Building the Autonomous Research Scientist: From Exploration Mechanisms to Scholarly Integrity

Yi R. (May) Fung

yrfung@ust.hk

Assistant Professor

Hong Kong University of Science and Technology

Jan 25, 2026

We’re Living in a Very Exciting Era of AI Advancements!

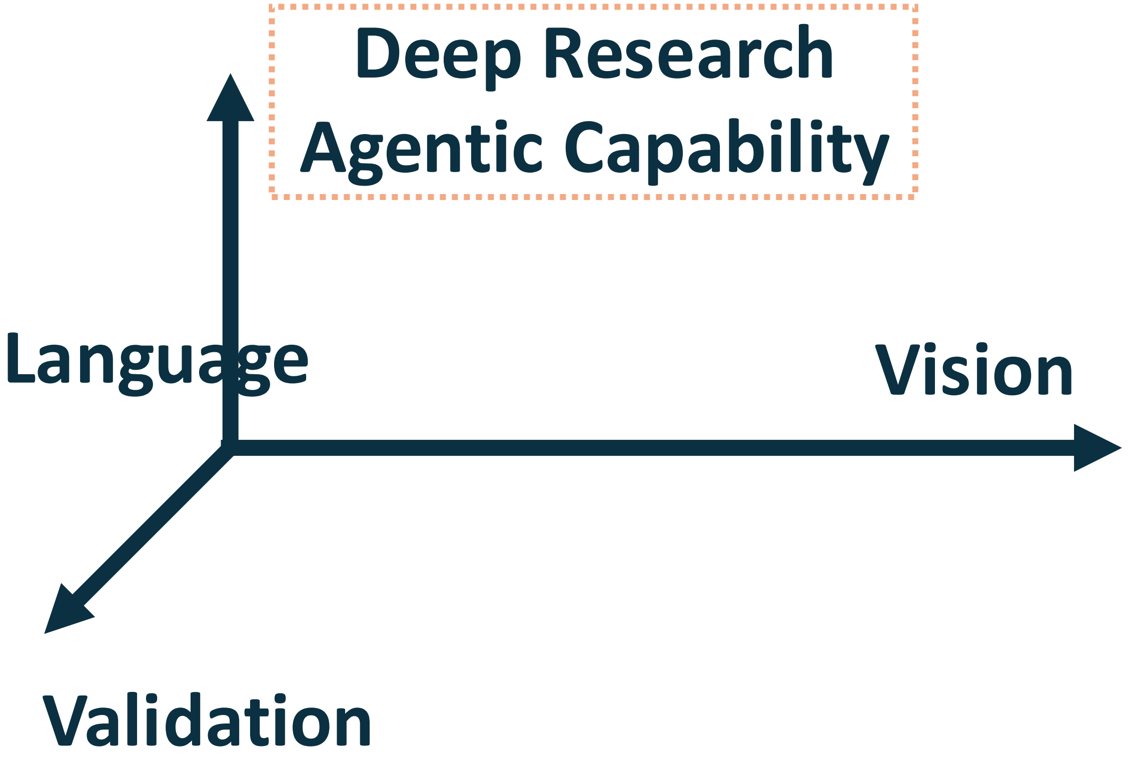
Growing Data

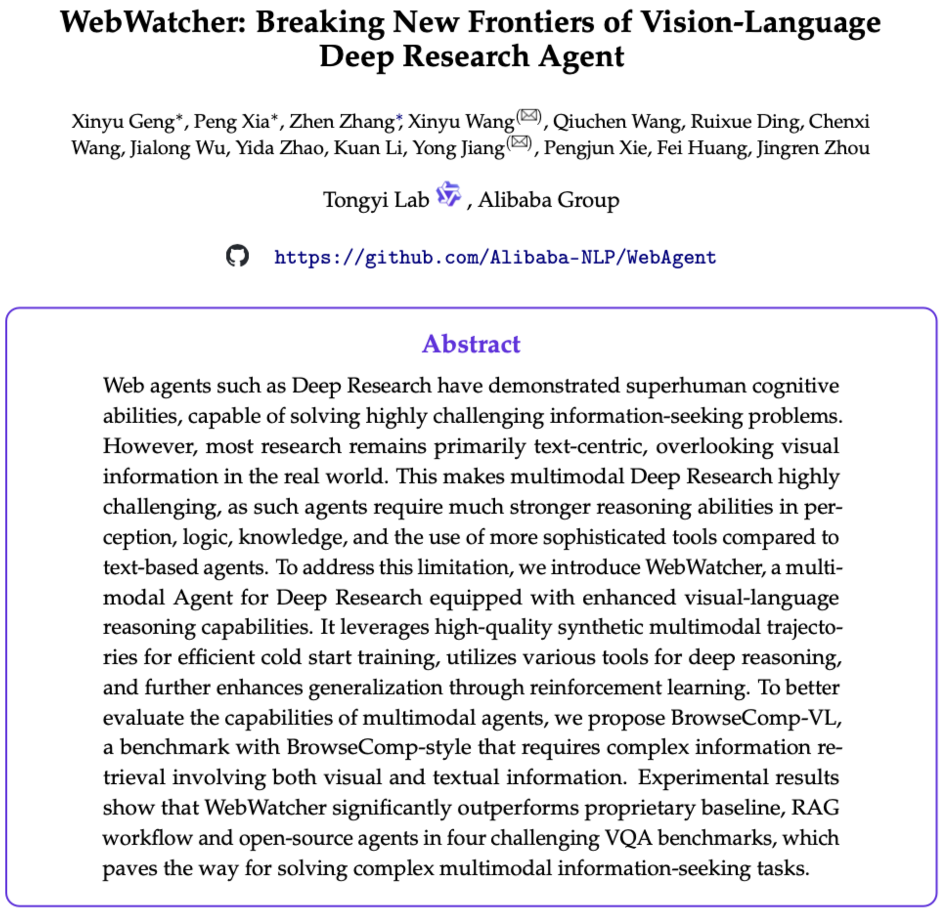
CoT Reasoning

Multi-Modal

“But what does the path to autonomous novel discovery with an AIScientist truly require? Exploration, with a visionary roadmap mapping, built on trust.”

Crucial Gaps and Paths Forward for Autonomous AI Research Scientists





WebWatcher: Breaking New Frontiers of Vision-Language Deep Research Agent

Xinyu Geng*, Peng Xia*, Zhen Zhang*, Xinyu Wang(✉), Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, Yong Jiang(✉), Pengjun Xie, Fei Huang, Jingren Zhou

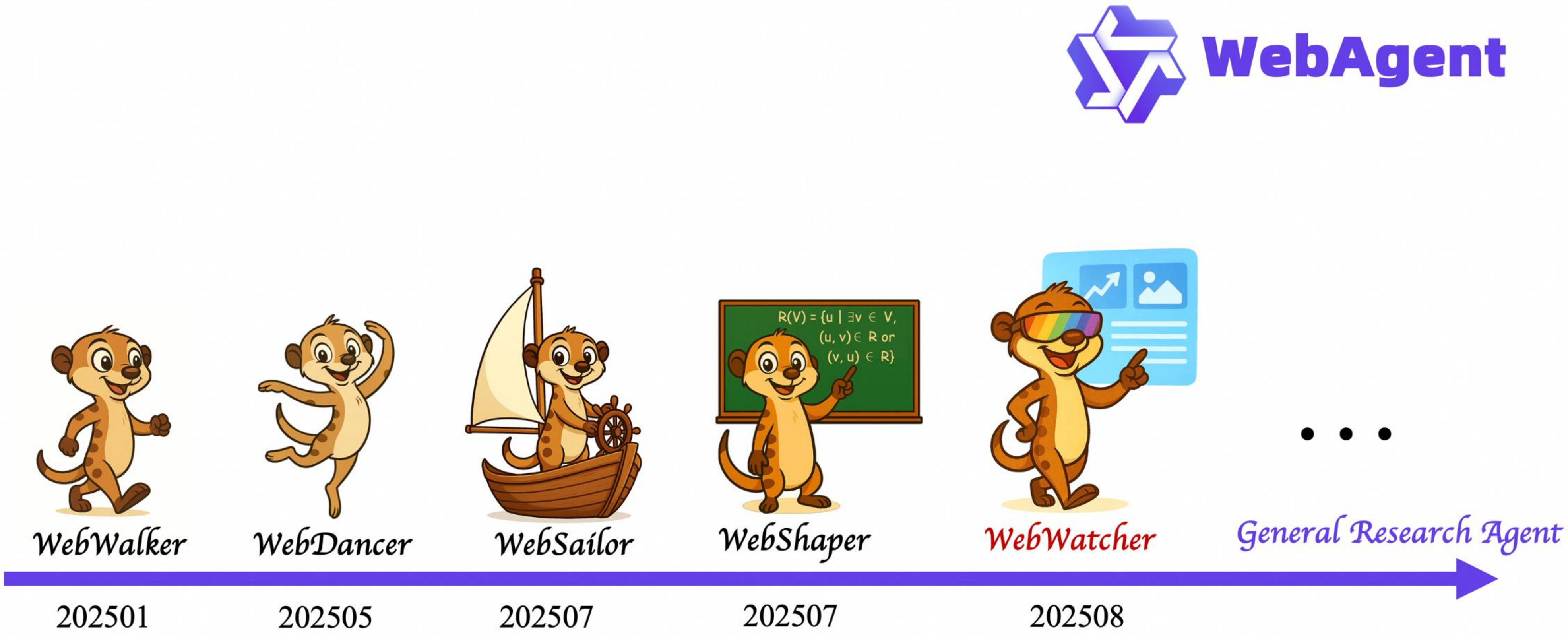
Tongyi Lab, Alibaba Group

https://github.com/Alibaba-NLP/WebAgent

Abstract

Web agents such as Deep Research have demonstrated superhuman cognitive abilities, capable of solving highly challenging information-seeking problems. However, most research remains primarily text-centric, overlooking visual information in the real world. This makes multimodal Deep Research highly challenging, as such agents require much stronger reasoning abilities in perception, logic, knowledge, and the use of more sophisticated tools compared to text-based agents. To address this limitation, we introduce WebWatcher, a multi-modal Agent for Deep Research equipped with enhanced visual-language reasoning capabilities. It leverages high-quality synthetic multimodal trajectories for efficient cold start training, utilizes various tools for deep reasoning, and further enhances generalization through reinforcement learning. To better evaluate the capabilities of multimodal agents, we propose BrowseComp-VL, a benchmark with BrowseComp-style that requires complex information retrieval involving both visual and textual information. Experimental results show that WebWatcher significantly outperforms proprietary baseline, RAG workflow and open-source agents in four challenging VQA benchmarks, which paves the way for solving complex multimodal information-seeking tasks.

