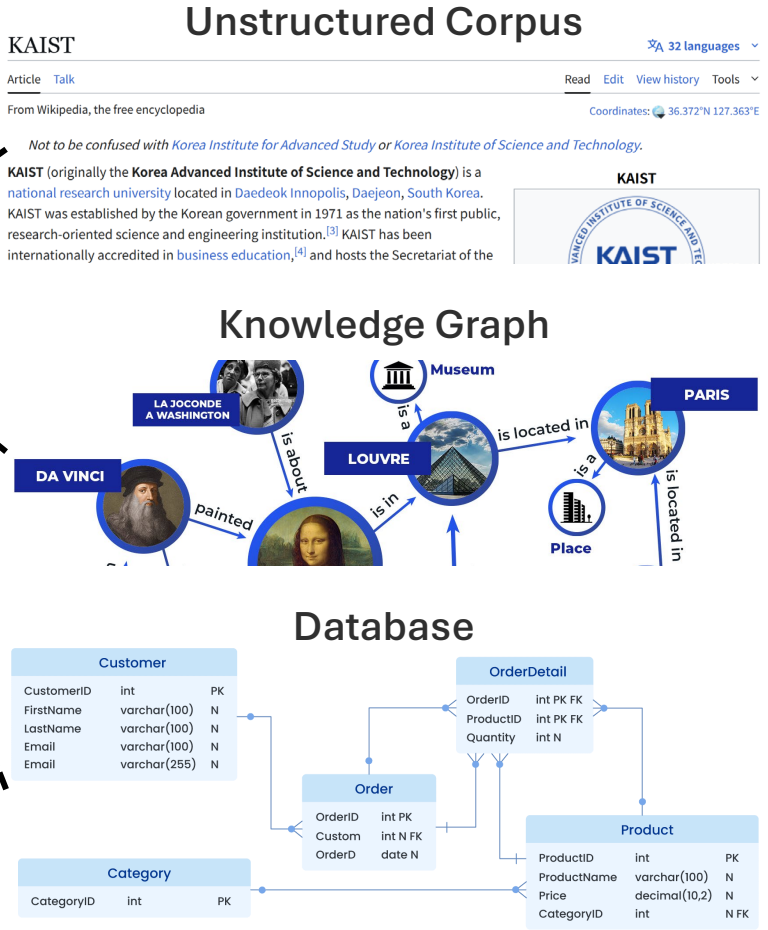
Recent approaches to reduce transformer inference latency?

Answer the user question by leveraging the retrieved knowledge from recent papers:
1. FlashAttention (…)
2. SpecInfer (…)
3. PagedAttention (…)

Question: Recent approaches to reduce …

Recent methods to reduce latency include FlashAttention for …

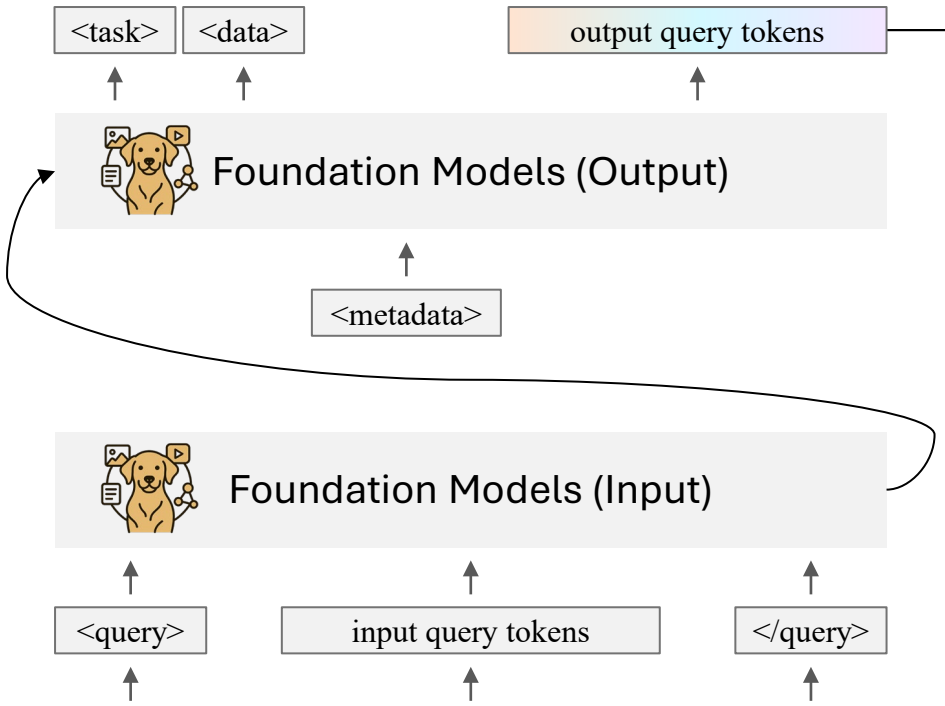




Unifying Access to Heterogeneous Knowledge Sources

I propose an any-to-any retrieval framework that unifies access to diverse knowledge sources, generating source-specific queries but also interleaving relevant metadata.

Sequence Layout for Universal Retriever



Which international collaborations has KAIST established in AI?

What international collaborations and partnerships in artificial intelligence has KAIST established with overseas universities, research institutes, and industry partners?

SELECT DISTINCT ?collab ?collabLabel ?partner ?partnerLabel ?country ?start ?end

WHERE {
BIND(<http://example.org/entity/KAIST> AS ?kaist)

...

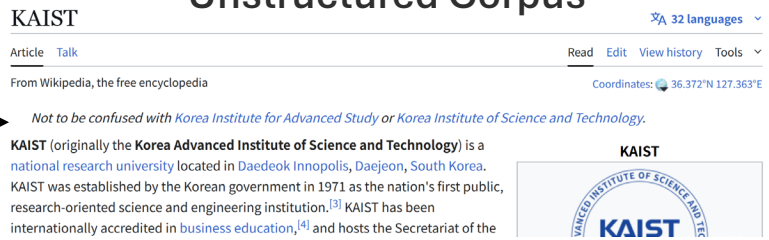
SELECT c.id, c.title, c.start_date, c.end_date,

FROM collaborations AS c

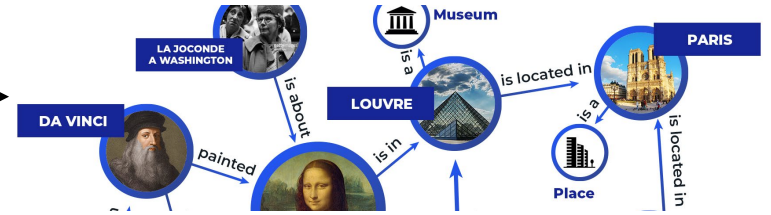
JOIN collaboration_participants AS cp1 ON cp1.collab_id = c.id

…

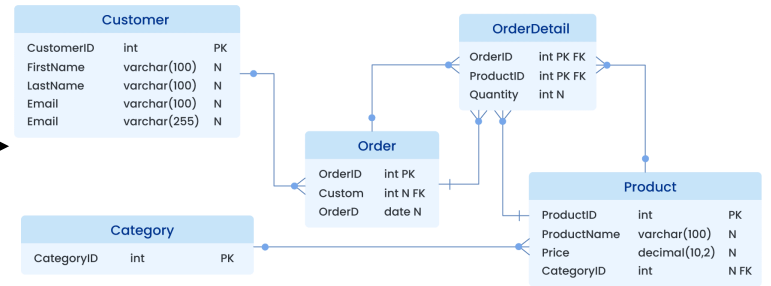
Unstructured Corpus



Knowledge Graph



Database

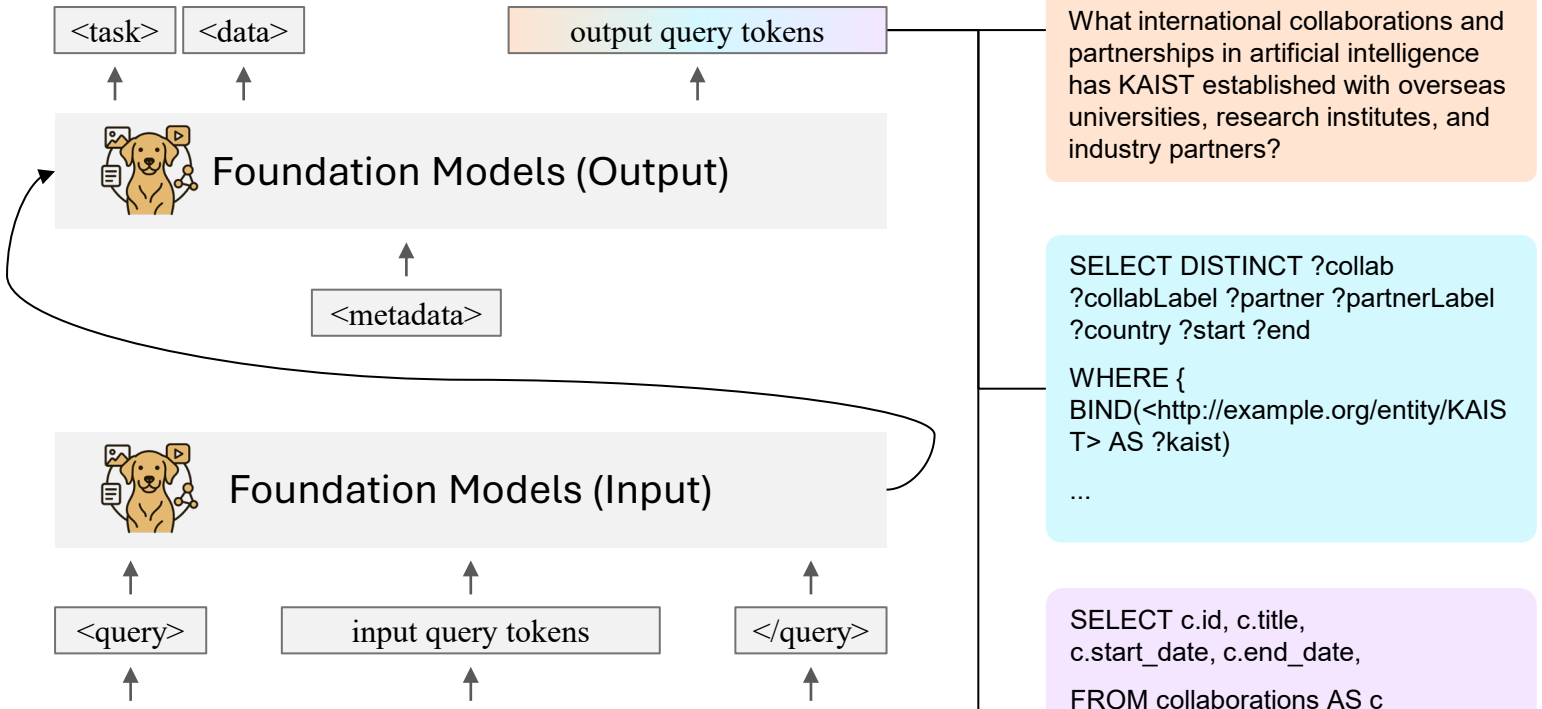


Baek et al., Universal Retrieval: Unifying Data Access Across Heterogeneous Knowledge Bases, In Preparation

Unifying Access to Heterogeneous Knowledge Sources

We validate Universal Retriever, showing that it achieves high special-token prediction accuracy and significantly outperforms single-modality retrievers.

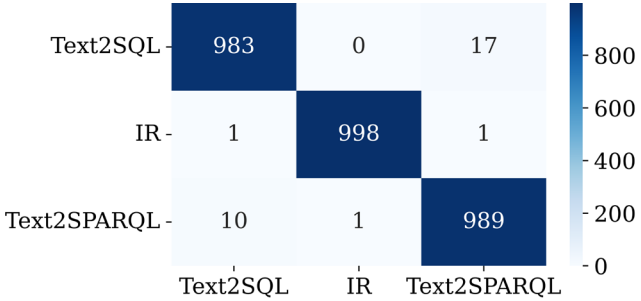
Sequence Layout for Universal Retriever



Special Token Results

Prediction Type	Accuracy
Task Token Prediction	99.00
Data Token Prediction	91.27

Confusion Matrix



Overall Results

Methods	Accuracy
Single-Modality Retriever	33.33
Universal Retriever (Ours)	56.93
Universal Retriever (Ours) w/ Oracle Special Tokens	63.33

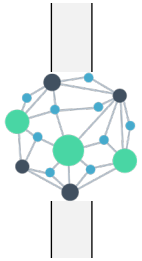
Baek et al., Universal Retrieval: Unifying Data Access Across Heterogeneous Knowledge Bases, In Preparation

Exploring Frontiers of Contextualization in Practice

I have also been committed to making foundation model contextualization methods useful and impactful to real-world applications, such as personalization and science.

