Frontiers of Retrieval-Augmented Language Models: Advancing AI Models for Real-World Impact

Large language models (LMs) demonstrate impressive capabilities, yet they often struggle in high-stakes, real-world applications like medical diagnosis or scholarly literature synthesis. Specifically, **Conventional LMs** (Figure 1, left) are prone to generating factual inaccuracies (i.e., hallucinations), lack transparency for output verification, and require costly retraining for updates. Scaling LMs alone does not resolve these issues. I introduced a transformative shift from monolithic LMs, pioneering **Retrieval-Augmented LMs** (Figure 1, middle). Retrieval-Augmented LMs retrieve external knowledge from large-scale text data (*datastores*) during inference, enabling them to transparently and accurately handle complex input, such as answering scientists' literature questions by consulting millions of up-to-date papers. I have established foundational work across three key areas:

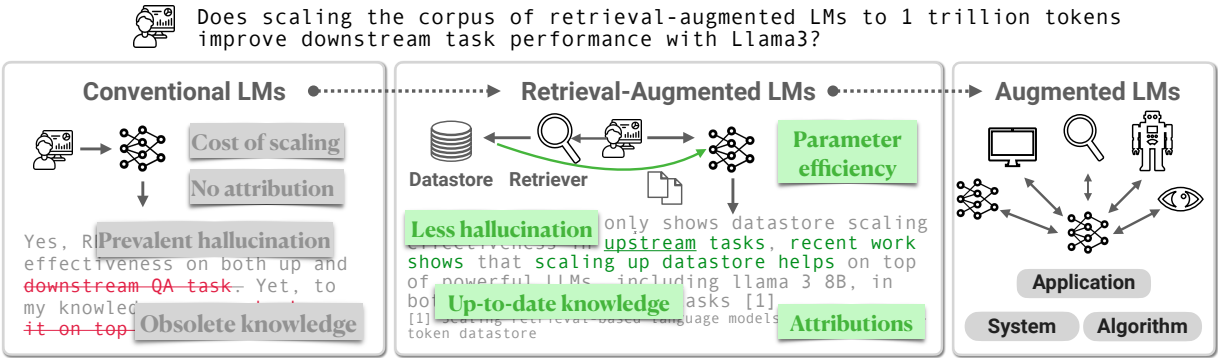

Figure 1: Roadmap from Conventional LMs to **Retrieval-Augmented LMs** to **Augmented LMs**.

- **Establishing the Necessity of Retrieval-Augmented LMs (§1):** My work was among the first to expose the limitations of scaling LMs, particularly hallucinations, and to propose Retrieval-Augmented LMs as a solution. I demonstrated that, while models like GPT-3 struggle with rare or long-tail knowledge, Retrieval-Augmented LMs enhance accuracy and provide transparent attributions for verification. Additionally, I showed that they are more compute-optimal and permit continuous knowledge updates without requiring retraining. [1]–[8]
- **Building the Foundation for Retrieval-Augmented LMs (§2):** To enhance the reliability and versatility of modern Retrieval-Augmented LMs, I designed new methods to advance both retrievers and LMs themselves. My framework, Self-RAG, introduces *dynamic retrieval and self-evaluation*, enabling smaller models to match ChatGPT's factuality. I also pioneered *instruction-following retrieval* and designed TART, the first retriever that adapts to user instructions in real time, establishing the foundation for today's text and multimodal retrievers. [9]–[17]
- **Making Real-World Impact through Retrieval-Augmented LMs (§3):** My research advances Retrieval-Augmented LM technology to create AI models that address: (1) *expert domains*: I developed OpenScholar, an open Retrieval-Augmented LM capable of synthesizing insights from millions of papers with expert-level accuracy, and (2) *low-resource languages*: I created the first multilingual Retrieval-Augmented LMs to enhance information access across the globe, along with benchmarks for under-resourced languages, including African languages. [18]–[28]

Broader Impacts: My research—recognized with honors such as an ACL Paper Award, an ICLR Top 1% Oral Presentation, and MIT Technology Review's Innovators Under 35 in Japan—has advanced and popularized Retrieval-Augmented LMs. My work is integrated into widely-used libraries with millions of downloads and has influenced high-impact tools like ChatGPT Search and Perplexity.

Future Work: I envision future AI models that seamlessly integrate complementary components like retrieval, verification tools, and specialized models, forming **Augmented LMs** (Figure 1, right). Achieving this will require coordinated advancements across AI stacks, systems, algorithms, and applications. Building on my Retrieval-Augmented LMs work, I will (§4) (1) advance scalable algorithms and training methods for Retrieval-Augmented and Augmented LMs, (2) optimize resource use and efficiency in these multi-component systems, and (3) partner with domain experts to apply these innovations in real-world settings, rigorously evaluating their effectiveness and risks.