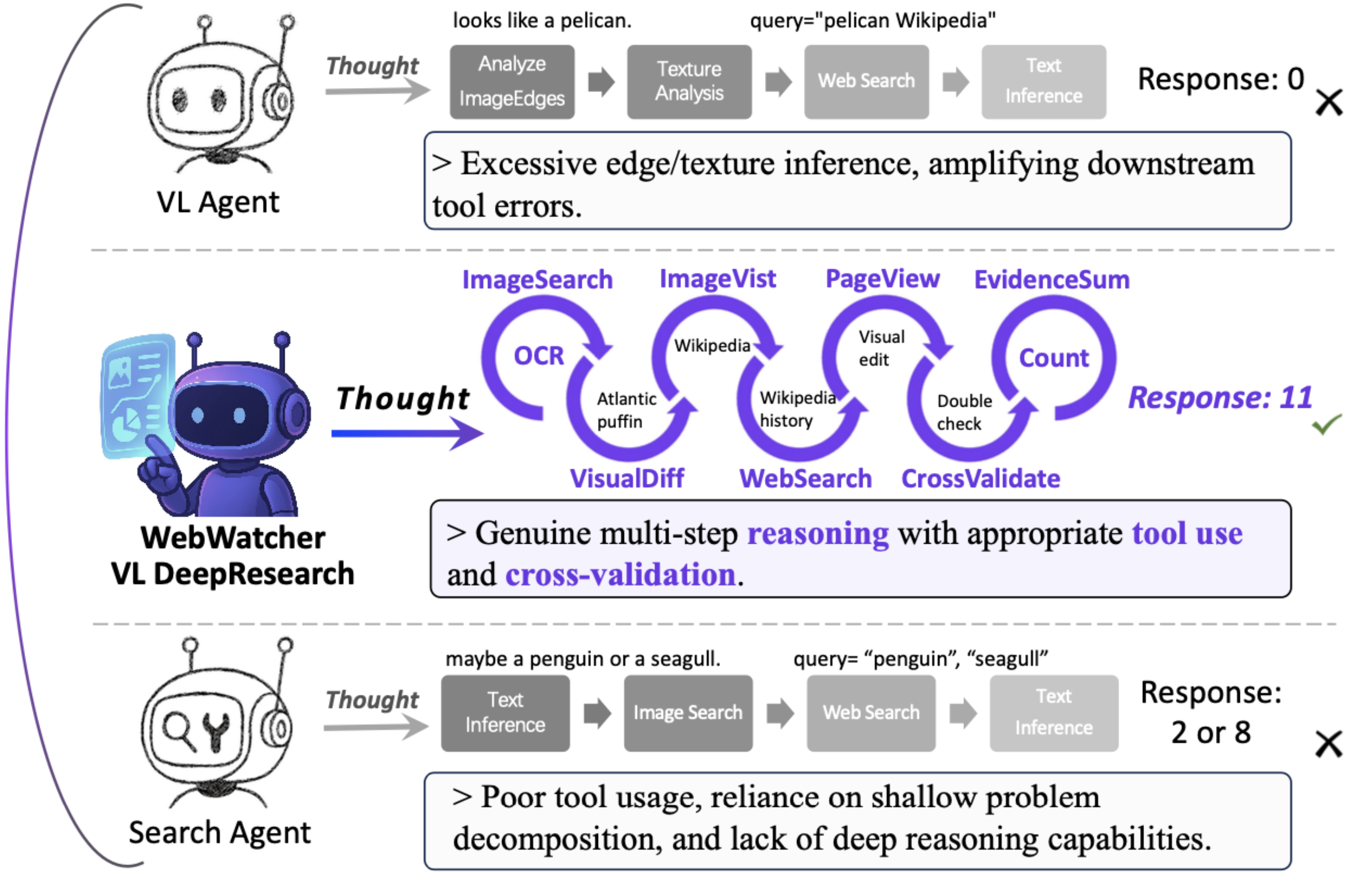
The Design of Retrieval-Augmented Web Agents



Comparison of Vision-Language Reasoning Agents

- **WebWatcher** resolves the GAIA case that defeats either vision-only reasoning or search-based agents, demonstrating the strength of multi-tool integration and in-depth reasoning generalization.

Question: On the Wikipedia page for the animal in the provided image, how many revisions from before 2020 had "visual edit" tags?
Answer: 11



> Excessive edge/texture inference, amplifying downstream tool errors.

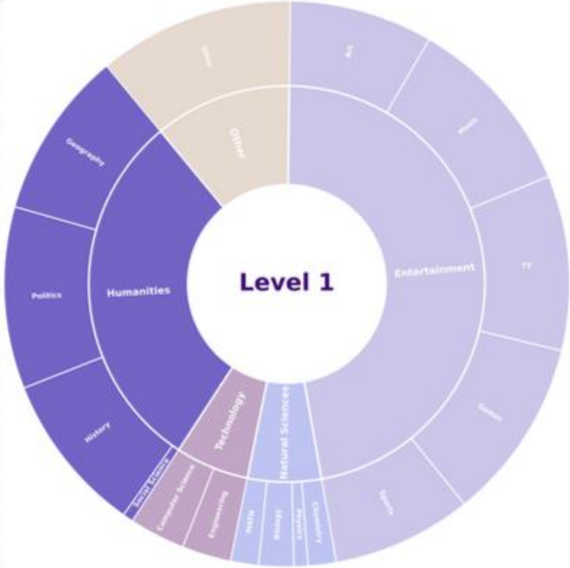
> Genuine multi-step **reasoning** with appropriate **tool use** and **cross-validation**.

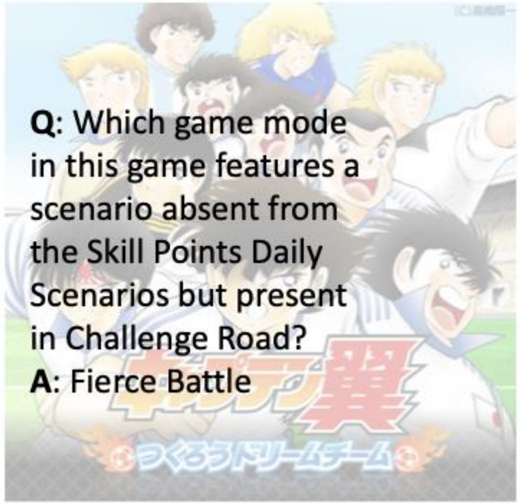
> Poor tool usage, reliance on shallow problem decomposition, and lack of deep reasoning capabilities.

Domain Distribution of Our BrowseComp-VL Dataset

❖ **Organized into 5 major domains (e.g., entertainment, humanities, technology, natural science and other) and comprising 17 fine-grained subfields**

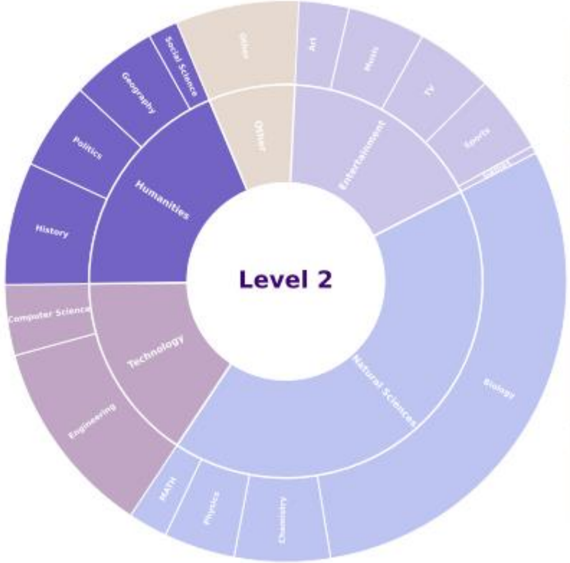
- ➢ **Lvl 1:** Questions require multi-hop reasoning but still reference explicit entities.
- ➢ **Lvl 2:** Questions are constructed with intentionally obscured or fuzzified entities and attributes.

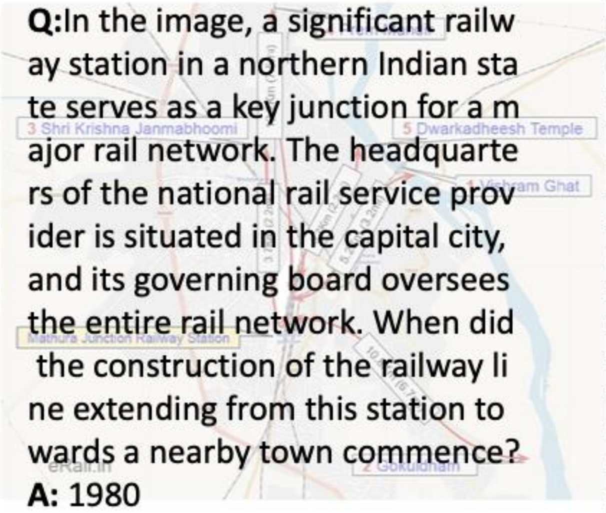




Q: Which game mode in this game features a scenario absent from the Skill Points Daily Scenarios but present in Challenge Road?