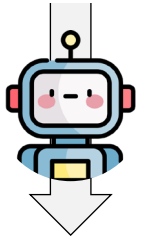
Personalization [1]

Knowledge Extraction



Entity	Availability
Latent Space	✓
Chain of Thought	✓
Sheaf Theory	✗

Personalization



Train Models to Do CoT in Continuous Latent Spaces

[1] Baek et al., WWW 2024; [2] Baek et al., NAACL 2024 (Oral Presentation); [3] Baek et al., ACL Findings 2025

Exploring Frontiers of Contextualization in Practice

I have also been committed to making foundation model contextualization methods useful and impactful to real-world applications, such as personalization and science.

Personalization [1]

Search Browsing

Knowledge Extraction

Entity	Availability
Latent Space	✓
Chain of Thought	✓
Sheaf Theory	✗

Personalization

Train Models to Do CoT in Continuous Latent Spaces

AI for Science [2]

Papers | Academic Graph | Concept (Entity) Store

Step 1. New Research Idea Generation

Step 2. Experimental Validation with Coding

Here is the new research idea with its experimental results …

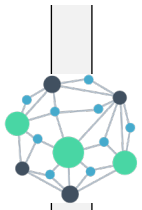
[1] Baek et al., WWW 2024; [2] Baek et al., NAACL 2024 (Oral Presentation); [3] Baek et al., ACL Findings 2025

Exploring Frontiers of Contextualization in Practice

I have also been committed to making foundation model contextualization methods useful and impactful to real-world applications, such as personalization and science.

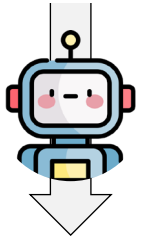
Personalization [1]

Knowledge Extraction



Entity	Availability
Latent Space	✓
Chain of Thought	✓
Sheaf Theory	✗

Personalization

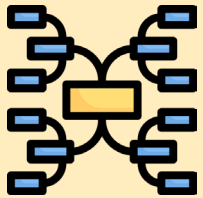


Train Models to Do CoT in Continuous Latent Spaces

AI for Science [2]

Papers

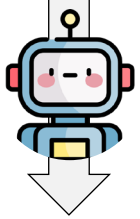
Academic Graph



Concept (Entity) Store

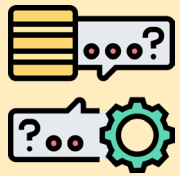
Step 1. New Research Idea Generation

Step 2. Experimental Validation with Coding



Here is the new research idea with its experimental results …

Text to SQL [3]



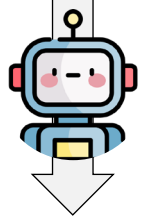
Queries

Schema

Knowledge

1) The Equity table contains columns MarketValue and Cost, …

2) The WACC formula is calculated as (EquityValue / TotalValue) * EquityCost + (DebtValue / TotalValue) * DebtCost * …

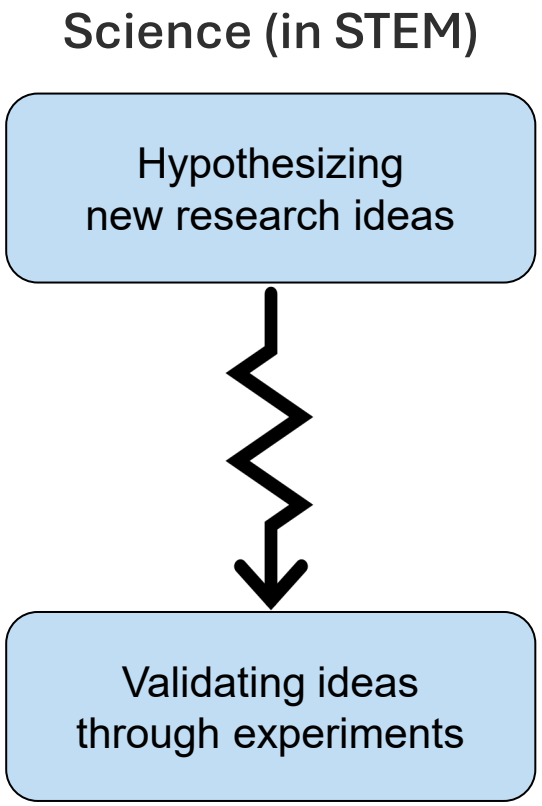


```
SELECT
  (e.Value / (e.Value + d.Value)) …
FROM Equity e
```


[1] Baek et al., WWW 2024; [2] Baek et al., NAACL 2024 (Oral Presentation); [3] Baek et al., ACL Findings 2025

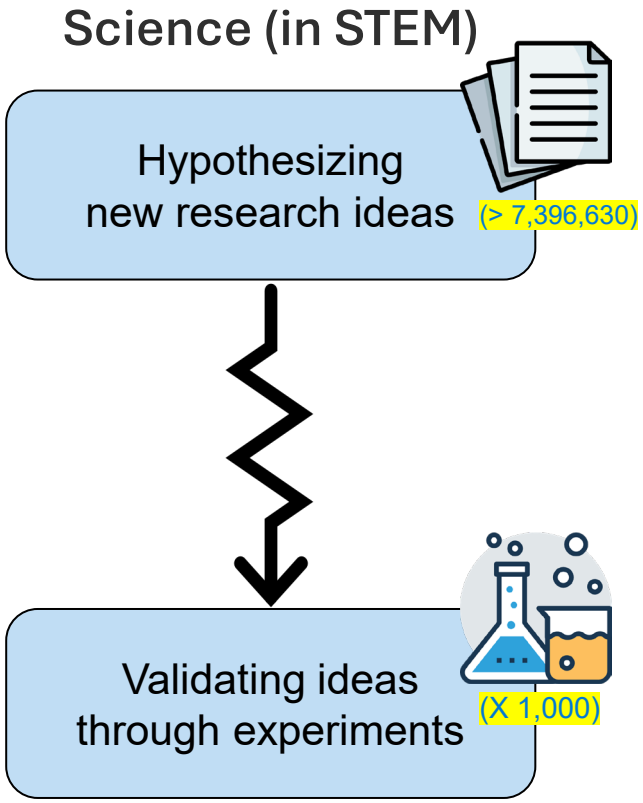
Accelerating Science through Model Contextualization

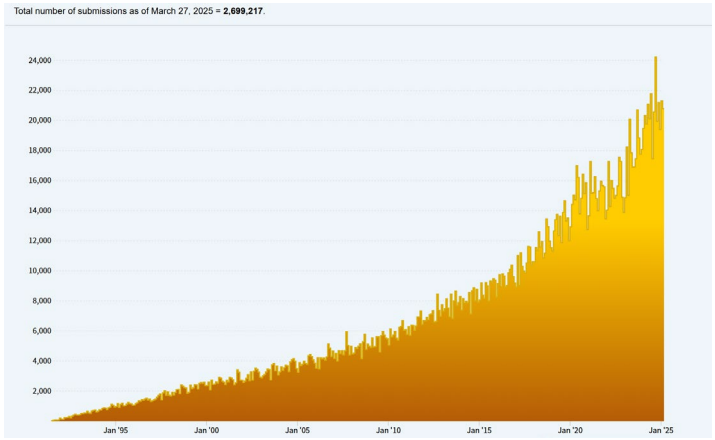
Scientific research plays a crucial role in driving innovation and improving human life, which involves hypothesizing new ideas and validating them with experimental trials.



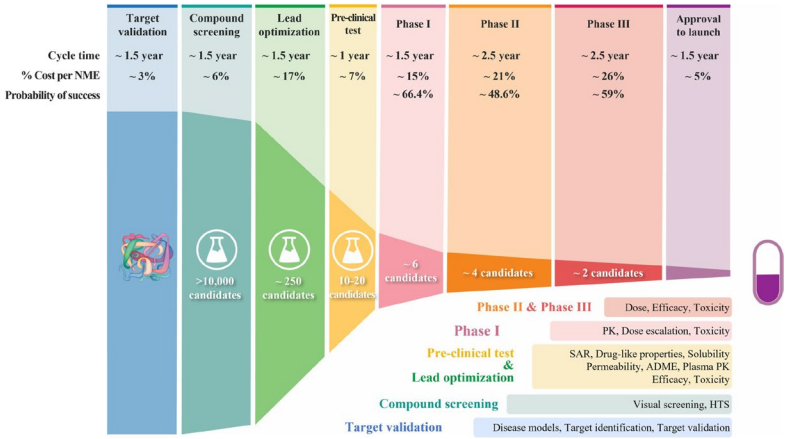
Accelerating Science through Model Contextualization

This process requires reading overwhelming amounts of knowledge over literature to formulate new ideas but also performing experimental validation of those new ideas.





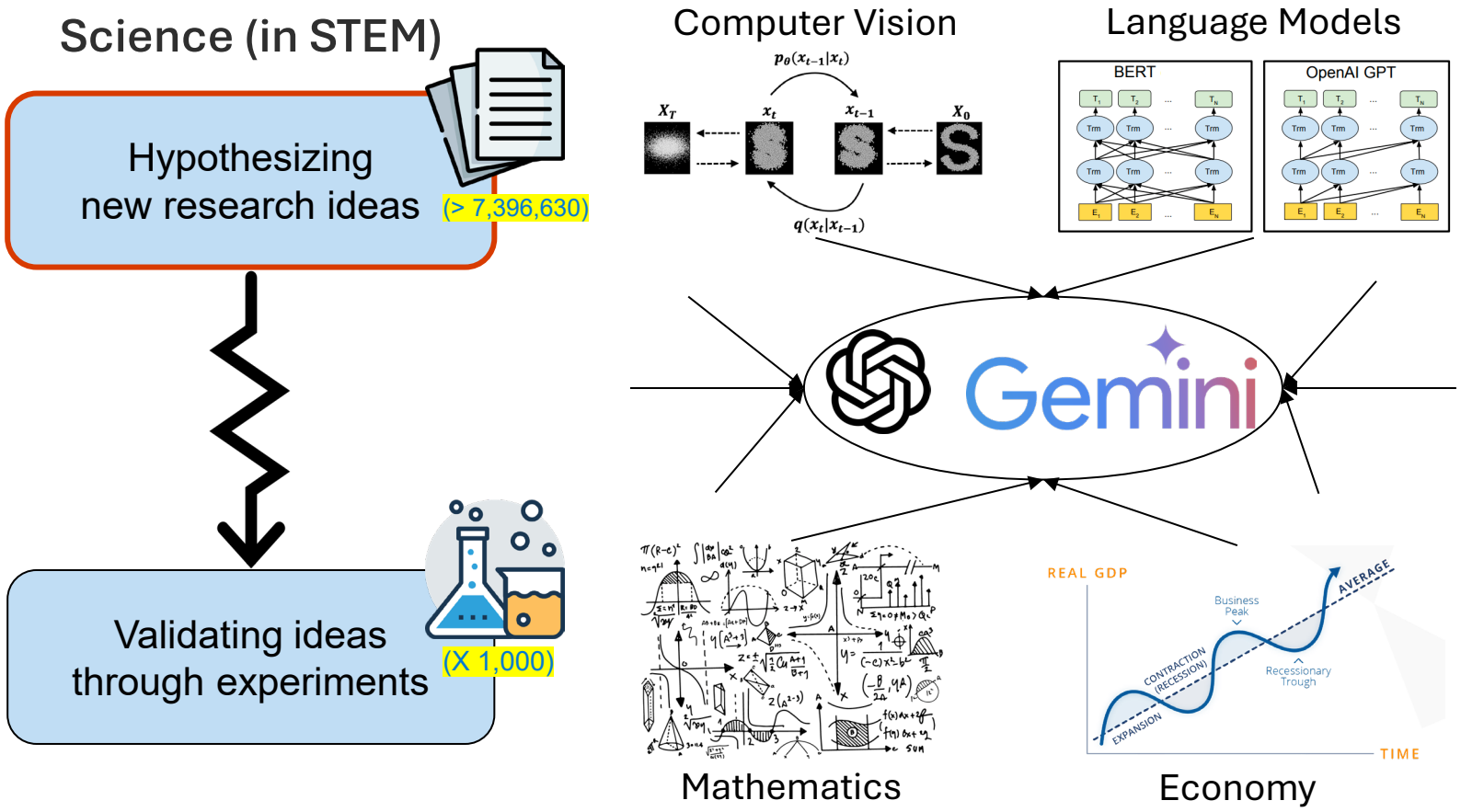
Overwhelming Literature



Validation Bottleneck

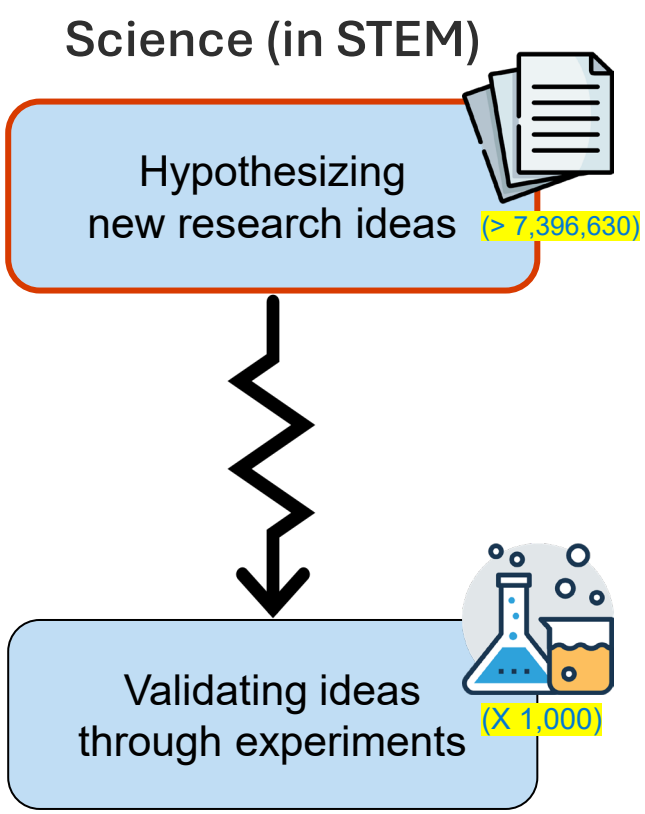
Accelerating Science through Model Contextualization

