Interview with Akari Asai - Beyond Scaling: Frontiers of Retrieval-Augmented Language Models

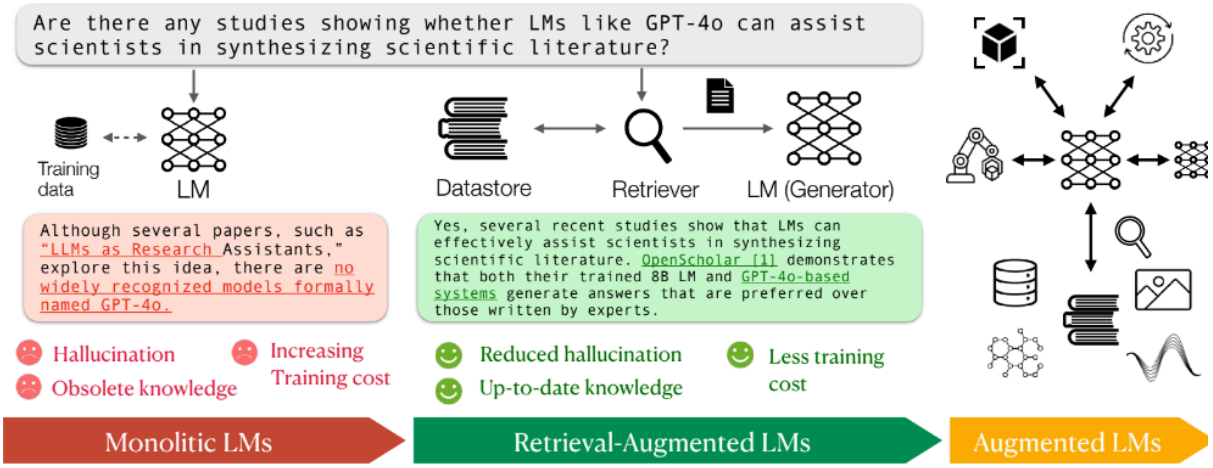
DOI: 10.1145/3834756.3834759

Akari Asai is a Research Scientist at the Allen Institute for AI (AI2) and an incoming Assistant Professor in the School of Computer Science at Carnegie Mellon University. She received her Ph.D. from the University of Washington. Her research advances reliable and up-to-date language systems, especially retrieval-augmented and agentic language models, and applies them to high-impact settings, including scientific discovery and information access for underrepresented languages. Akari's work has been published at top-tier NLP and ML venues and in Nature, and has been recognized with multiple paper awards, the IBM Global Fellowship, and industry grants. She is a recipient of the AAAI/ACM SIGAI Doctoral Dissertation Award, was named to MIT Technology Review's Innovators Under 35 and Forbes 30 Under 30 (Asia, Science), and has been featured by outlets including Forbes and MIT Technology Review.

You were awarded the 2025 AAAI Doctoral Dissertation Award. What was the topic of your dissertation research, and why was this an interesting area of study to you?

My PhD dissertation was on Retrieval-Augmented Language Models. Language Models, or LMs, have made remarkable progress in recent years by scaling up training data and model sizes. However, they still face critical limitations: they hallucinate facts, rely on outdated knowledge, and their outputs are often difficult to verify. These issues are especially problematic in high-stakes domains like scientific research.

My thesis argued that overcoming these challenges requires moving beyond monolithic LMs toward Augmented LMs: systems that are designed, trained, and deployed alongside complementary modules. Specifically, my work pioneered Retrieval-Augmented LMs, which locate relevant knowledge from large-scale text collections and incorporate it at inference time, rather than relying solely on what the model has memorized during training. By grounding generation in retrieved evidence, these systems can produce more accurate, up-to-date, and verifiable outputs.



The pathway from monolithic to augmented LMs. Image credits: Akari Asai

This was a particularly exciting area because, when I started working on it, the prevailing belief in the community was that simply scaling models larger would eventually solve these problems. Our work challenged that assumption and showed that retrieval augmentation offers a fundamentally more effective and efficient path forward. Retrieval-Augmented LMs, such as retrieval-augmented generation (RAG), have since been widely adopted across both academia and industry, powering systems like generative search engines.

What were the main contributions or findings of your dissertation?

My dissertation addressed three core questions about Retrieval-Augmented LMs: why they are necessary, how to build strong foundations for them, and how to deploy them for real-world impact.

First, we conducted one of the earliest large-scale studies on LM hallucinations. We showed that even very large models with hundreds of billions of parameters struggle to reliably recall less popular, long-tail facts, and that simply scaling models up is insufficient to fix this. In contrast, Retrieval-Augmented LMs significantly reduce hallucinations by accessing external knowledge during inference, while also improving training efficiency.