A: Fierce Battle

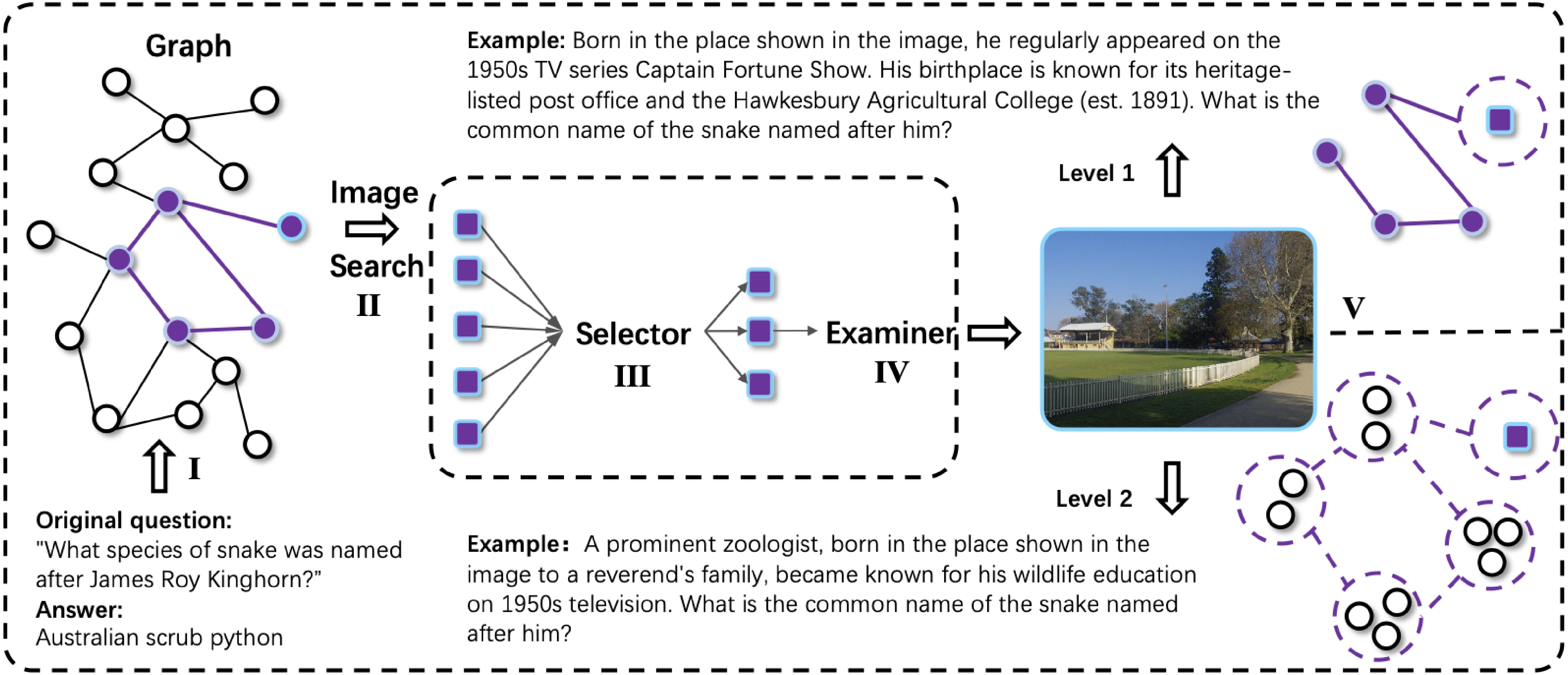




Q:In the image, a significant railway station in a northern Indian state serves as a key junction for a major rail network. The headquarters of the national rail service provider is situated in the capital city, and its governing board oversees the entire rail network. When did the construction of the railway line extending from this station towards a nearby town commence?

A: 1980

Data Generation Pipelines for WebWatcher

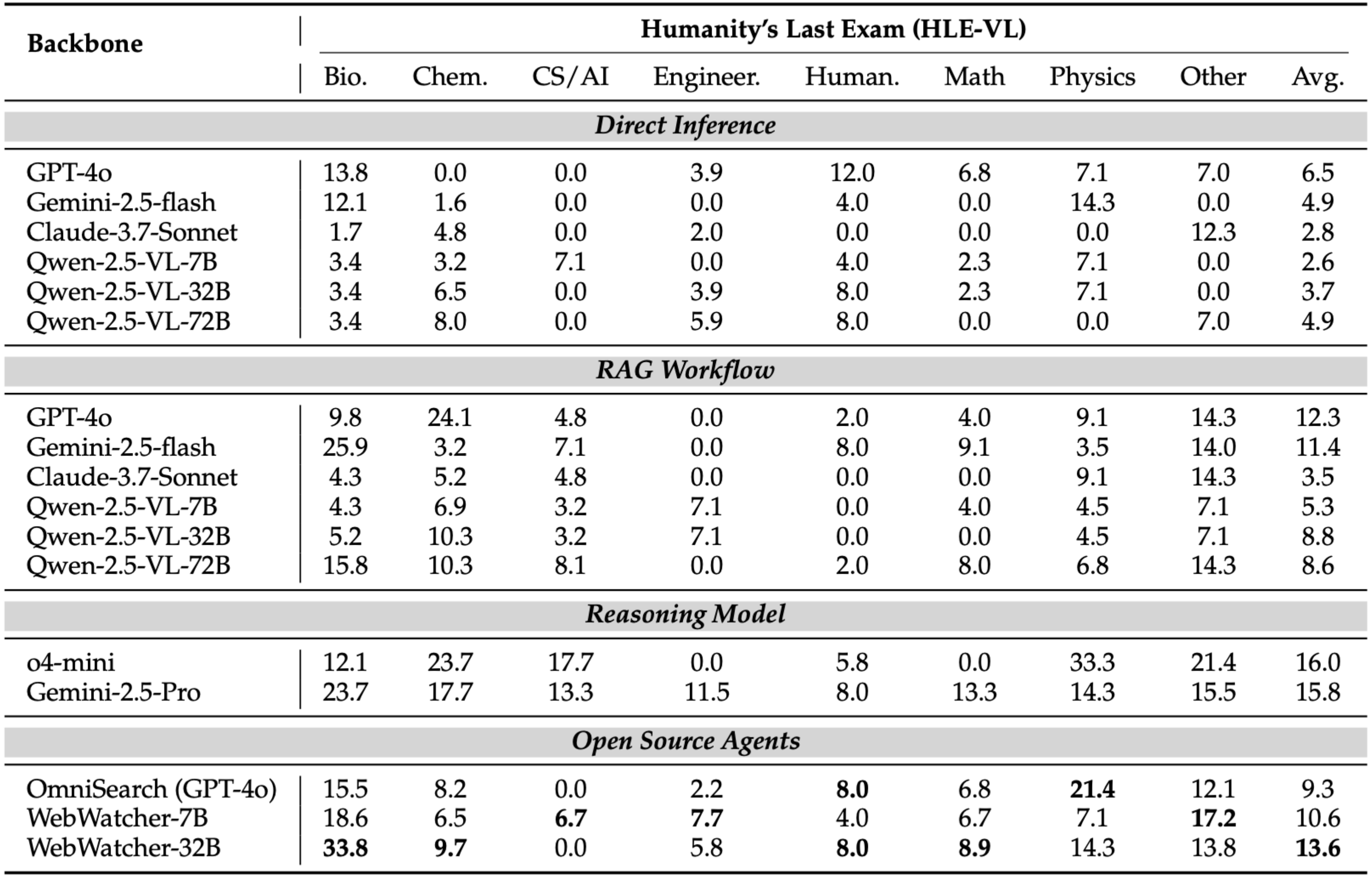


Example: Born in the place shown in the image, he regularly appeared on the 1950s TV series Captain Fortune Show. His birthplace is known for its heritage-listed post office and the Hawkesbury Agricultural College (est. 1891). What is the common name of the snake named after him?

Original question:
"What species of snake was named after James Roy Kinghorn?"
Answer:
Australian scrub python

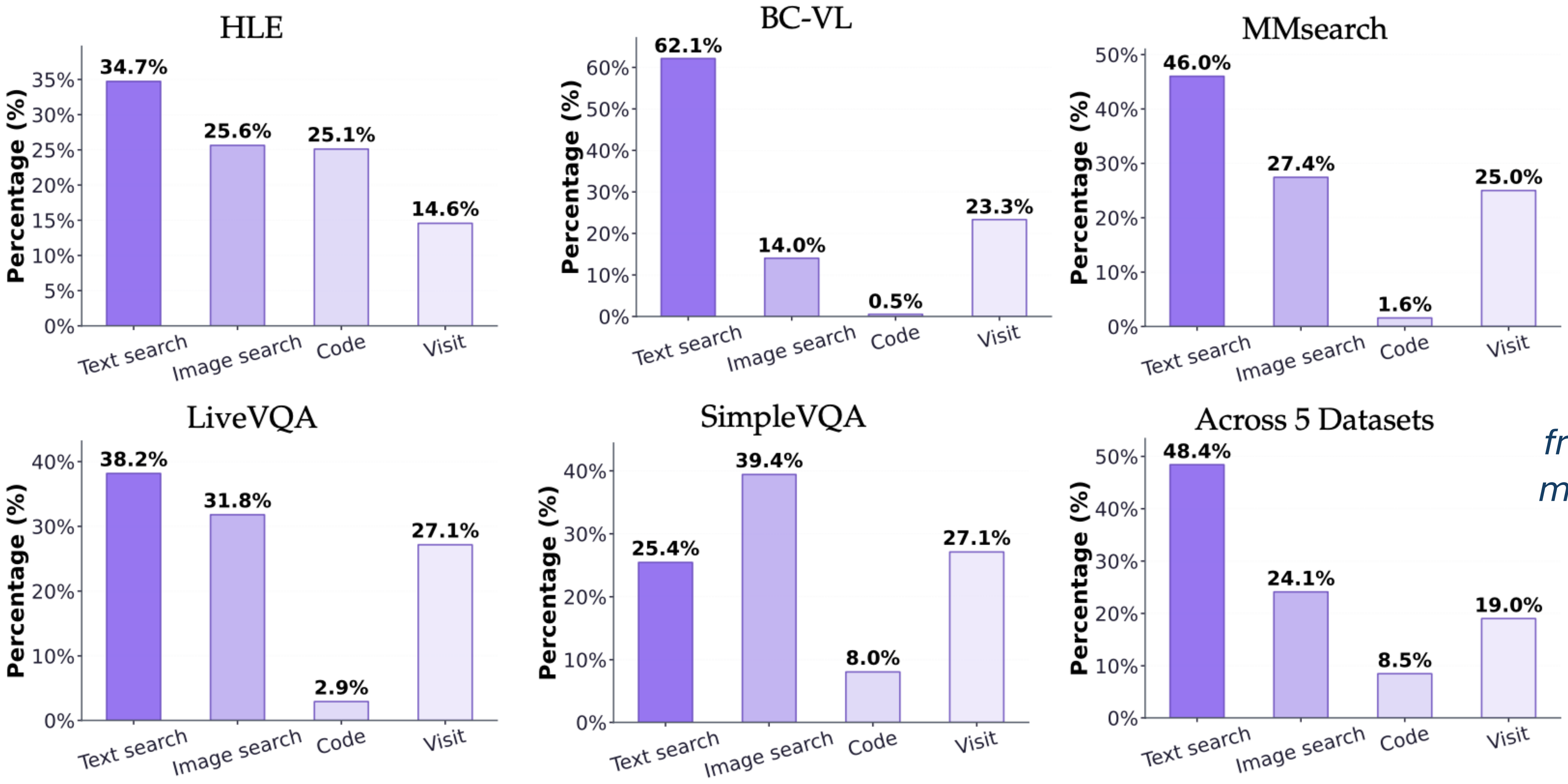
Example： A prominent zoologist, born in the place shown in the image to a reverend's family, became known for his wildlife education on 1950s television. What is the common name of the snake named after him?