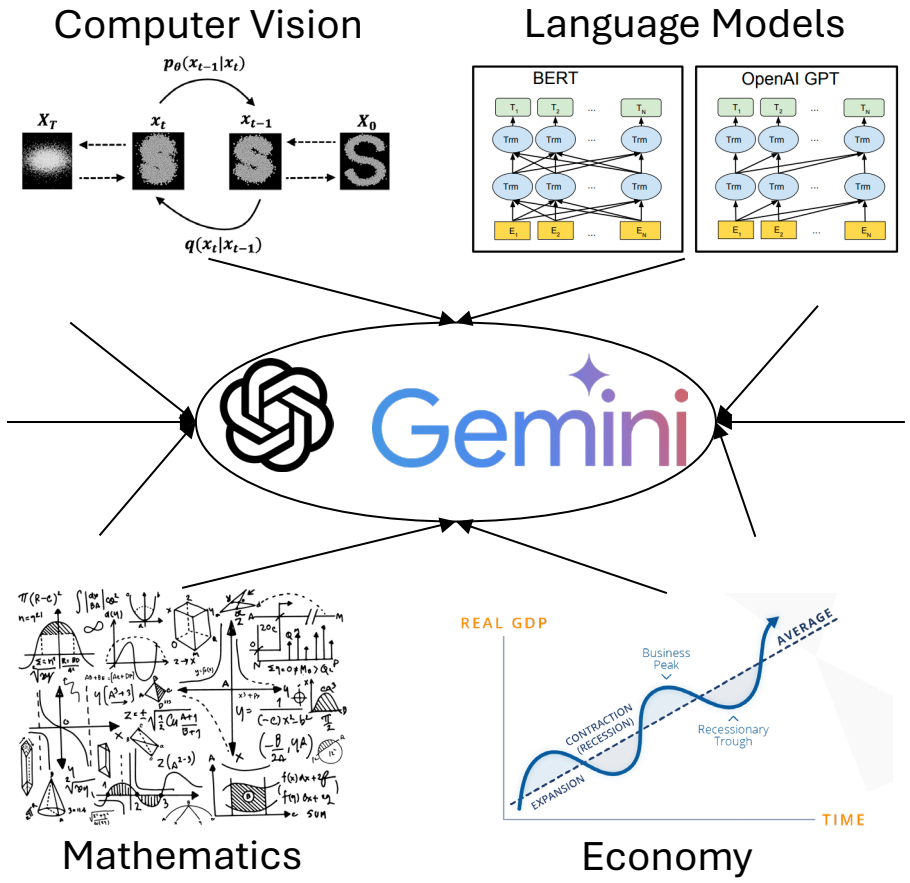
Models can process data at scales far beyond human capabilities and contain wider ranges of knowledge than any individual expert, potentially useful for idea generation.

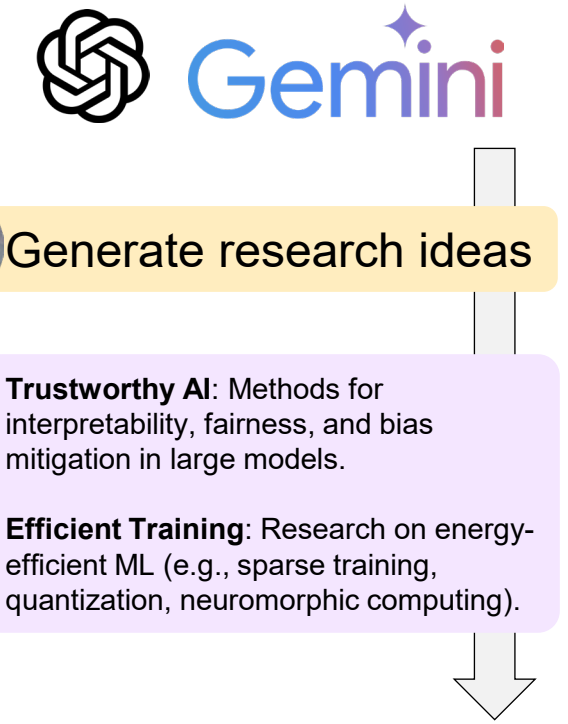


Accelerating Science through Model Contextualization

Yet, in the absence of proper contextualization, foundation models often produce ideas that are overly general, ambiguous, or misaligned with the intended goals.

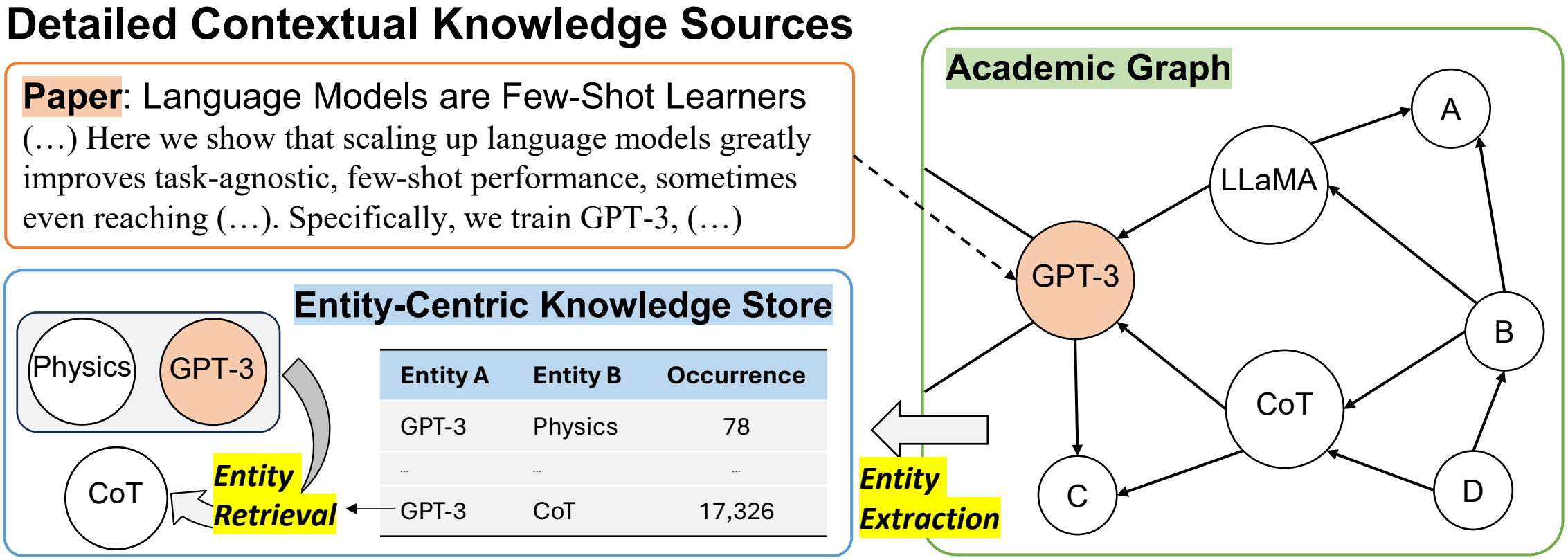






Contextual Knowledge Sources for Idea Generation

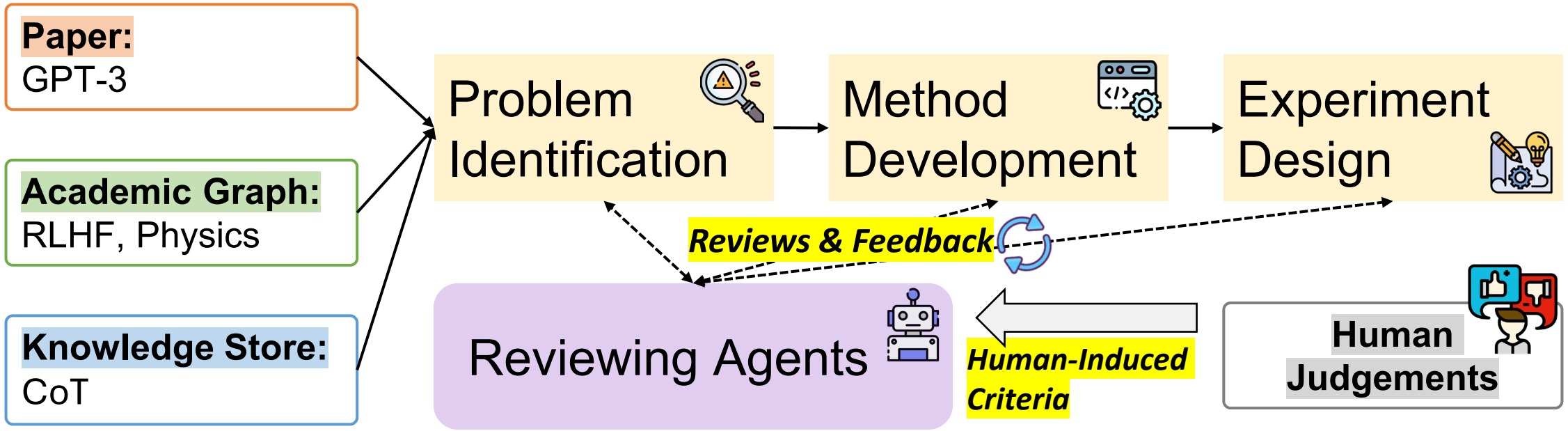
To support research idea generation, we contextualize foundation models with a core paper, their references and citations, and entities retrieved based on shared concepts.



Baek et al., ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models, NAACL 2025 (Oral, > 217 Citations, #6 Most Influential Paper)

Process of Research Idea Generation and Refinement

With contextual knowledge, we break the research process into problem identification, method development, and experiment design, and generate them accordingly.



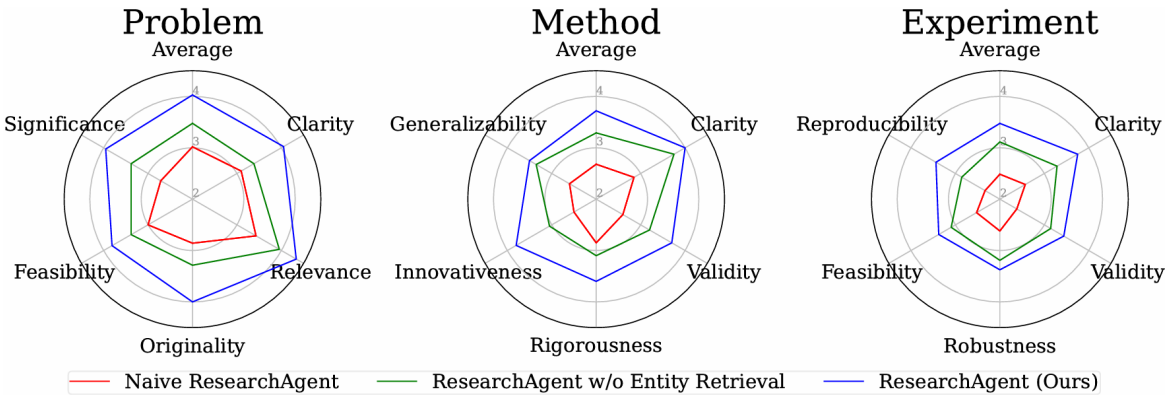
We further model how humans refine ideas through peer discussions by introducing Reviewing Agents, which are instantiated with criteria derived from human judgments.