1 Establishing the Necessity of Retrieval-augmented LMs

Despite their impressive capabilities and growing utility, LMs remain plagued by significant uncertainties about their potential for catastrophic failures and whether scaling alone can mitigate these risks. I systematically analyzed LM limitations and proposed Retrieval-Augmented LMs to overcome them. My work was among the first to develop Retrieval-Augmented LMs, demonstrating that they (1) **significantly reduce hallucinations** and (2) **improve training efficiency** through compute-efficient scaling while enabling knowledge updates without retraining. To promote adoption, I led the first tutorial on this topic at ACL [6], which has been integrated into university courses, and delivered over 25 invited talks and nine guest lectures at five universities as a leading expert.

Evaluating and Mitigating LM Hallucinations: A fundamental challenge for LMs is hallucination, i.e., outputs containing factual errors, but scant research explained when and why this occurs. I conducted the first large-scale quantitative analysis of hallucinations related to factual knowledge memorization in LMs [1], showing that scaling primarily helps models memorize common knowledge (red line in Figure 2), but they still struggle with long-tail knowledge (blue line), i.e., information underrepresented in pre-training data. Even models with billions of parameters, such as GPT-3, could not overcome long-tail hallucinations. To address this, I developed one of the earliest Retrieval-Augmented Generation (RAG) systems for LMs, proving that retrieving passages at *inference time* could significantly reduce hallucinations, especially in long-tail cases. This work, awarded the ***ACL Best Video Award: Most Viewed***, sparked numerous follow-up studies, including our own analysis on real user-LM interactions [5]. In this follow-up, we introduced a fine-grained hallucination taxonomy, showing that even models like ChatGPT hallucinate in over 50% of complex queries. Our Retrieval-Augmented LM, **FAVA**, trained for hallucination detection successfully identified and mitigated such diverse hallucinations.

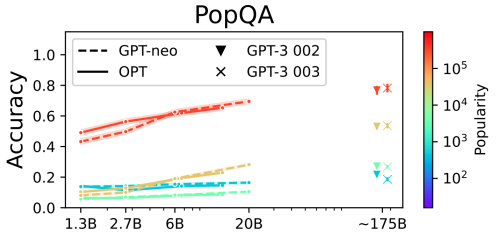

Figure 2: LM performance on questions with varying subject popularity.

Improving Efficiency with Retrieval-Augmented LMs: Another significant drawback of LMs is their inefficiency in training. Encoding extensive world knowledge directly into model parameters requires substantial computational resources for initial pretraining and continuous updates to prevent knowledge obsolescence. As shown in Figure 3, Retrieval-Augmented LMs (red line), leveraging a 1.4 trillion-token datastore, achieve comparable performance to much larger conventional LMs (blue line) across multiple tasks. This significantly reduces the FLOPs needed for pretraining with only a modest increase in FLOPs required for building the datastore [4]. Furthermore, I show that Retrieval-Augmented LMs can incorporate up-to-date knowledge without retraining simply by updating the datastore, reducing errors in responses to time-sensitive user queries (e.g., politics) by 40% [8].

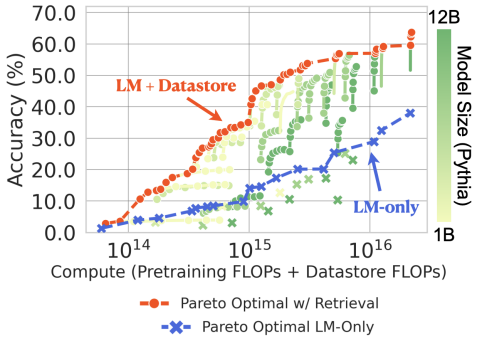

Figure 3: Retrieval-Augmented LMs require less training and indexing FLOPs.

2 Building the Foundation for Retrieval-augmented LMs

Retrieval-Augmented LMs consist of three core components: a datastore, a retriever, and an LM for generation, as shown in Figure 1 (middle). My research has advanced all three, laying a principled foundation for modern Retrieval-Augmented LMs. Specifically, I (1) designed **LMs optimized for retrieval augmentation**, and (2) developed **a new generation of retrieval systems with improved versatility and robustness**. My work transformed earlier models designed for specific tasks such as Question Answering (QA) into a general framework capable of solving a broader range of problems.