Empirical Performance of WebWatcher



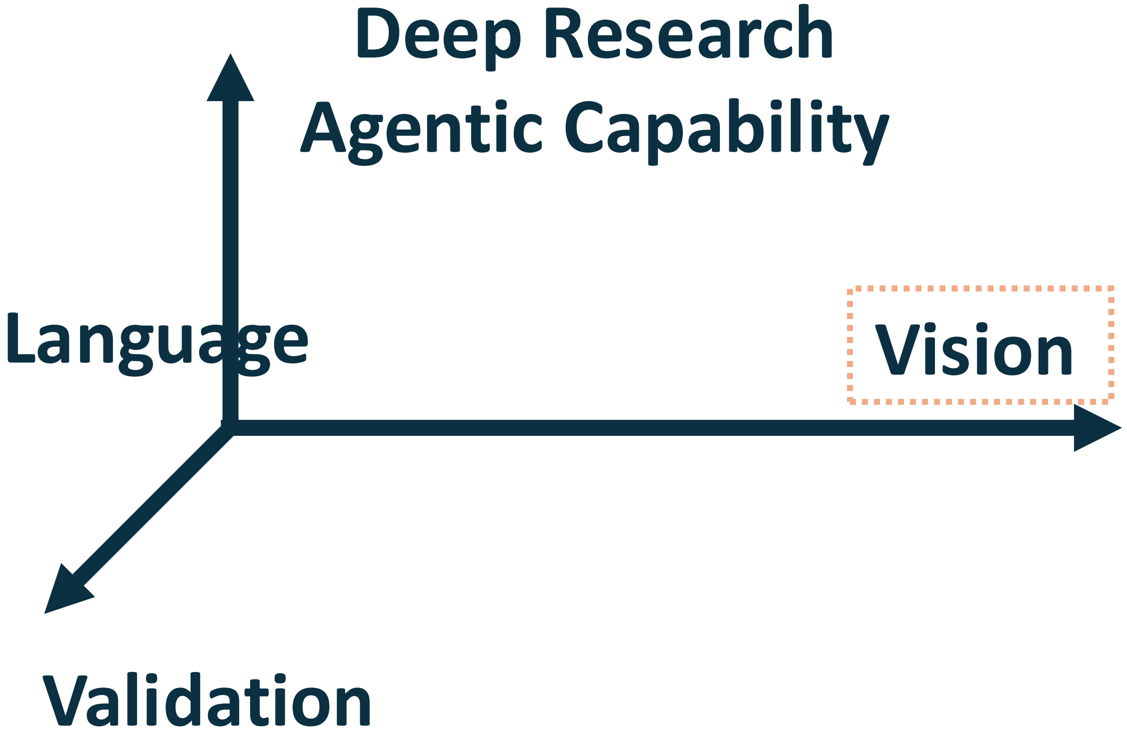
Backbone	**Humanity's Last Exam (HLE-VL)**								
	Bio.	Chem.	CS/AI	Engineer.	Human.	Math	Physics	Other	Avg.
Direct Inference									
GPT-4o	13.8	0.0	0.0	3.9	12.0	6.8	7.1	7.0	6.5
Gemini-2.5-flash	12.1	1.6	0.0	0.0	4.0	0.0	14.3	0.0	4.9
Claude-3.7-Sonnet	1.7	4.8	0.0	2.0	0.0	0.0	0.0	12.3	2.8
Qwen-2.5-VL-7B	3.4	3.2	7.1	0.0	4.0	2.3	7.1	0.0	2.6
Qwen-2.5-VL-32B	3.4	6.5	0.0	3.9	8.0	2.3	7.1	0.0	3.7
Qwen-2.5-VL-72B	3.4	8.0	0.0	5.9	8.0	0.0	0.0	7.0	4.9
RAG Workflow									
GPT-4o	9.8	24.1	4.8	0.0	2.0	4.0	9.1	14.3	12.3
Gemini-2.5-flash	25.9	3.2	7.1	0.0	8.0	9.1	3.5	14.0	11.4
Claude-3.7-Sonnet	4.3	5.2	4.8	0.0	0.0	0.0	9.1	14.3	3.5
Qwen-2.5-VL-7B	4.3	6.9	3.2	7.1	0.0	4.0	4.5	7.1	5.3
Qwen-2.5-VL-32B	5.2	10.3	3.2	7.1	0.0	0.0	4.5	7.1	8.8
Qwen-2.5-VL-72B	15.8	10.3	8.1	0.0	2.0	8.0	6.8	14.3	8.6
Reasoning Model									
o4-mini	12.1	23.7	17.7	0.0	5.8	0.0	33.3	21.4	16.0
Gemini-2.5-Pro	23.7	17.7	13.3	11.5	8.0	13.3	14.3	15.5	15.8
Open Source Agents									
OmniSearch (GPT-4o)	15.5	8.2	0.0	2.2	**8.0**	6.8	**21.4**	12.1	9.3
WebWatcher-7B	18.6	6.5	**6.7**	**7.7**	4.0	6.7	7.1	**17.2**	10.6
WebWatcher-32B	**33.8**	**9.7**	0.0	5.8	**8.0**	**8.9**	14.3	13.8	**13.6**

Distribution of Tool Calls by WebWatcher in VL Agent Benchmarks



Y-axis reflects fraction of total calls made to tool in x-axis

Crucial Gaps and Paths Forward for Autonomous AI Research Scientists



Deep Research Agentic Capability

Language

Vision

Validation

Exploring Agentic Multimodal Large Language Models: A Survey for AIScientists

Jinglin Jian[†] Yi R. Fung[†] Denghui Zhang[◇] Yiqian Liang[§]
Qingyu Chen[♡] Zhiyong Lu[¶] Qingyun Wang[*]
[†]University of Illinois at Urbana-Champaign [§]University of Cambridge
[♡]Yale University [◇]Stevens Institute of Technology [*]William & Mary
[¶]National Library of Medicine, National Institutes of Health
{jj50,yifung2}@illinois.edu, dzhang42@stevens.edu, yl992@cam.ac.uk
qingyu.chen@yale.edu, zhiyong.lu@nih.gov, qwang16@wm.edu

Abstract

The emergence of agentic Multimodal Large Language Models (MLLMs) has catalyzed a new paradigm in scientific discovery, enabling systems to autonomously understand, reason, and act across diverse modalities. Agentic MLLMs are emerging as the next frontier for AIScientists, systems capable of assisting or even independently conducting every stage in the scientific research cycle. This paper presents a systematic taxonomy and framework for scientific MLLM agents development, covering multimodal perception, training and inference methodologies, evaluation benchmarks, and human–AIScientist collaboration. We further outline persistent challenges such as data scarcity, reliability, and interpretability, emphasizing the importance of transparent and controllable interaction frameworks. By framing agentic MLLMs as collaborative partners that augment rather than replace human scientists, we advocate for a balanced vision of AIScientists, one that advances discovery while promoting democracy and inclusivity in scientific progress.

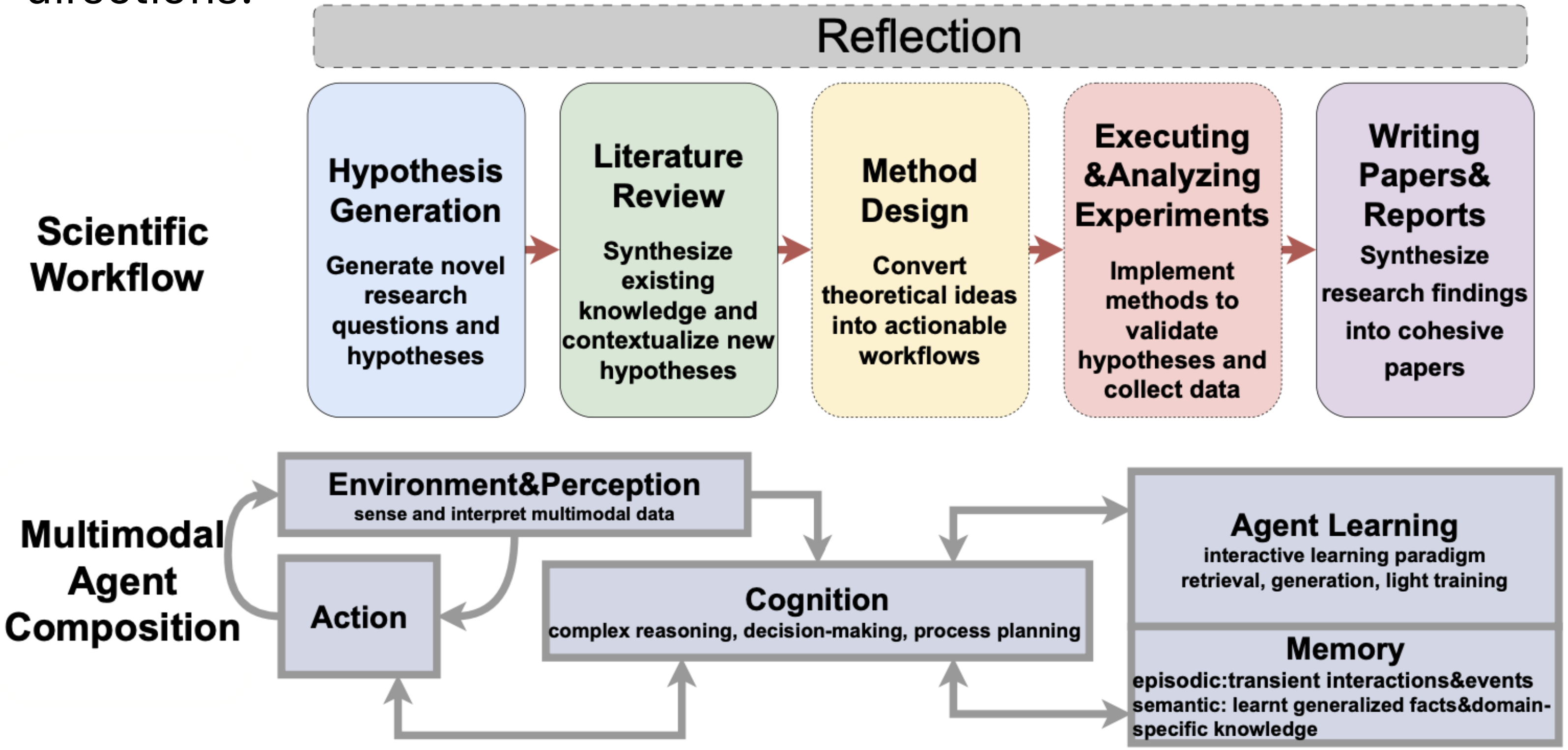
Prize in Chemistry[1] has also been awarded to researchers who utilized AI in pioneering new pathways for drug discovery.