Baek et al., ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models, NAACL 2025 (Oral, > 217 Citations, #6 Most Influential Paper)

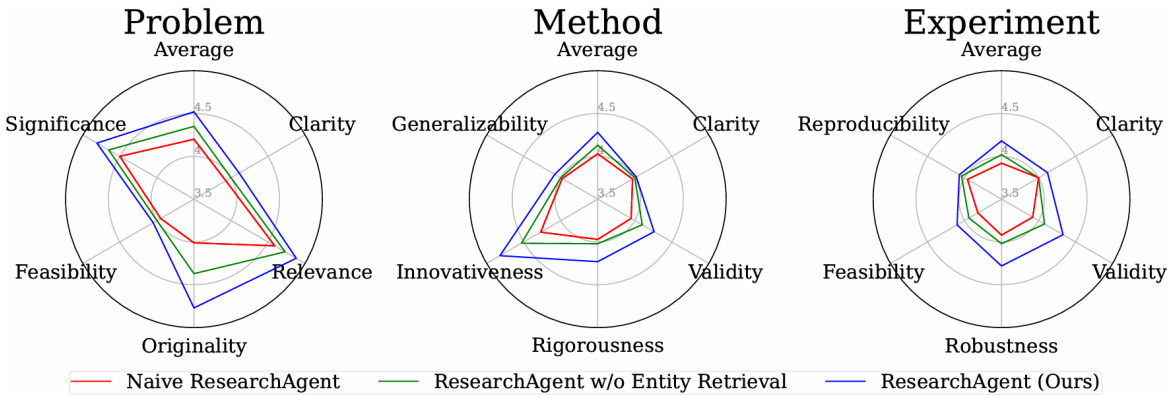
When Contextualized, Models Spark Stronger Ideas

We validate research ideas with both human and model-based evaluations, showing that contextualizing models with salient elements helps generate stronger ideas.

Human Evaluation Results



Model-based Evaluation Results

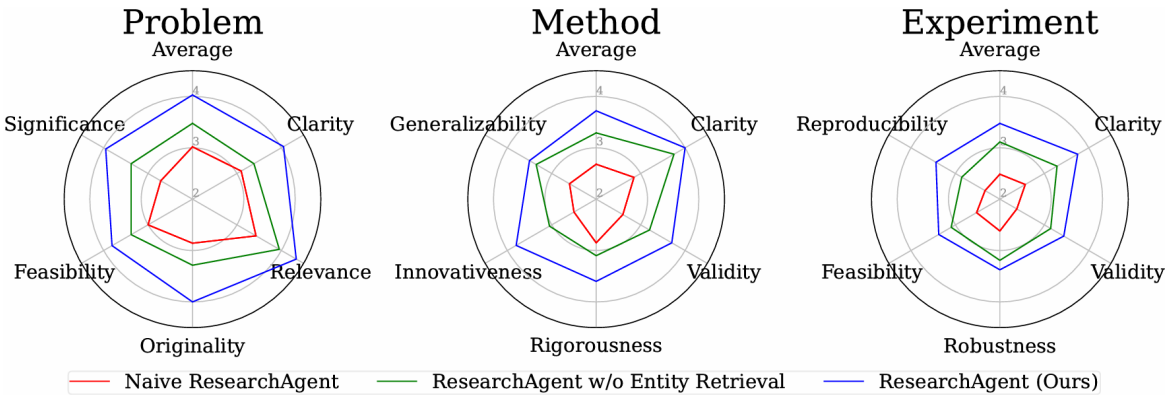


Baek et al., ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models, NAACL 2025 (Oral, > 217 Citations, #6 Most Influential Paper)

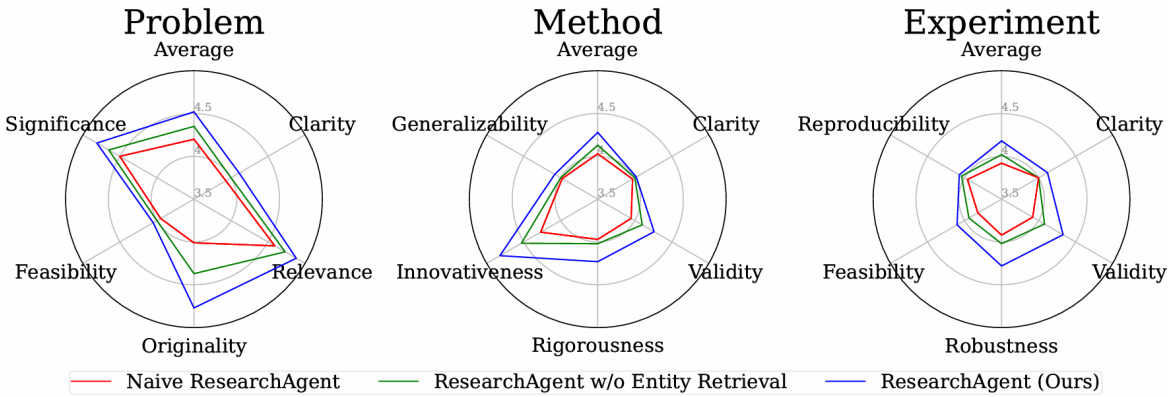
When Contextualized, Models Spark Stronger Ideas

We validate research ideas with both human and model-based evaluations, showing that contextualizing models with salient elements helps generate stronger ideas.

Human Evaluation Results



Model-based Evaluation Results



The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery (Lu et al., 2024, from SAKANA AI)
Can LLMs Generate Novel Research Ideas? (Si et al., 2024, from Stanford Univ.)
Agent Laboratory: Using LLM Agents as Research Assistants (Schmidgall et al., 2025, from AMD)

Baek et al., ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models, NAACL 2025 (Oral, > 217 Citations, #6 Most Influential Paper)

Impact: Media Coverage and Open-Source Release

ResearchAgent has rapidly gained broad recognition: cited widely, featured in Nature, highlighted among top systems, adopted in products, and highlighted by the media.

Cited no less than 217 (while published this year)

ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models — 217 — 2025
J Baek, SK Jauhar, S Cucerzan, SJ Hwang
NAACL 2025

10 AI Systems for Research

10 AI Systems for Scientific Research

Today, we offer you to explore these 10 AI systems for scientific research:

4. ResearchAgent implements LLMs to automate idea generation, methods, and experiment design, and ReviewingAgents' feedback to refine ideas:

ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models (2404.07738)

Talk at MSR

Media (MARKTECHPOST)

ResearchAgent: Transforming the Landscape of Scientific Research Through AI-Powered Idea Generation and Iterative Refinement

By Sana Hassan - April 14, 2024

Quoted in Nature (2025) for Expert Commentary

paper, he adds. Ionescu told *Nature* he would give the AI-generated paper a rating of 2 or 3.

Park judges the overlap with his paper to be much stronger than Hoover's and Ionescu's ratings. He says he would give it a score of 5 on Gupta's scale, and adds that it "reflects a strong methodological resemblance that I consider noteworthy." Even so, this does not necessarily align with what he sees as the legal or ethical definition of plagiarism, he told *Nature*.

"There's no one way to prove idea plagiarism."

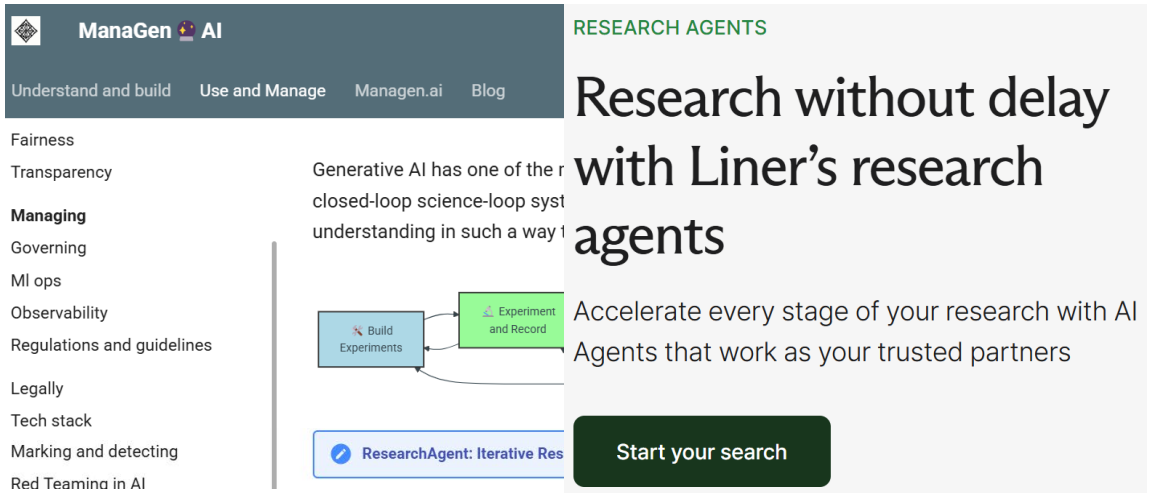
already said in its paper that, in general, The AI Scientist makes citation mistakes; that it should cite more related papers; and that researchers

their study. Turnitin identified none of the source papers their human experts had spotted, whereas OpenScholar found only one, they say.

But human reviewers disagree about this sort of thing, too, says Jinheon Baek, a graduate student in AI at KAIST. At conferences, he says, he's seen reviewers argue about what counts as original in research papers. "Novelty is very subjective," he says.

600 | Nature | Vol 644 | 21 August 2025

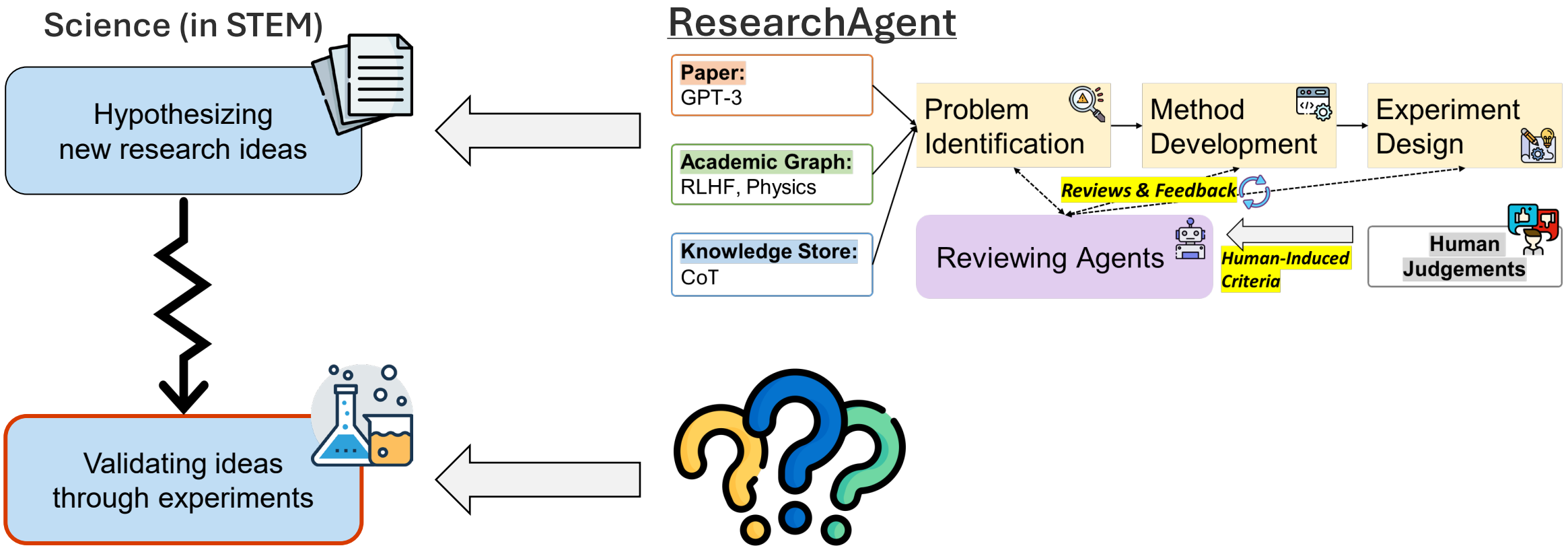
Implemented in Open-Source Library and Product



Baek et al., ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models, NAACL 2025 (Oral, > 217 Citations, #6 Most Influential Paper)