Designing and Training LMs for Retrieval Augmentation: While combining off-the-shelf LMs with retrieval models (i.e., RAG) can be effective, it has limitations: LMs often fail to filter irrelevant

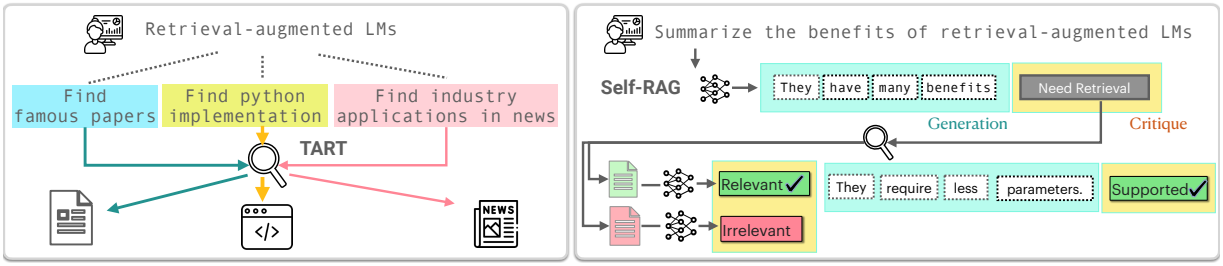

Figure 4: Overview of **TART** (left) and **Self-RAG** (right).

context, generate unsupported outputs, while unnecessary retrieval can degrade both efficiency and performance. I build novel training and inference algorithms as well as new architectures that help LMs more economically leverage retrieved information [9]–[11]. One such method, **Self-RAG** [9] (Figure 4 right), recognized as an ***ICLR Oral Presentation (Top 1%)*** and an ***Best Paper Honorable Mention at the NeurIPS workshop***, trains LMs to adaptively retrieve—retrieving only when necessary—distinguish between helpful and unhelpful context, and self-critique outputs based on their faithfulness to cited evidence, leveraging learned special tokens and controlled decoding. Despite its smaller size (8B and 13B parameters), Self-RAG outperforms much larger LMs across tasks, producing responses more reliably supported by citations. It has gained significant traction in academia and industry, with support from major libraries like LlamaIndex and LangChain (15M+ monthly installations) and recognition in *Forbes* as a key RAG advancement for real-world applications.

Developing Versatile and Robust Retrievers: Despite advances in neural retrieval, challenges remain in achieving versatility, i.e., the ability to identify relevant context across diverse tasks. Fact-based tasks like QA often rely on word similarity, while tasks such as mathematical reasoning require retrieval that is based on deeper reasoning patterns even when surface similarity is low. I have aimed to develop a robust, general-purpose retrieval system that can adapt to various information needs, enabling more versatile Retrieval-Augmented LMs [12], [13]. Notably, I pioneered the retrieval-with-instruction paradigm [12] and introduced **TART** (Figure 4 left), the first retrieval system that dynamically adjusts its behavior based on user-provided instructions, much like LMs adapt flexibly to instructions. To support this paradigm shift, I built the first large-scale dataset for instruction-tuning for retrieval systems, enabling comprehensive training and evaluation. TART sparked numerous follow-up studies with many of today's top-performing text retrievers and multimodal retrieval systems. Traditional retrieval systems also often lack robustness for complex queries—such as multi-hop questions or obscure entities—leading to missed crucial implied information. Therefore, I developed **PathRetriever** [14], a graph-based system that combines neural models with symbolic graph traversal to enhance both reliability and explainability. This system was ***integrated into Salesforce's COVID Search in 2020***, providing researchers with critical COVID-related information. I also developed **LUKE** [15] (***Most Influential EMNLP 2020 Paper No.8; downloaded 1.4 million times on HuggingFace***), introducing entity-aware pre-training to enhance text representations with entity knowledge, improving retrieval robustness for entity-centric queries.

3 Making Real-World Impact through Retrieval-augmented LMs

In addition to building these LM advancements, I demonstrate their transformative power in solving complex real-world problems in domains where LMs struggle due to limited or underrepresented data. My research addresses challenges in two critical areas: (1) **expert-domain tasks**, aiding researchers in synthesizing scientific literature and enhancing LM-based code generation, and (2) **low-resource languages**, improving information access equity across diverse world languages.

Developing Trustworthy Retrieval-Augmented LMs for Science and Code: LMs can be unreliable in high-stakes applications like software development assistance and scientific literature synthesis. I led a team from UW, AI2 Semantic Scholar, Meta, and five universities to develop **OpenScholar** [18], the first open Retrieval-Augmented LM to enhance cross-disciplinary literature synthesis. OpenScholar utilizes our datastore with 45 million open-access papers, trained scientific models, and a self-feedback framework to boost reliability and citation accuracy. We evaluated OpenScholar with **ScholarBench**, a large-scale benchmark for literature synthesis across fields like computer science, physics, biomedicine, and neuroscience, developed with Ph.D. experts. OpenScholar outperformed

GPT4o and was preferred by experts in over 70% of cases. Its public demo[1] attracted ***11k+ users in one week and featured in VentureBeat***. Additionally, we introduced **CodeRAG-Bench** [19] to enable comprehensive evaluations of Retrieval-Augmented LMs for code generation tasks.