The concept of the "AIScientist," first introduced by Lu et al. (2024a), envisions a framework for fully automated scientific discovery. Powered by agentic multimodal large language models (MLLMs), these systems aim to support the entire research lifecycle: understanding papers (Radensky et al., 2024; Wang et al., 2024f), generating new hypotheses (Wang et al., 2024a; Li et al., 2024g; Baek et al., 2024a), planning and conducting experiments (Boiko et al., 2023; Tom et al., 2024), analyzing results, drafting manuscripts (Wang et al., 2024e), and reviewing papers (Du et al., 2024; Weng et al., 2024).

To fully realize the potential of agentic MLLMs, it is essential to explore innovative methodologies tailored to their training and deployment within scientific domains. Challenges such as cross-disciplinary, limited data availability in scientific fields, and the prevalence of long-tail distributions (Wang et al., 2024c) necessitate the development of advanced modeling techniques. These include contrastive

Preliminaries of the Modern Scientific Discovery Workflow

- ❖ To support the end-to-end research lifecycle, agentic MLLMs employ a closed-loop system of **Environment & Perception**, **Cognition**, and **Action**, bolstered by **Episodic** and **Semantic Memory** for continuous learning.
- ❖ This workflow drives the discovery process from **Hypothesis Generation** through **Experiment Execution**, utilizing an iterative **Reflection** layer to synthesize findings into cohesive reports and refine future research directions.



Reflection

Scientific Workflow

Hypothesis Generation
Generate novel research questions and hypotheses

Literature Review
Synthesize existing knowledge and contextualize new hypotheses

Method Design
Convert theoretical ideas into actionable workflows

Executing &Analyzing Experiments
Implement methods to validate hypotheses and collect data

Writing Papers& Reports
Synthesize research findings into cohesive papers

Multimodal Agent Composition

Environment&Perception
sense and interpret multimodal data

Action

Cognition
complex reasoning, decision-making, process planning

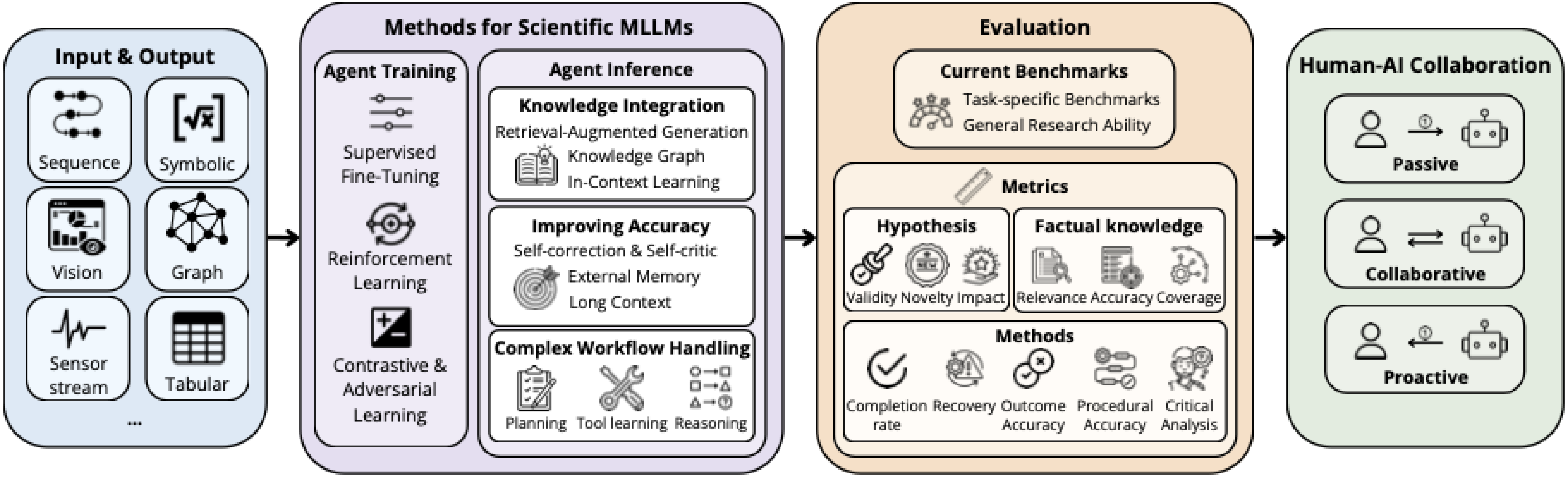
Agent Learning
interactive learning paradigm
retrieval, generation, light training

Memory
episodic:transient interactions&events
semantic: learnt generalized facts&domain-specific knowledge

Overview of the Agentic MLLM Framework for Scientific Discovery

❖ To handle the complexity of modern discovery, scientific MLLMs process diverse modalities including **symbolic equations**, **molecular graphs**, and **sensor streams**, utilizing specialized **reinforcement** and **contrastive learning** to manage the long-tail distributions inherent in scientific data.

❖ High-fidelity results are further enabled through advanced **workflow handling** and **knowledge integration**, with success measured against robust metrics such as hypothesis **novelty**, **impact**, and **procedural accuracy**.

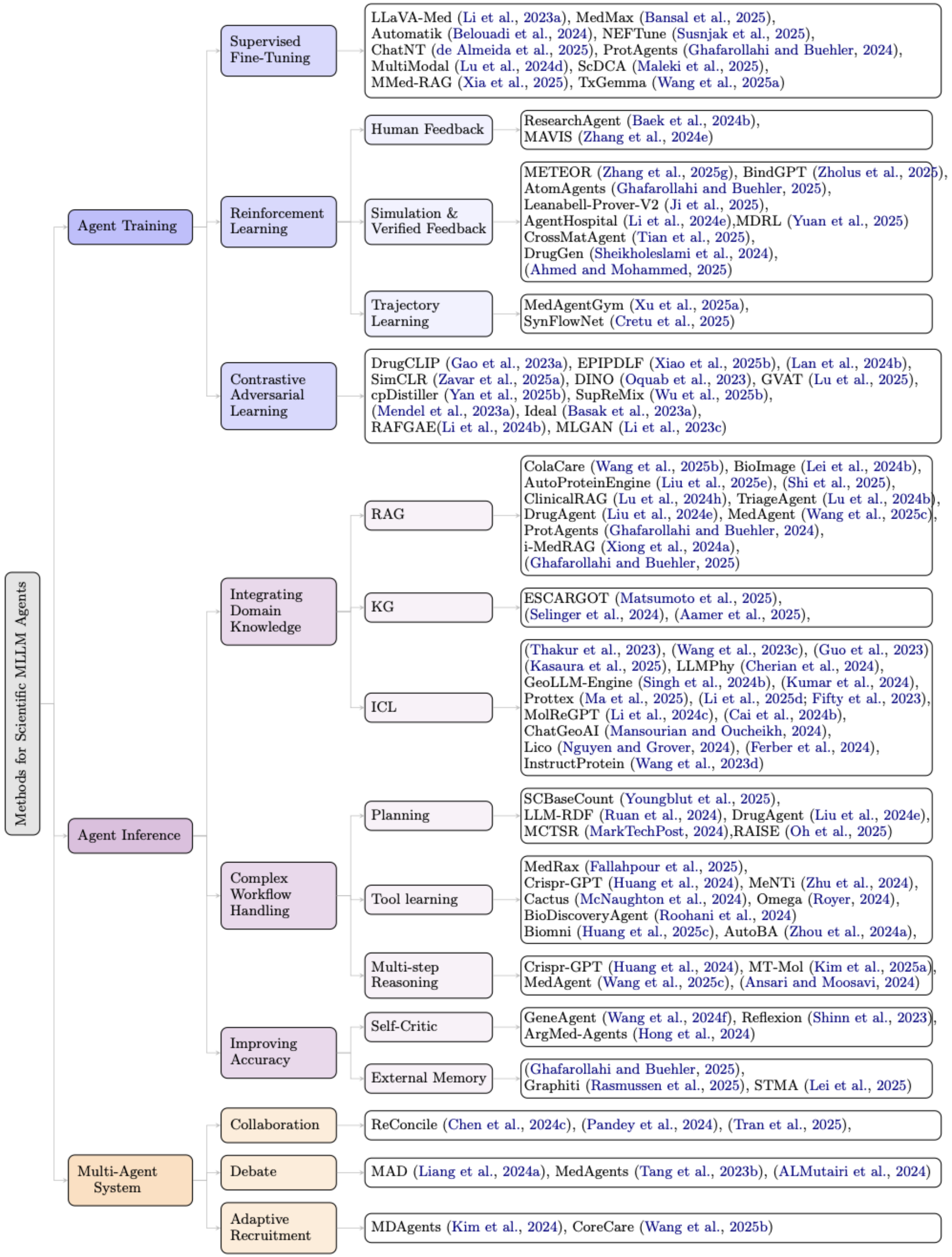


Taxonomy of Methods for Scientific MLLM Agents

❖ Current SOTA methods (e.g., ProtAgents, MedAgent) utilize supervised finetuning training tailored for domain specific research.

❖ To ground reasoning and enable autonomous discovery, agentic MLLMs integrate domain knowledge through techniques like **RAG** and **In-Context Learning** while leveraging specialized frameworks for **complex reasoning** and **autonomous tool use**.

❖Future-facing discovery utilizes multi-agent systems that employ debate (MAD), collaboration (ReConcile), and adaptive recruitment (MDAgents) to improve accuracy and reduce bias.