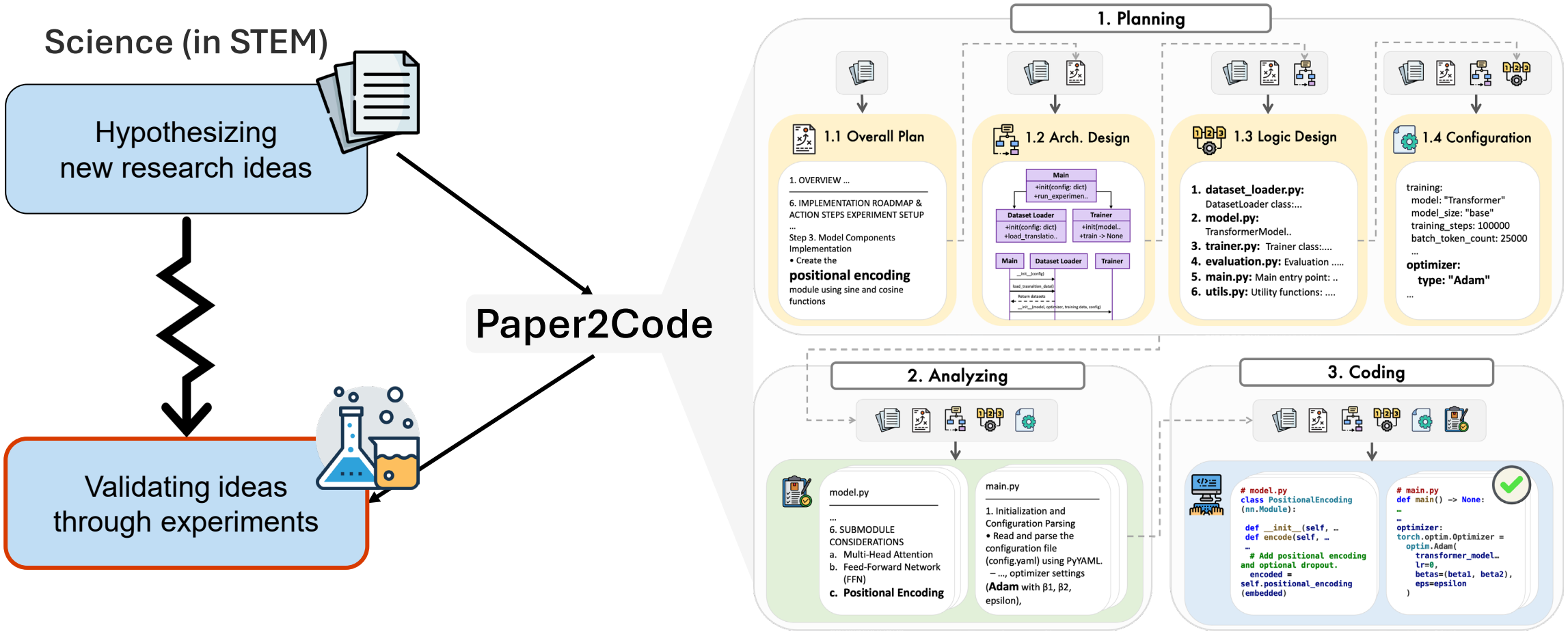
Extensions of ResearchAgent

While ResearchAgent can empower the ideation process,
science also requires implementing and executing ideas to validate them.



Extensions of ResearchAgent

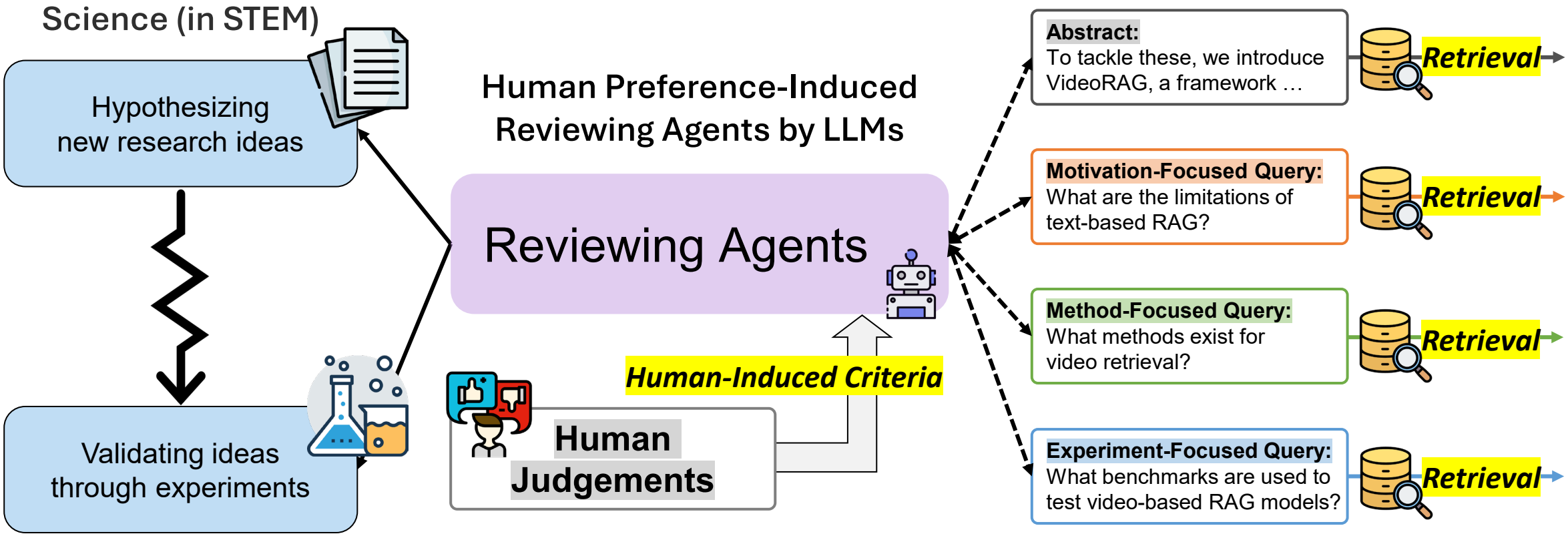
To enable implementation and execution, we propose PaperCoder, which automatically converts ideas (or papers) into code with multiple steps and agents.



Seo, Baek et al., Paper2Code: Automating Code Generation from Scientific Papers in Machine Learning, DL4C @ NeurIPS 2025

Extensions of ResearchAgent

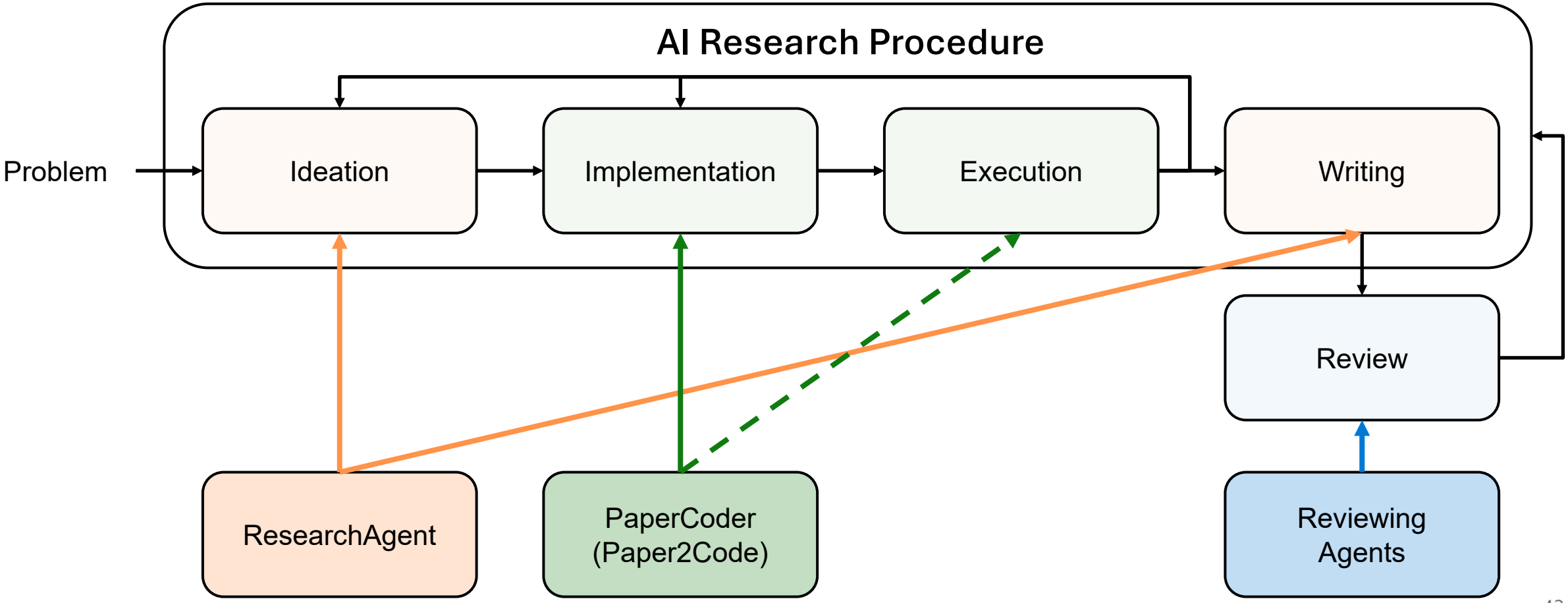
Reviewing Agents are designed to iteratively improve research ideas, and as their performance hinges on contextual information, we introduce a new retrieval approach.



Park, Baek et al., Chain of Retrieval: Multi-Aspect Iterative Search Expansion and Post-Order Search Aggregation for Full Paper Retrieval, arXiv 2025

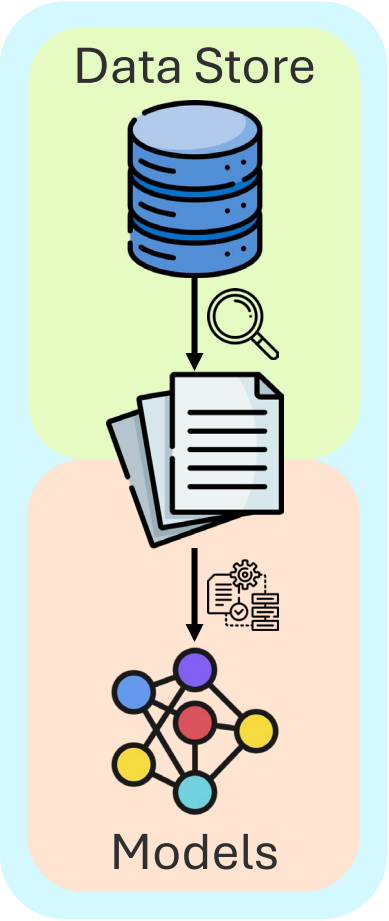
AI-Empowered R&D (for AI)

We believe that ResearchAgent, Reviewing Agents, and PaperCoder can accelerate and empower the entire R&D life cycle for AI, transforming how we do (AI) research.



Summary

In my PhD study, I aim to go beyond (scaling) parameters (and in scope),
with emphasis on contextualization, knowledge retrieval, and their applications.



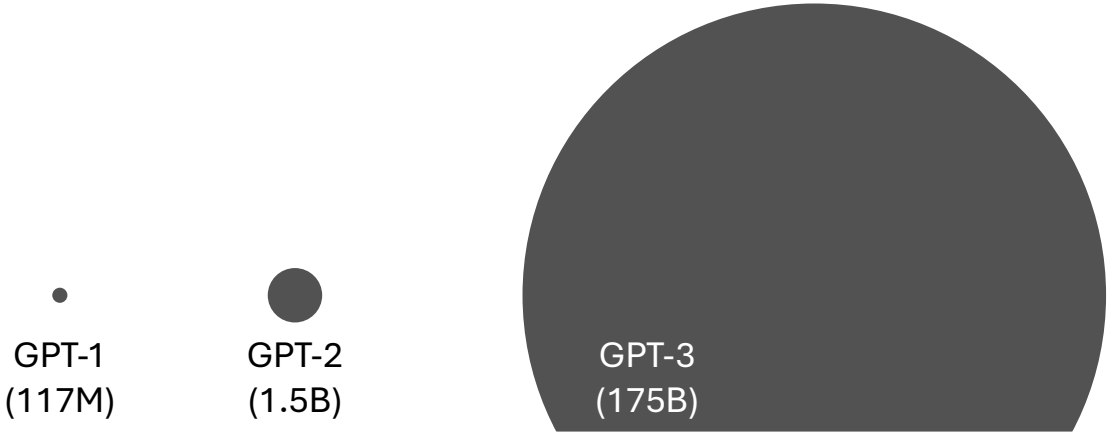
Part 1. Advancing Foundation Models with Contextualization

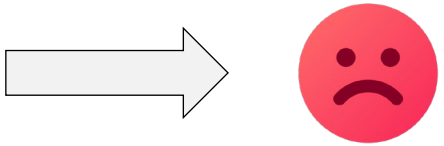
Part 2. Representing and Retrieving Knowledge for Contextualization

Part 3. Expanding Contextualization to Real-World Applications

Why Contextualization Still Matters Beyond Training

Even though smaller open-weight models become more accessible, contextualization remains valuable; for example, it cannot scale to per-user training for personalization.