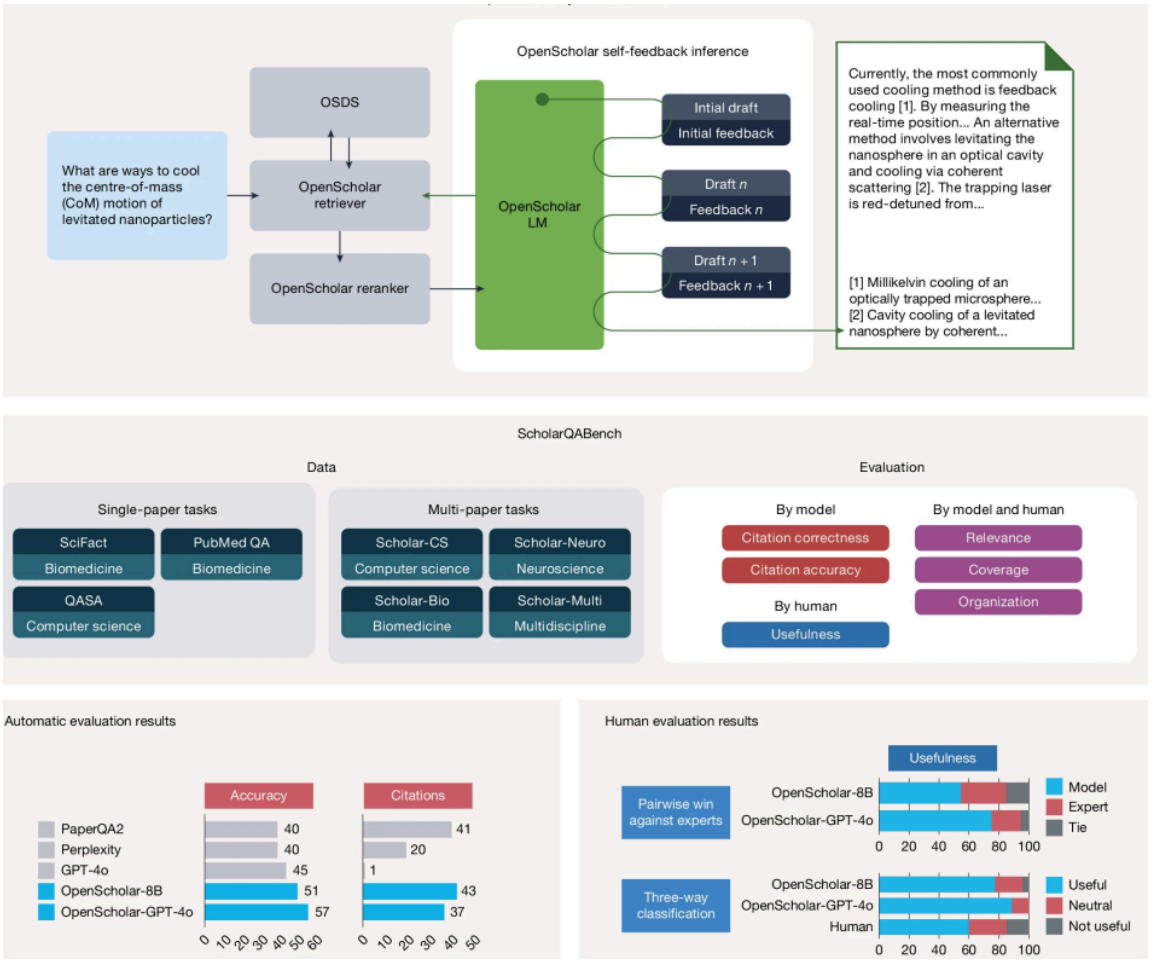
Second, we built new technical foundations for Retrieval-Augmented LMs. The most prominent contribution here is Self-RAG, which trains LMs to dynamically decide when to retrieve, when to generate, and how to self-evaluate their own outputs. This made Retrieval-Augmented LMs far more flexible and reliable than earlier approaches that simply concatenated retrieved documents to a prompt. Self-RAG has been integrated into major LLM libraries, such as LlamaIndex and LangChain. We also developed the first instruction-following retrieval systems, which enable retrievers to adapt to diverse user needs across tasks. This approach has since become the de facto standard for state-of-the-art embedding systems.