Methods for Scientific MLLM Agents

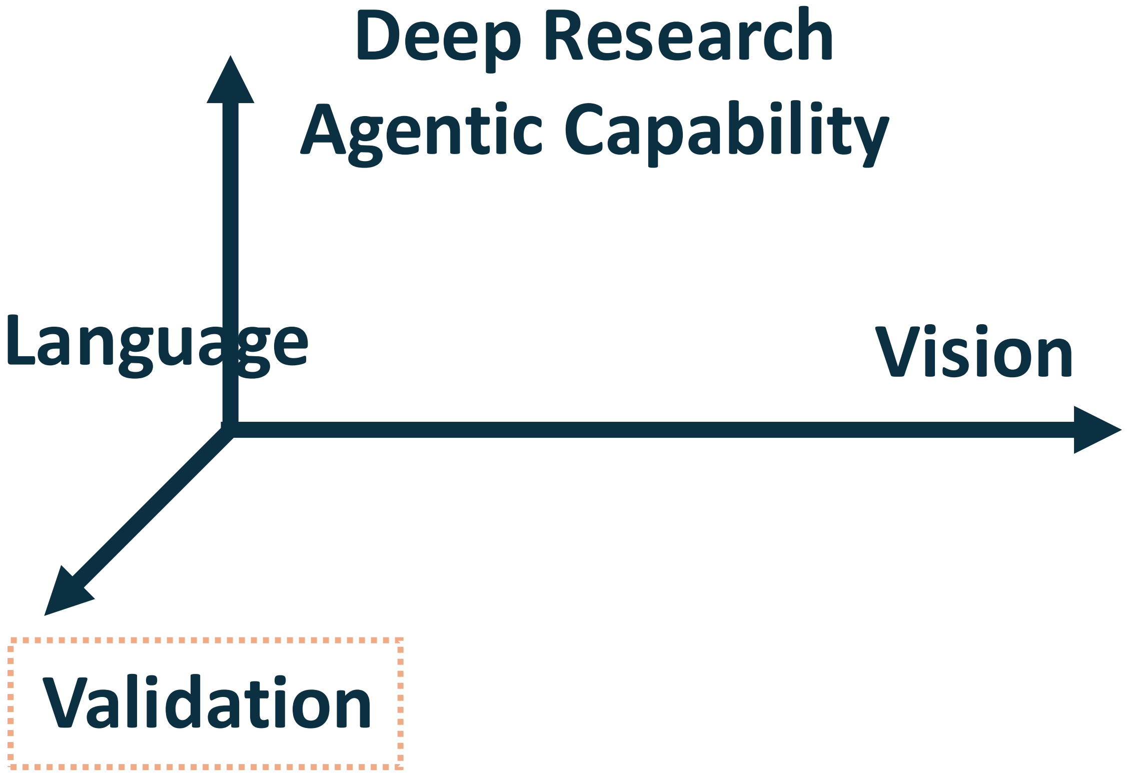
- Agent Training
  - Supervised Fine-Tuning
    - LLaVA-Med (Li et al., 2023a), MedMax (Bansal et al., 2025), Automatik (Belouadi et al., 2024), NEFTune (Susnjak et al., 2025), ChatNT (de Almeida et al., 2025), ProtAgents (Ghafarollahi and Buehler, 2024), MultiModal (Lu et al., 2024d), ScDCA (Maleki et al., 2025), MMed-RAG (Xia et al., 2025), TxGemma (Wang et al., 2025a)
  - Reinforcement Learning
    - Human Feedback
      - ResearchAgent (Baek et al., 2024b), MAVIS (Zhang et al., 2024e)
    - Simulation & Verified Feedback
      - METEOR (Zhang et al., 2025g), BindGPT (Zholus et al., 2025), AtomAgents (Ghafarollahi and Buehler, 2025), Leanabell-Prover-V2 (Ji et al., 2025), AgentHospital (Li et al., 2024e),MDRL (Yuan et al., 2025) CrossMatAgent (Tian et al., 2025), DrugGen (Sheikholeslami et al., 2024), (Ahmed and Mohammed, 2025)
    - Trajectory Learning
      - MedAgentGym (Xu et al., 2025a), SynFlowNet (Cretu et al., 2025)
  - Contrastive Adversarial Learning
    - DrugCLIP (Gao et al., 2023a), EPIPDLF (Xiao et al., 2025b), (Lan et al., 2024b), SimCLR (Zavar et al., 2025a), DINO (Oquab et al., 2023), GVAT (Lu et al., 2025), cpDistiller (Yan et al., 2025b), SupReMix (Wu et al., 2025b), (Mendel et al., 2023a), Ideal (Basak et al., 2023a), RAFGAE(Li et al., 2024b), MLGAN (Li et al., 2023c)
- Agent Inference
  - Integrating Domain Knowledge
    - RAG
      - ColaCare (Wang et al., 2025b), BioImage (Lei et al., 2024b), AutoProteinEngine (Liu et al., 2025e), (Shi et al., 2025), ClinicalRAG (Lu et al., 2024h), TriageAgent (Lu et al., 2024b), DrugAgent (Liu et al., 2024e), MedAgent (Wang et al., 2025c), ProtAgents (Ghafarollahi and Buehler, 2024), i-MedRAG (Xiong et al., 2024a), (Ghafarollahi and Buehler, 2025)
    - KG
      - ESCARGOT (Matsumoto et al., 2025), (Selinger et al., 2024), (Aamer et al., 2025),
    - ICL
      - (Thakur et al., 2023), (Wang et al., 2023c), (Guo et al., 2023) (Kasaura et al., 2025), LLMPhy (Cherian et al., 2024), GeoLLM-Engine (Singh et al., 2024b), (Kumar et al., 2024), Prottex (Ma et al., 2025), (Li et al., 2025d; Fifty et al., 2023), MolReGPT (Li et al., 2024c), (Cai et al., 2024b), ChatGeoAI (Mansourian and Oucheikh, 2024), Lico (Nguyen and Grover, 2024), (Ferber et al., 2024), InstructProtein (Wang et al., 2023d)
  - Complex Workflow Handling
    - Planning
      - SCBaseCount (Youngblut et al., 2025), LLM-RDF (Ruan et al., 2024), DrugAgent (Liu et al., 2024e), MCTSR (MarkTechPost, 2024),RAISE (Oh et al., 2025)
    - Tool learning
      - MedRax (Fallahpour et al., 2025), Crispr-GPT (Huang et al., 2024), MeNTi (Zhu et al., 2024), Cactus (McNaughton et al., 2024), Omega (Royer, 2024), BioDiscoveryAgent (Roohani et al., 2024) Biomni (Huang et al., 2025c), AutoBA (Zhou et al., 2024a),
    - Multi-step Reasoning
      - Crispr-GPT (Huang et al., 2024), MT-Mol (Kim et al., 2025a), MedAgent (Wang et al., 2025c), (Ansari and Moosavi, 2024)
  - Improving Accuracy
    - Self-Critic
      - GeneAgent (Wang et al., 2024f), Reflexion (Shinn et al., 2023), ArgMed-Agents (Hong et al., 2024)
    - External Memory
      - (Ghafarollahi and Buehler, 2025), Graphiti (Rasmussen et al., 2025), STMA (Lei et al., 2025)
- Multi-Agent System
  - Collaboration
    - ReConcile (Chen et al., 2024c), (Pandey et al., 2024), (Tran et al., 2025),
  - Debate
    - MAD (Liang et al., 2024a), MedAgents (Tang et al., 2023b), (ALMutairi et al., 2024)
  - Adaptive Recruitment
    - MDAgents (Kim et al., 2024), CoreCare (Wang et al., 2025b)

Ongoing Challenges and Future Directions

- ❖ **Addressing Data Scarcity:** A primary hurdle remains the scarcity of high-quality multimodal scientific data and the prevalence of **long-tail distributions** that challenge standard model generalization.
- ❖ **Reliability and Interpretability:** For AI to be a trusted partner, we must develop **transparent and controllable interaction frameworks** that allow human scientists to audit the AI's reasoning.
- ❖ **Human-AI Collaboration:** The vision for AIScientists is a **Collaborative Partner** model, shifting from passive tools to proactive agents that augment, rather than replace, human expertise.
- ❖**Democratizing Discovery**: By reducing the technical barriers to complex experimentation, agentic MLLMs can promote **inclusivity and democracy** in the global scientific progress.

Jian, J., Fung, Y. R., Zhang, D., Liang, Y., Chen, Q., Lu, Z., & Wang, Q. (2025). Exploring Agentic Multimodal Large Language Models: A Survey for AIScientists. *Authorea Preprints*.

Crucial Gaps and Paths Forward for Autonomous AI Research Scientists

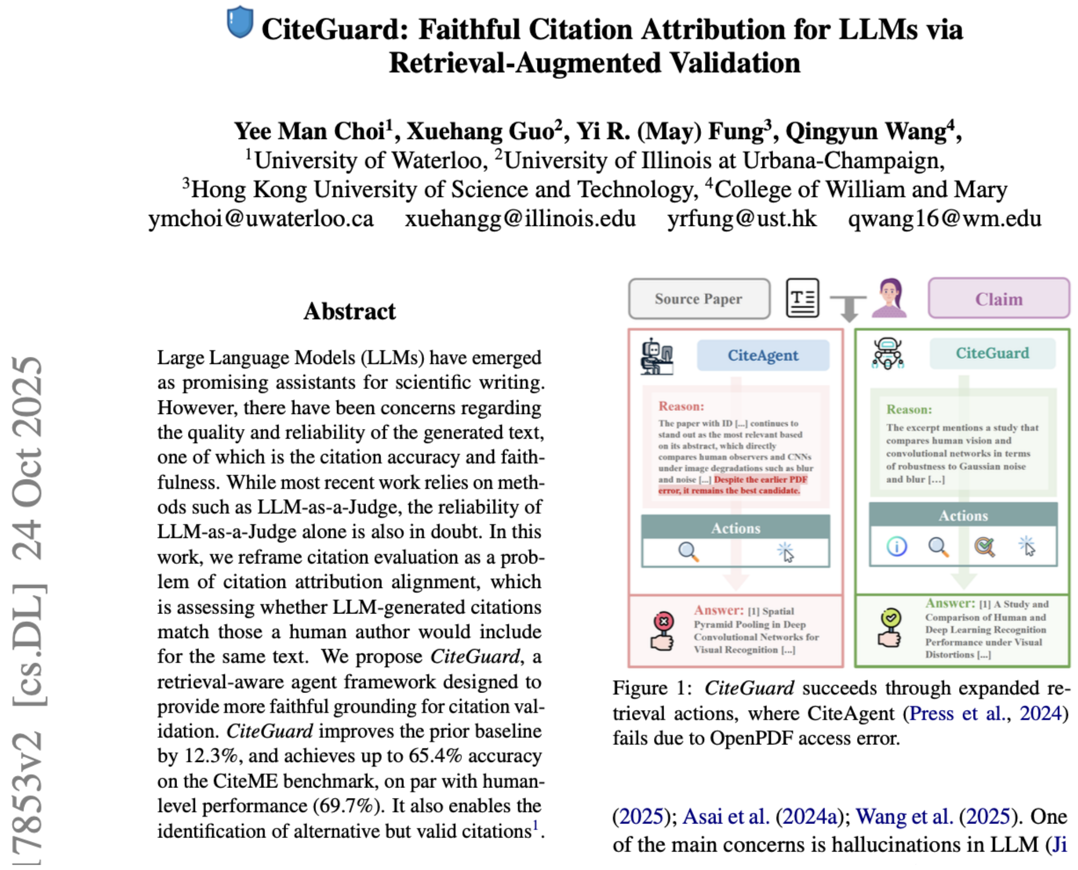


Deep Research Agentic Capability

Language

Vision

Validation



CiteGuard: Faithful Citation Attribution for LLMs via Retrieval-Augmented Validation