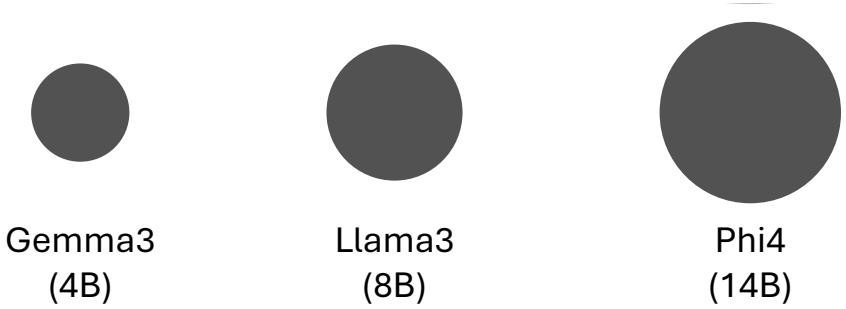


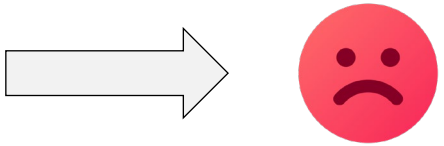


Too costly to train:

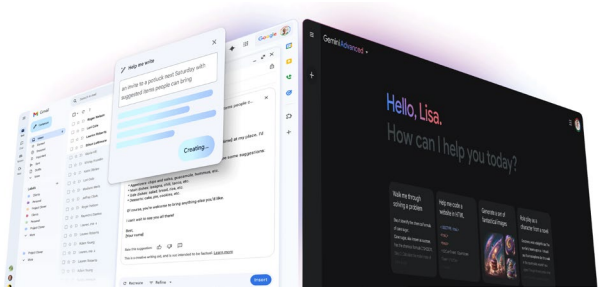
- Training: 48 x H100 (80GB)
- Inference: 8 x H100 (80GB)

Alternatives?



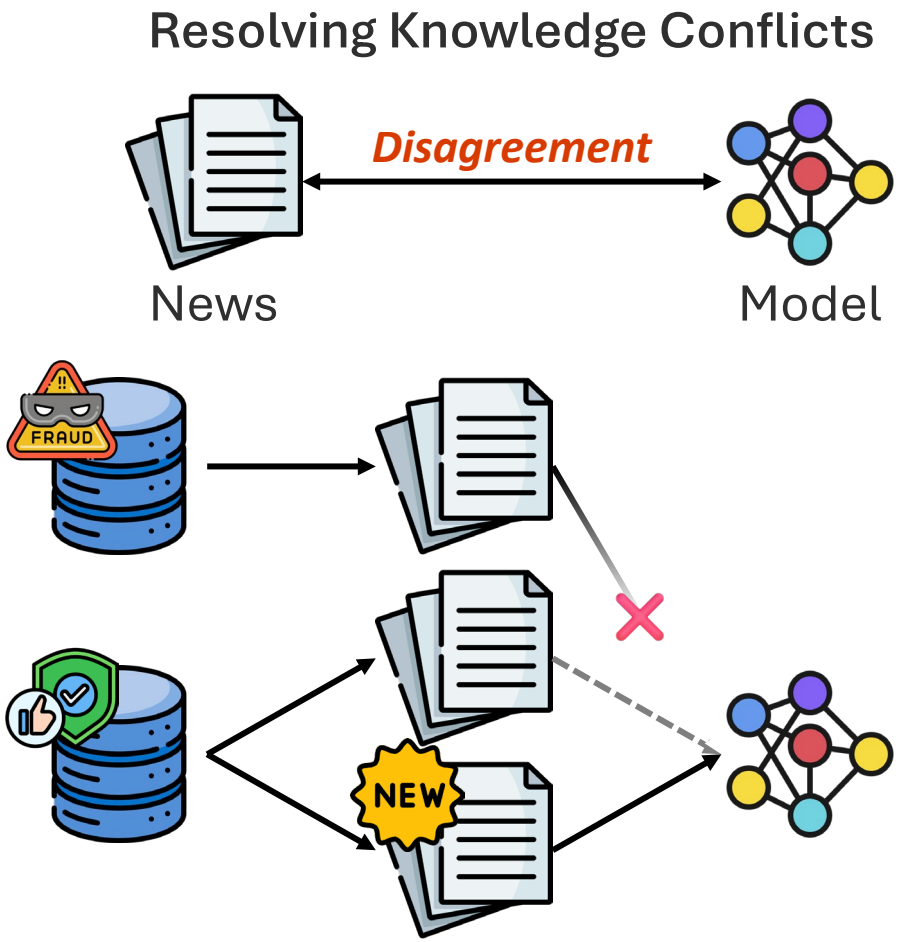


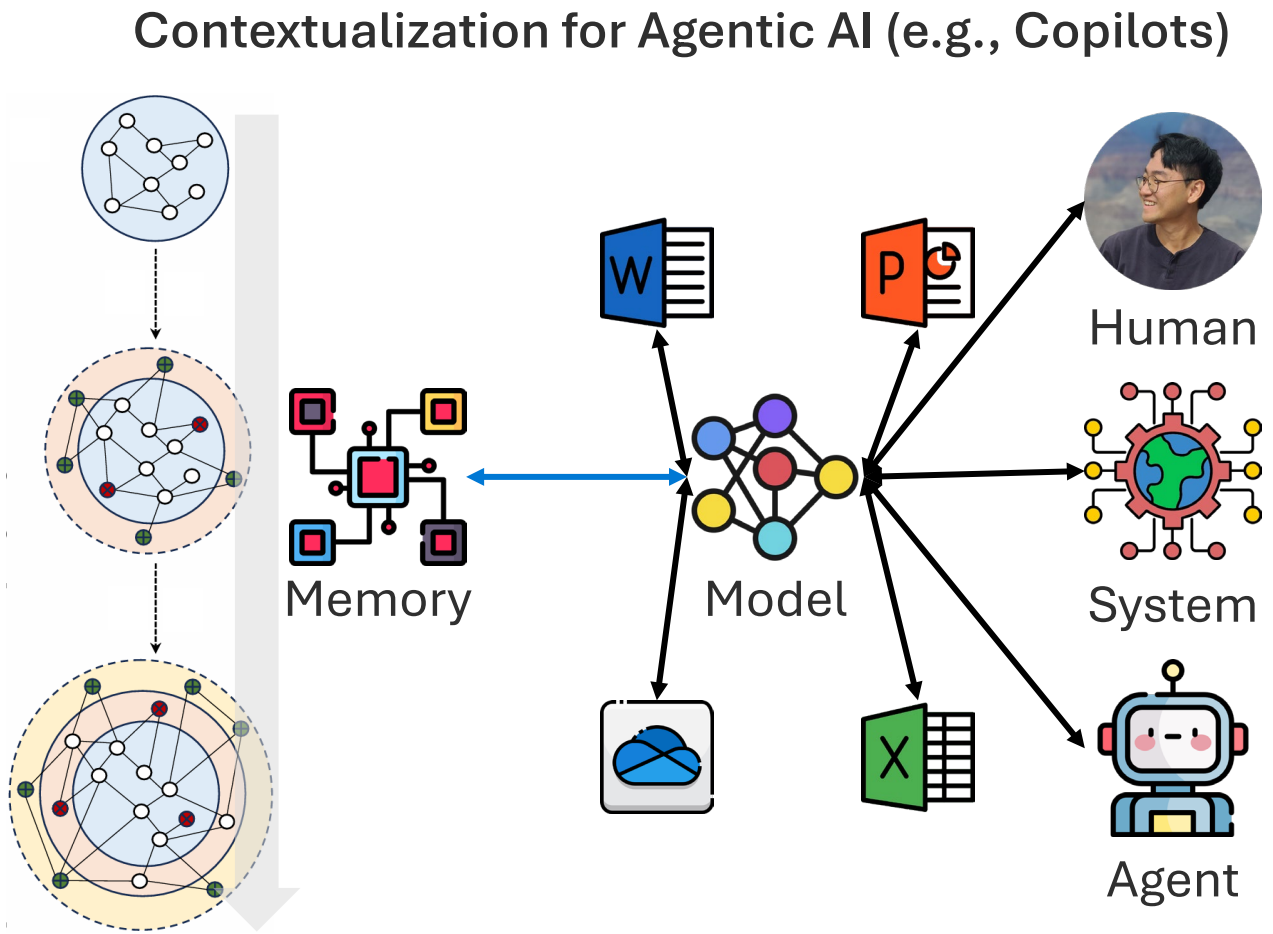
Still costly per user



Is Current Contextualization Enough?

Next-generation contextualization should not only handle knowledge conflicts but also incorporate memory to leverage past interactions, possibly via post-optimization.

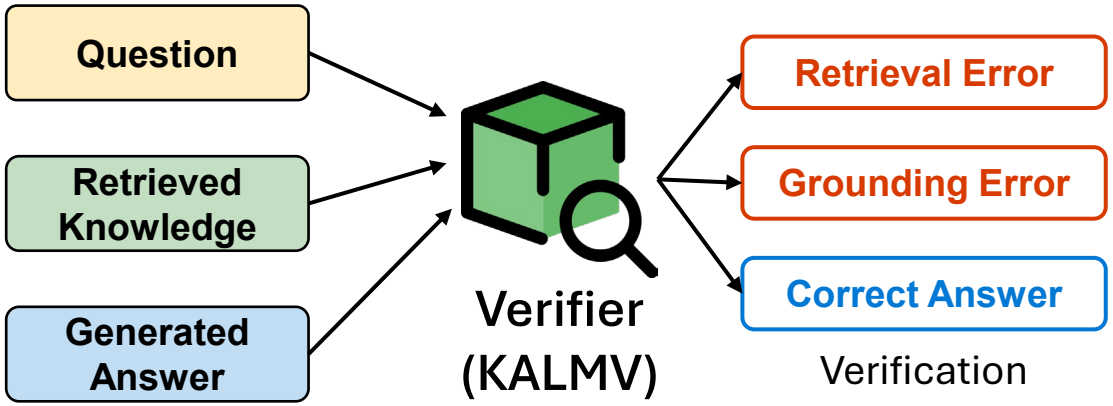


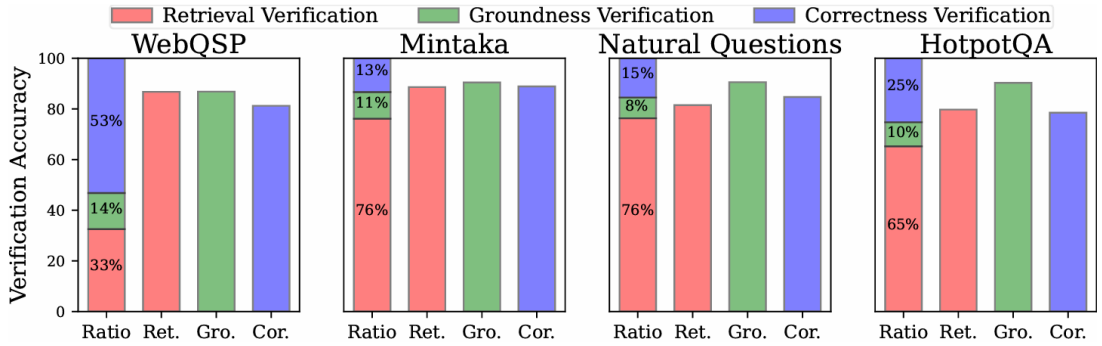


Verification Is Easier than Generation — Is it?

Recall that verification is generally easier than generation in the context of retrieval and grounding errors; but does this advantage persist in other settings?