Finally, we demonstrated real-world impact. Our flagship application, OpenScholar, was just published in Nature and is the first fully open Retrieval-Augmented LM designed to help scientists synthesize scientific literature. OpenScholar searches over 45 million open-access

papers and generates citation-backed responses. In expert evaluations, scientists preferred OpenScholar's answers over those written by human experts 51% of the time, and the system achieves citation accuracy on par with domain experts, while GPT-4o hallucinates citations 78 to 90% of the time. We also released the first public demo of its kind so the community could interact with the system, and more than 30,000 researchers and practitioners used it. We also made contributions to retrieval-augmented code generation and cross-lingual information access, particularly for under-resourced languages.

How has your research developed since then? What are you focusing on now?

I continue pushing the vision of Augmented LMs: systems where language models are designed in conjunction with other models and tools, and natively trained to use them. Rather than treating retrieval or other modules as an afterthought or a plug-in, I believe we need to fundamentally rethink how we train and deploy these models to work with external components.



Overview of OpenScholar (top) and ScholarQABench (middle). The results (bottom) show that OpenScholar with a trained 8B or GPT-4o substantially outperforms other systems and is preferred over experts more than 50% of the time in human evaluations. Image credits: [Asai et. al., Nature, 2026.]

One major direction I have pursued since my dissertation has been advancing deep research agents. Deep Research Agents, which can perform many searches autonomously and produce long-form, detailed reports, have become popular and widely adopted. However, how to build systems that are competitive with proprietary deep research agents, such as OpenAI Deep Research, has remained underexplored. Building on the foundations of OpenScholar, we developed DR Tulu, the first open model directly trained for long-form deep research using large-scale reinforcement learning. DR Tulu performs multi-step search and synthesis to produce comprehensive, well-attributed reports. Despite being only an 8-billion-parameter model, it matches or exceeds proprietary deep research systems like OpenAI Deep Research and Gemini across science, healthcare, and general domains, and we released all the code, data, and models openly. I am also exploring new frontiers and applications of such agentic systems, especially for scientific discovery, expanding OpenScholar efforts.

How have you seen your field of research evolve in recent years?

The field has changed enormously since I started my PhD in 2019. That year, one of the first successful pre-trained LMs, BERT, had just been released. Before that, we were designing new network architectures and training models separately on each task. BERT introduced the paradigm of pre-training and fine-tuning, and then LLMs emerged and demonstrated that even without fine-tuning, these models can perform highly competitively across many tasks. In that sense, both the problems people work on and how we approach them have changed dramatically.

At the same time, my longer-term ambition hasn't changed much since my undergraduate days. I have always wanted to build language technology systems that help people access the information they need to empower human beings through better tools. I am excited that LMs are now powerful enough to serve as knowledge intermediaries that can help people navigate vast amounts of information. OpenScholar showed that expert scientists can genuinely accelerate their work using our systems. I would like to continue developing reliable and adaptive agentic systems for real-world challenges, as well as addressing the remaining fundamental challenges of LMs, from understanding their risks to exploring new architectures.

Which future directions or open questions excite you most?

LMs today are remarkably powerful, but there remains a significant gap between closed, proprietary systems and open models, especially for long-horizon, challenging tasks. I am excited about exploring how we can build open foundation models for such tasks, spanning learning and inference algorithms as well as the infrastructure to support them, and that help people adapt these models for their own goals. Many real-world tasks, such as those in scientific discovery, require adaptation to local environments or specific domains, which is often not easily achieved by simply prompting the best proprietary models. As these models are increasingly integrated into local environments to solve real tasks, it becomes critical to explore how they perform in out-of-domain scenarios and how to make them robust and adaptable. I would also like to continue working on real-world applications, particularly in scientific discovery, where I see enormous potential for these systems to make a tangible difference.

Were you always set on being a researcher?

Not at all. When I was an undergraduate, I actually switched my major from Economics to Computer Science, and initially I wanted to become a software engineer. After doing engineering internships at a few companies, I found it really exciting that we could directly contribute to products used by millions or even billions of users e.g., search engines. But I started feeling that I wished we could share more openly what we had discovered internally, and that I wanted to work on truly open-ended problems where the solution isn't yet known. That desire drove me to pursue a research career and a PhD, and it later heavily influenced my commitment to open research and my decision to go into academia.

Do you have any advice for early career researchers in your field?

The field moves incredibly quickly, especially in LLMs, and junior researchers can feel a lot of pressure to publish more and more papers at a faster pace on mainstream topics. But I have found it is really important to take time to discover your own interests and develop a longer-term research agenda beyond short-term projects to learn areas deeply and invest time in high-impact work.

My most cited paper and my Nature paper both took a year or more to complete. While I was working on them, I worried that I was spending too much time on a single project. But in retrospect, that investment was necessary, and those projects ended up opening up many opportunities. So I would encourage early career researchers to be patient, to focus on depth and rigor, and to trust that doing things well will pay off in the long run.