Yee Man Choi[1], Xuehang Guo[2], Yi R. (May) Fung[3], Qingyun Wang[4],
[1]University of Waterloo, [2]University of Illinois at Urbana-Champaign,
[3]Hong Kong University of Science and Technology, [4]College of William and Mary
ymchoi@uwaterloo.ca xuehangg@illinois.edu yrfung@ust.hk qwang16@wm.edu

Abstract

Large Language Models (LLMs) have emerged as promising assistants for scientific writing. However, there have been concerns regarding the quality and reliability of the generated text, one of which is the citation accuracy and faithfulness. While most recent work relies on methods such as LLM-as-a-Judge, the reliability of LLM-as-a-Judge alone is also in doubt. In this work, we reframe citation evaluation as a problem of citation attribution alignment, which is assessing whether LLM-generated citations match those a human author would include for the same text. We propose *CiteGuard*, a retrieval-aware agent framework designed to provide more faithful grounding for citation validation. *CiteGuard* improves the prior baseline by 12.3%, and achieves up to 65.4% accuracy on the CiteME benchmark, on par with human-level performance (69.7%). It also enables the identification of alternative but valid citations[1].

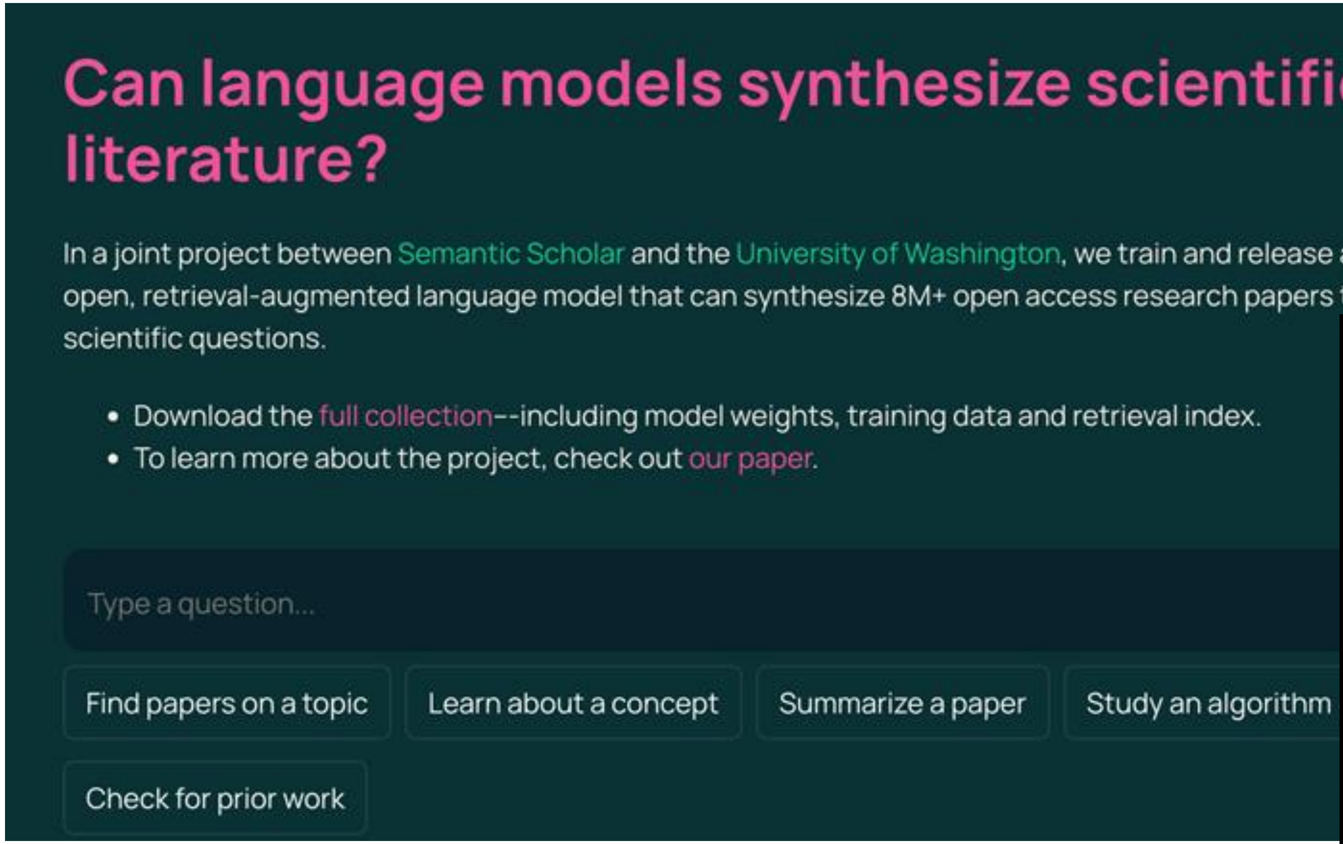
Source Paper
Claim
CiteAgent
Reason:
The paper with ID [...] continues to stand out as the most relevant based on its abstract, which directly compares human observers and CNNs under image degradations such as blur and noise [...] Despite the earlier PDF error, it remains the best candidate.
Actions
Answer: [1] Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition [...]
CiteGuard
Reason:
The excerpt mentions a study that compares human vision and convolutional networks in terms of robustness to Gaussian noise and blur [...]
Actions
Answer: [1] A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions [...]

Figure 1: *CiteGuard* succeeds through expanded retrieval actions, where CiteAgent (Press et al., 2024) fails due to OpenPDF access error.

(2025); Asai et al. (2024a); Wang et al. (2025). One of the main concerns is hallucinations in LLM (Ji

AI4Research

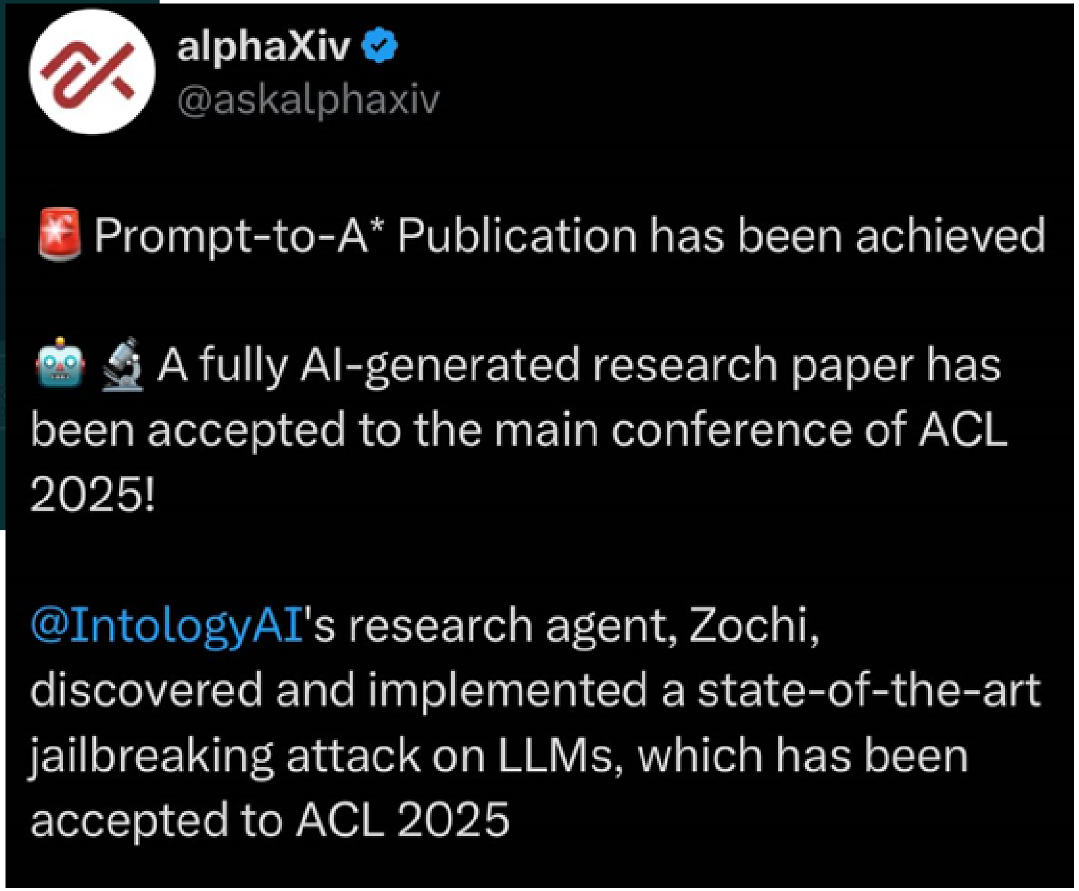
Edison



Can language models synthesize scientific literature?

In a joint project between Semantic Scholar and the University of Washington, we train and release a fully open, retrieval-augmented language model that can synthesize 8M+ open access research papers to answer scientific questions.

- Download the full collection--including model weights, training data and retrieval index.
- To learn more about the project, check out our paper.



alphaXiv
@askalphaxiv

🚨 Prompt-to-A* Publication has been achieved

🤖 🔬 A fully AI-generated research paper has been accepted to the main conference of ACL 2025!