Verification in Prior Model Contextualization [1]



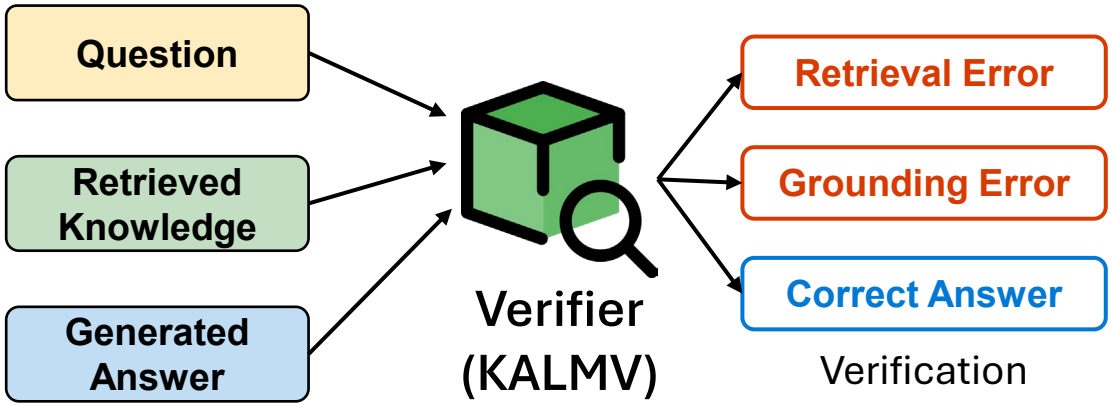


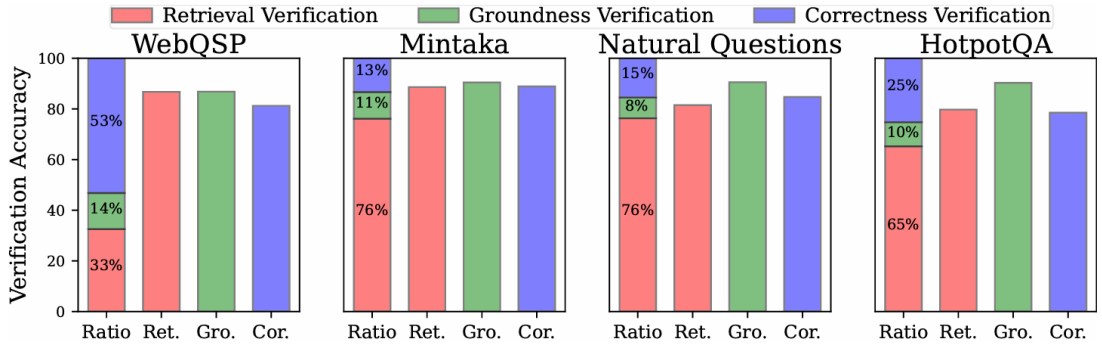
[1] Baek et al., Knowledge-Augmented Language Model Verification, EMNLP 2023

Verification Is Easier than Generation — Is it?

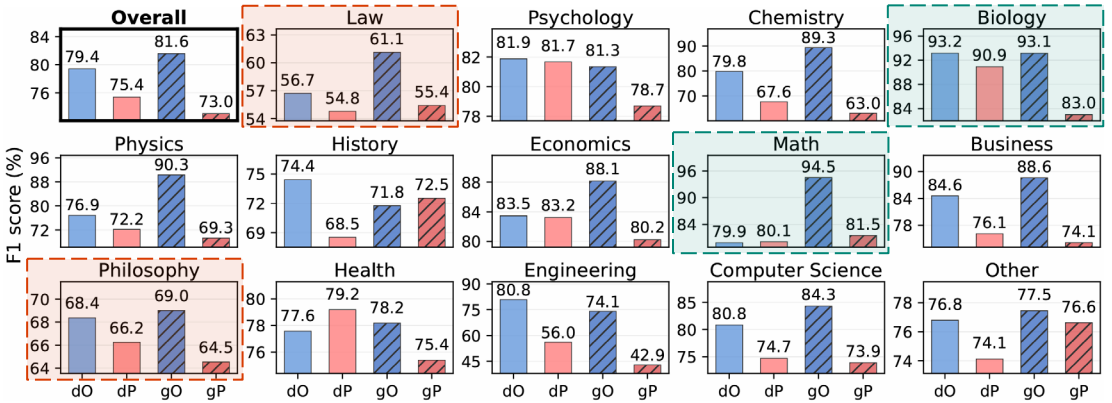
Reward models struggle in areas where interpretations vary across jurisdictions, framings, and perspectives, which could be pronounced in assessing research ideas.

Verification in Prior Model Contextualization [1]

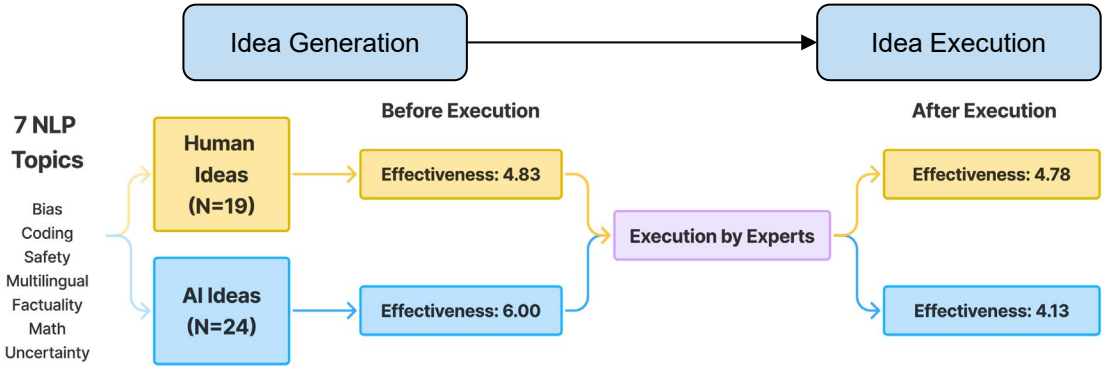




Verification Results in Various Domains [2]



Verification in (AI for) Science (Si et al., 2025)

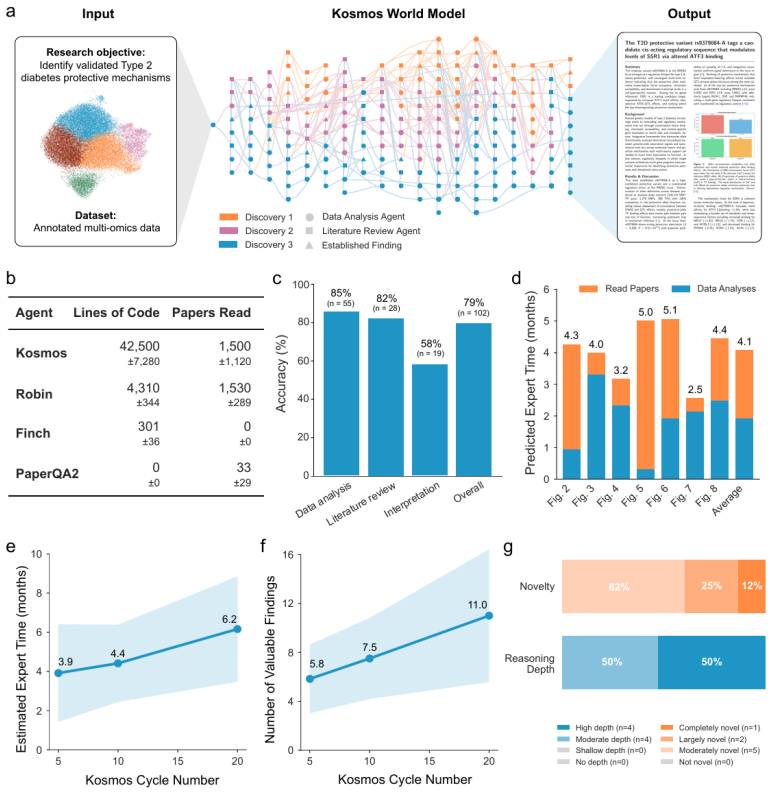


[1] Baek et al., Knowledge-Augmented Language Model Verification, EMNLP 2023; [2] Lee, …, Baek et al., Rethinking Reward Models for Multi-Domain Test-Time Scaling, arXiv 2025

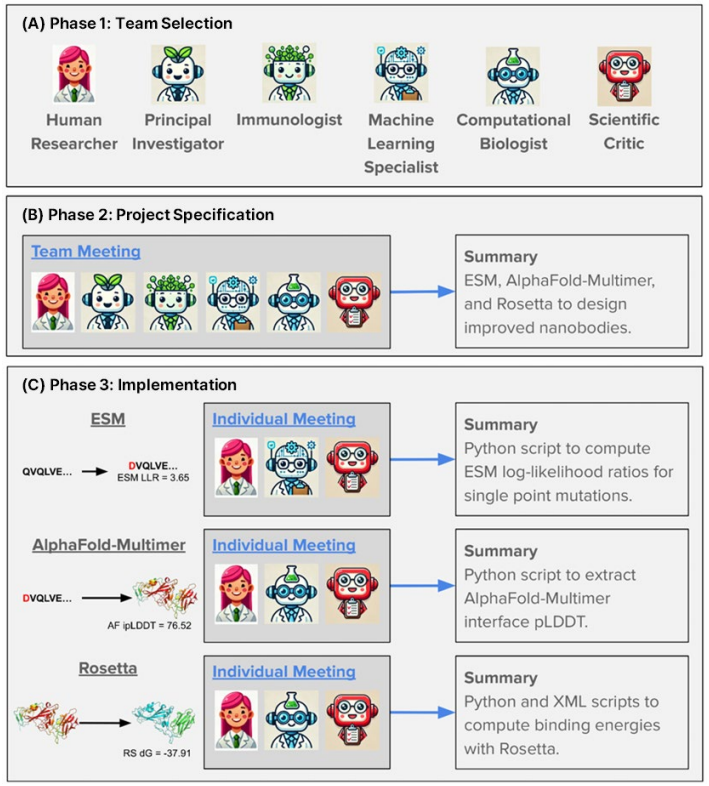
What Should We Do for Science (Powered by AI)?

AI outputs can still be surprisingly meaningful (as shown in Kosmos and by Prof. Ryu), but their verification with AI might not be trustworthy: humans must remain in the loop.

Kosmos: An AI Scientist [1]



The Virtual Lab [2]



AI-Assisted Math Proof [3]

Ernest Ryu
@ErnestRyu

I used ChatGPT to solve an open problem in convex optimization.

Part I

(1/N)

7:14 AM · Oct 22, 2025 · **1.2M** Views

Ernest Ryu
@ErnestRyu

Preprint on using ChatGPT to resolve a 42-year-old open problem (point convergence of Nesterov's accelerated gradient method) is out.

Mathematical results are complete, though still need to expand the discussion of historical context & prior work.
(1/2)

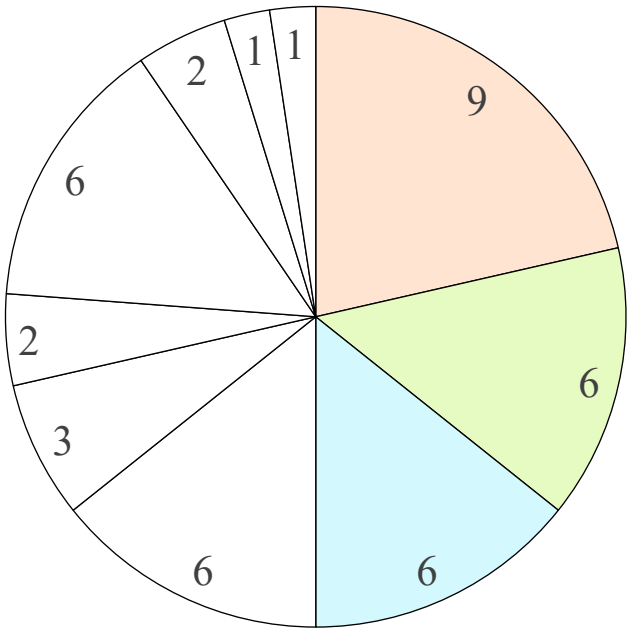
arxiv.org
Point Convergence of Nesterov's Accelerated Gradient Me...
The Nesterov accelerated gradient method, introduced in 1983, has been a cornerstone of optimization theory and ...

12:39 AM · Oct 29, 2025 · **82.8K** Views

[1] Mitchener et al., Kosmos: An AI Scientist for Autonomous Discovery, arXiv 2025; [2] Swanson et al., The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies, Nature 2025;
[3] Jang & Ryu, Point Convergence of Nesterov's Accelerated Gradient Method: An AI-Assisted Proof, arXiv 2025

Publication List

Papers by Category

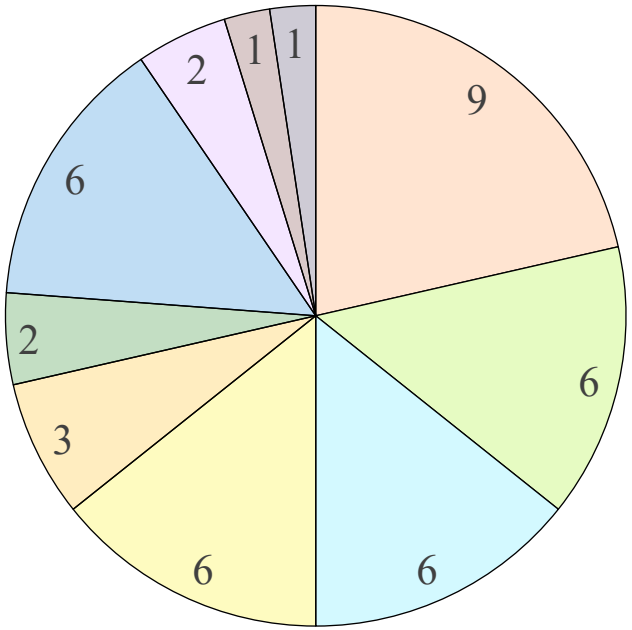


	Paper Title	Covered
01	Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering	🟢
02	Knowledge-Augmented Language Model Verification	🟢
03	VideoRAG: Retrieval-Augmented Generation over Video Corpus	🟢
04	Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity (⭐)	🟧
05	Efficient Real-time Refinement of Language Model Text Generation	🟧
06	UniversalRAG: Retrieval-Augmented Generation over Corpora of Diverse Modalities and Granularities	🟧
07	Knowledge Graph-Augmented Language Models for Knowledge-Grounded Dialogue Generation	❌
08	KALA: Knowledge-Augmented Language Model Adaptation (🎙️)	❌
09	Knowledge-Augmented Reasoning Distillation for Small Language Models in Knowledge-Intensive Tasks	❌
10	Unified Multi-Modal Interleaved Document Representation for Retrieval	🟢
11	Direct Fact Retrieval from Knowledge Graphs without Entity Linking	🟢
12	Universal Retrieval: Unifying Data Access Across Heterogeneous Knowledge Bases	🟢
13	Database-Augmented Query Representation for Information Retrieval (🎙️)	❌
14	Augmenting Document Representations for Dense Retrieval with Interpolation and Perturbation (🎙️)	❌
15	Unsupervised Document Expansion for Information Retrieval with Stochastic Text Generation	❌
16	Knowledge-Augmented Large Language Models for Personalized Contextual Query Suggestion	🟢
17	ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models (🎙️ | ✨)	🟢
18	Knowledge Base Construction for Knowledge-Augmented Text-to-SQL	🟢
19	Paper2Code: Automating Code Generation from Scientific Papers in Machine Learning	🟧
20	Chain of Retrieval: Multi-Aspect Iterative Search Expansion and Post-Order Search Aggregation for Full Paper Retrieval	🟧
21	Retrieval-Augmented Data Augmentation for Low-Resource Domain Tasks	❌

⭐: Ranked #7 Most Influential Paper at NAACL 2024 (as of Sep 2025); ✨: Ranked #6 Most Influential Paper at NAACL 2025 (as of Sep 2025); 🎙️: Oral Presentation.

Publication List

Papers by Category

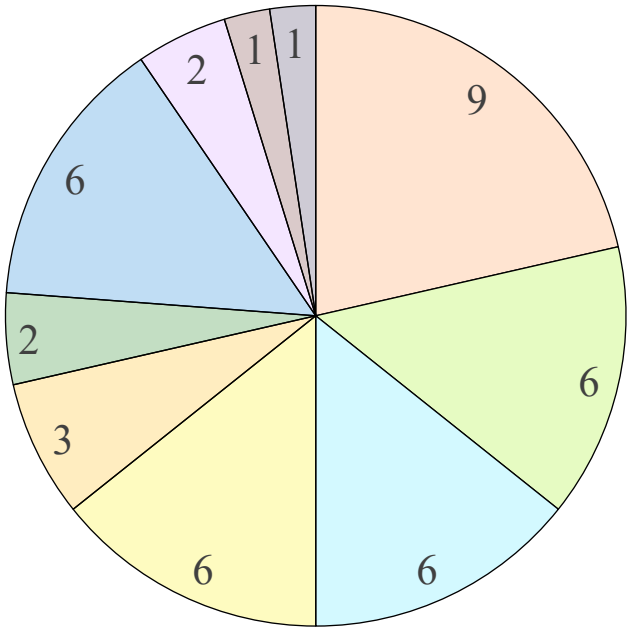


	Paper Title	Covered
22	Sketch-of-Thought: Efficient LLM Reasoning with Adaptive Cognitive-Inspired Sketching	✗
23	Towards Better Understanding of Program-of-Thought Reasoning in Cross-Lingual and Multilingual Environments	✗
24	An Empirical Study of Multilingual Reasoning Distillation for Question Answering	✗
25	CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark (🎙️)	✗
26	CaMMT: Benchmarking Culturally Aware Multimodal Machine Translation	✗
27	Exploring The Spatial Reasoning Ability of Neural Models in Human IQ Test	✗
28	Test-Time Self-Adaptive Small Language Models for Question Answering	✗
29	Phrase Retrieval for Open-Domain Conversational Question Answering with Conversational Dependency Modeling	✗
30	Realistic Conversational Question Answering with Answer Selection based on Calibrated Confidence and Uncertainty	✗
31	Revisiting In-Context Learning with Long Context Language Models	✗
32	Efficient Long Context Language Model Retrieval with Compression	✗
33	Personalized Subgraph Federated Learning	✗
34	Graph Self-supervised Learning with Accurate Discrepancy Learning	✗
35	Edge Representation Learning with Hypergraphs	✗
36	Accurate Learning of Graph Representations with Graph Multiset Pooling	✗
37	Learning to Extrapolate Knowledge: Transductive Few-shot Out-of-Graph Link Prediction	✗
38	Object Detection in Aerial Images with Uncertainty-Aware Graph Network	✗
39	System Prompt Optimization with Meta-Learning	✗
40	Task-Adaptive Neural Network Retrieval with Meta-Contrastive Learning (🔧)	✗
41	Rethinking Code Refinement: Learning to Judge Code Efficiency	✗
42	The BiGGen Bench: A Principled Benchmark for Fine-grained Evaluation of Language Models with Language Models (🏆)	✗

🏆 : Received the Best Paper Award at NAACL 2025; 🎙️ : Oral Presentation; 🔧 : Spotlight Presentation.

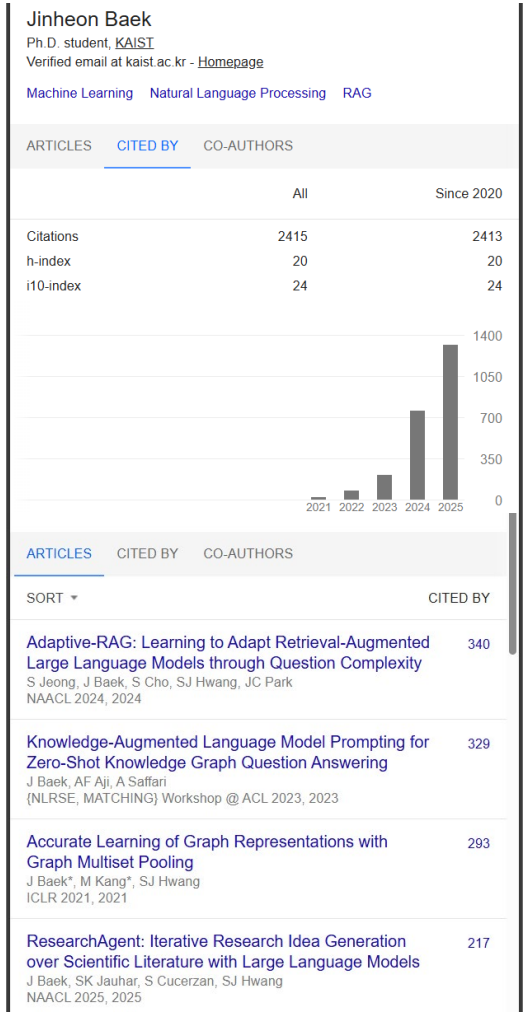
Achievements in My Ph.D. Journey

Papers (> 40)



- Contextualization
- Retrieval
- Applications
- Reasoning
- QA & Conversation
- Long-Context Modeling
- Graph Learning
- Meta-Learning
- Code
- Benchmark

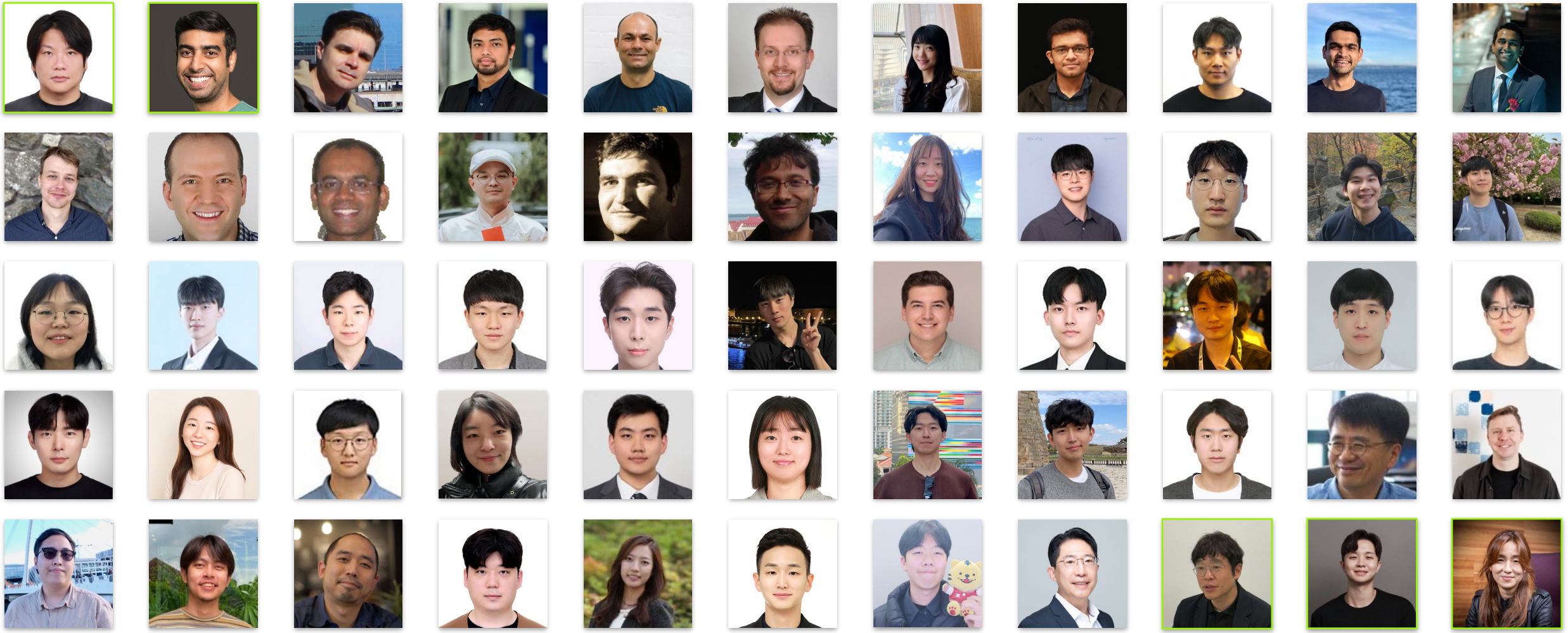
Citations (> 2400)



*Ranked 36th Among All Koreans

Yejin Choi (1.000, 90%)	Youngjae Yu (0.212, 99%)
Sung Ju Hwang (0.817, 100%)	Moontae Lee (0.212, 99%)
Jinwoo Shin (0.625, 100%)	Kee-Eung Kim (0.212, 100%)
Honglak Lee (0.471, 89%)	Sungjin Ahn (0.212, 87%)
Kyunghyun Cho (0.394, 99%)	Sewon Min (0.192, 97%)
Eunho Yang (0.365, 100%)	Chulhee Yun (0.192, 98%)
Gunhee Kim (0.327, 100%)	Hwanjun Song (0.183, 94%)
Sewoong Oh (0.327, 99%)	Jaehong Yoon (0.183, 99%)
Minjoon Seo (0.288, 86%)	Kangwook Lee (0.183, 99%)
Se-Young Yun (0.279, 99%)	Noseong Park (0.173, 91%)
Kimin Lee (0.269, 82%)	Jong Chul Ye (0.173, 99%)
Sungroh Yoon (0.269, 100%)	Dongyeop Kang (0.163, 99%)
Juho Lee (0.269, 71%)	Jaewoo Kang (0.163, 100%)
Jaegul Choo (0.260, 100%)	Sangdoo Yun (0.154, 99%)
Sungsoo Ahn (0.260, 99%)	Jung-Woo Ha (0.154, 99%)
Seung-won Hwang (0.240, 100%)	Youngchul Sung (0.154, 99%)
Eunsol Choi (0.240, 96%)	SangKeun Lee (0.144, 100%)
Jinkyoo Park (0.221, 99%)	Jinheon Baek (0.144, 97%)

*: According to the number of publications at NeurIPS, ICML, ICLR, ACL, EMNLP, and NAACL from 2020 to 2025.



KOREA UNIVERSITY

KAIST

amazon

Microsoft

IBM

Google