@IntologyAI's research agent, Zochi, discovered and implemented a state-of-the-art jailbreaking attack on LLMs, which has been accepted to ACL 2025

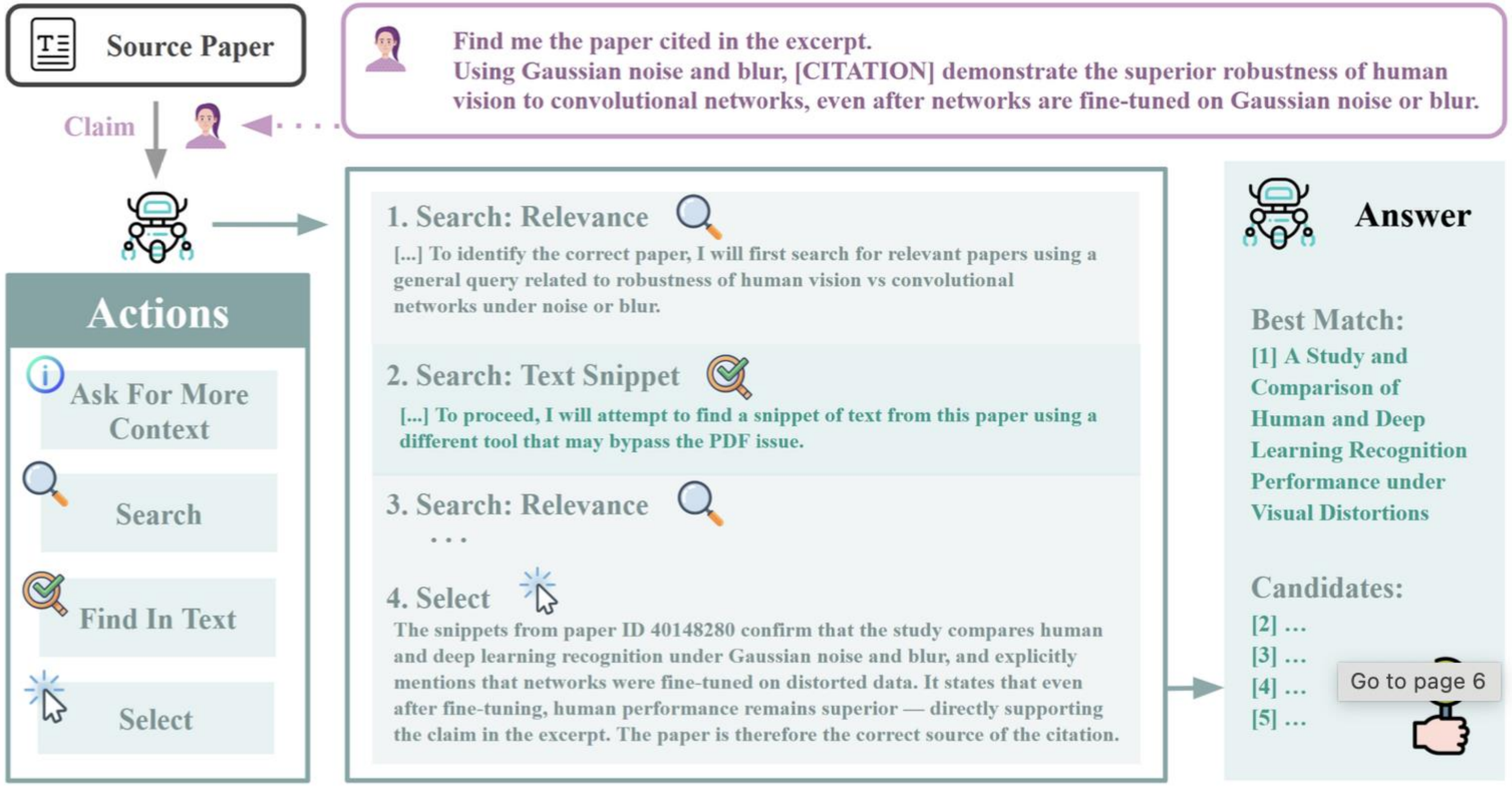
But there are still some concerns:

Let’s look at this from the scope of **generated related work section**:

- How reliable is the generated text?
  - Is using LLM-as-a-Judge good enough?
  - Ye et.al, Justice or Prejudice? Quantifying **Biases** in LLM-as-a-Judge

- How close is the generated text to human written ones?
  - Do we have comparative analysis like human would have? Or simply just **a summary per paper** in each sentence?
  - Li et.al, ChatCite: LLM Agent with Human Workflow Guidance for Comparative Literature Summary
    - Comparative Analysis Score (LLM-as-a-Judge)

Our Proposed Framework, CiteGuard, Leverages Expanded Toolset for Agentic Validation of Scientific Text



Find me the paper cited in the excerpt.
Using Gaussian noise and blur, [CITATION] demonstrate the superior robustness of human vision to convolutional networks, even after networks are fine-tuned on Gaussian noise or blur.

1. Search: Relevance

[...] To identify the correct paper, I will first search for relevant papers using a general query related to robustness of human vision vs convolutional networks under noise or blur.

2. Search: Text Snippet

[...] To proceed, I will attempt to find a snippet of text from this paper using a different tool that may bypass the PDF issue.

3. Search: Relevance

. . .

4. Select

The snippets from paper ID 40148280 confirm that the study compares human and deep learning recognition under Gaussian noise and blur, and explicitly mentions that networks were fine-tuned on distorted data. It states that even after fine-tuning, human performance remains superior — directly supporting the claim in the excerpt. The paper is therefore the correct source of the citation.

Answer

Best Match:
[1] A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions

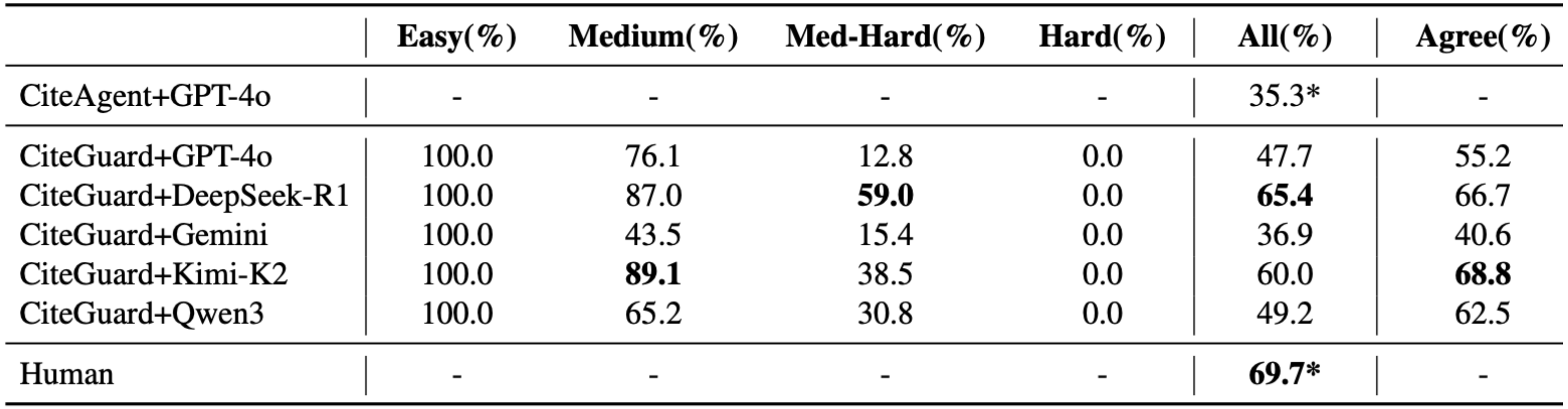
Candidates:
[2] …
[3] …
[4] …
[5] …

Go to page 6

Choi, Y. M., Guo, X., Fung, Y. R., & Wang, Q. (2025). CiteGuard: Faithful Citation Attribution for LLMs via Retrieval-Augmented Validation. *arXiv preprint arXiv:2510.17853*.

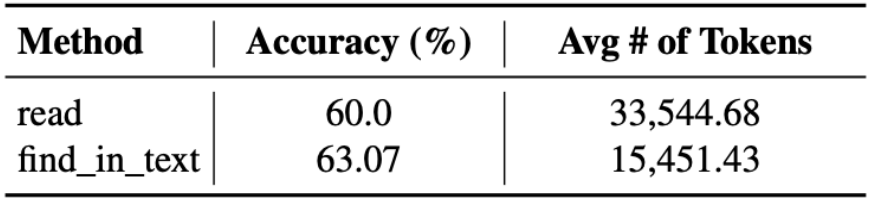
Experimental Results and Discussion

- **CiteGuard** demonstrates that retrieval-augmented agents can achieve citation attribution fidelity on par with human authors (65.4% vs 69.7%).



	Easy(%)	**Medium(%)**	**Med-Hard(%)**	**Hard(%)**	**All(%)**	**Agree(%)**
CiteAgent+GPT-4o	-	-	-	-	35.3*	-
CiteGuard+GPT-4o	100.0	76.1	12.8	0.0	47.7	55.2
CiteGuard+DeepSeek-R1	100.0	87.0	**59.0**	0.0	**65.4**	66.7
CiteGuard+Gemini	100.0	43.5	15.4	0.0	36.9	40.6
CiteGuard+Kimi-K2	100.0	**89.1**	38.5	0.0	60.0	**68.8**
CiteGuard+Qwen3	100.0	65.2	30.8	0.0	49.2	62.5
Human	-	-	-	-	**69.7***	-

- **Tool Efficiency Analysis in Precision vs. Cost:** The *find_in_text* action is not only more accurate (**63.07%**) than the standard read action (**60.0%**) but also substantially more token-efficient, reducing consumption by over **50%** (15k vs. 33k tokens).



Method	**Accuracy (%)**	**Avg # of Tokens**
read	60.0	33,544.68
find_in_text	63.07	15,451.43

- **Future Work:** The consistent failure across all models on "Hard" samples (0.0%) suggests that complex, multi-hop reasoning for citation verification remains an open challenge for the next generation of autonomous research agents.

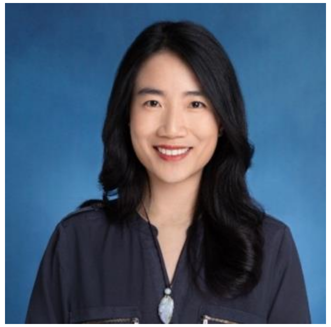
Summary of Key Takeaways

❖ Today's proof-of-concepts are evolving into tomorrow's collaborators. This requires:

- **Eyes and Hands:** *Webwatcher* provides active exploration beyond static text.
- **A Mind and Blueprint:** Our *survey for AIScientists* structures the intelligence and capabilities needed.
- **A Voice of Trust:** *CiteGuard* ensures this voice is credible and verifiable.

❖ The future of autonomous discovery is collaborative, multimodal, and agentic in tool-enhanced interaction with the real-world environment.

We are a Young and Energetic Team



Yi R. (May) Fung, PI
https://mayrfung.github.io

Zhitao He

Shijue Huang

Zhaochen Su

Zeyu Qin

Rui Min

Yuchen Huang

Xinyu Geng

Overarching Goal: Advance **human-centered trustworthy AI** with **multimedia knowledge reasoning** capability and **scalable alignment** principles for helping **solve real-world problems.**