Improving Information Access across Languages with Retrieval-Augmented LMs: Many LMs are optimized for English [21]. Furthermore, with web resources varying widely across languages [20], most world languages lack access to essential information, creating barriers to access to education, healthcare, and social participation. To address this, I pioneered multilingual Retrieval-Augmented LMs that retrieve and reason across languages and created benchmarks for languages with varying resources. I introduced the first end-to-end multilingual Retrieval-Augmented LM, **CORA** [22] that trains a unified multilingual retriever and LM using large-scale synthetic data, bypassing the need for language-specific or translation models. CORA set new state-of-the-art results in multilingual open-domain QA across 26 languages. I also developed **XORQA** [23], the first large-scale cross-lingual open-retrieval QA dataset covering seven different languages. These efforts led to collaborations on the first QA dataset for 10 African languages [24] and a cross-lingual product QA system for Amazon [25]. Additionally, I led the inaugural workshop on multilingual information access [26] and organized shared tasks on multilingual Retrieval-Augmented generation [27].

4 Future Work

As AI takes on increasingly complex, high-stakes roles, I envision a new generation of LMs that I refer to as **Augmented LMs** (Figure 1 right), i.e., models designed to seamlessly integrate diverse components like retrieval mechanisms, verification tools, and specialized domain models. Building on my work in Retrieval-Augmented LMs, my future research will push the boundaries of Augmented LMs, leveraging expertise across the AI stack, from infrastructure and algorithms to applications. This work will involve close collaboration with domain experts in areas such as biomedicine and systems engineering to ensure rigorous validation and impact in expert-driven, high-stakes settings.

Developing New Architectures and Training for Augmented LMs: Current AI architectures and training approaches, designed for monolithic LMs, fall short in addressing the complex input and inference demands of Augmented LMs. Advancing these models will require fundamentally *new architectures that dynamically adapt to diverse signals and real-time contexts.* I aim to explore improved caching mechanisms and transformer alternatives like state-space models for efficient long-context processing, as well as designs that more effectively integrate multimodal input [28]. Achieving optimal performance will also involve *tailored training strategies for multi-component systems*, with objectives that incorporate feedback across components and use advanced synthetic data generation.

Improving Infrastructures for Augmented LMs: Augmented LMs bring unique system-level challenges that require infrastructure capable of supporting diverse computational needs. For instance, unlike conventional monolithic LMs, which primarily depend on GPU-intensive tasks with minimal data transfer, Retrieval-Augmented LMs need optimized CPU and memory usage, high-speed interconnects, and efficient data flow across nodes [4]. To address unique requirements, I plan to collaborate with systems experts to pinpoint inefficiencies in current resource allocation. Together, we will *design infrastructure that optimally allocates resources across the entire Augmented LM pipeline*, adjusting to different stages of inference, informed by insights from algorithmic advancements.

Integrating and Advancing Augmented LMs for New Expert Domains: I am committed to *developing AI models that address the specific needs of specialized fields and less-common languages*. Building trustworthy AI for these areas requires custom advancements and rigorous evaluation with domain experts beyond Computer Science, as demonstrated with OpenScholar. Moving forward, I will design Augmented LMs that tackle the unique challenges of expert domains by collaborating directly with professionals. For instance, I plan to work closely with clinicians to develop medical Augmented LMs for diagnosing rare conditions while rigorously assessing potential risks to ensure safety. By aligning models to these distinct needs and carefully monitoring real-world outcomes, I aim to create robust, responsible AI systems that provide meaningful support to yield innovative findings in specialized fields.

[1] https://openscholar.allen.ai/

References

[1] Mallen*, **Asai***, Zhong, Das, Hajishirzi, and Khashabi, "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[2] **Asai** and Choi, "Challenges in information-seeking qa: Unanswerable questions and paragraph retrieval," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[3] Chen, **Asai***, Mireshghallah*, Min, Grimmelmann, Choi, Hajishirzi, Zettlemoyer, and Koh, "Copybench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024.

[4] Shao, He, **Asai**, Shi, Dettmers, Min, Zettlemoyer, and Koh, "Scaling retrieval-based language models with a trillion-token datastore," in *Neural Information Processing Systems (NeurIPS)*, 2024.

[5] Mishra, **Asai**, Balachandran, Wang, Neubig, Tsvetkov, and Hajishirzi, "Fine-grained hallucination detection and editing for language models," in *Conference on Language Modeling (COLM)*, 2024.

[6] **Asai**, Min, Zhong, and Chen, "Retrieval-based language models and applications," in *Annual Meeting of the Association for Computational Linguistics (ACL, Tutorial)*, 2023.

[7] **Asai**, Zhong, Chen, Koh, Zettlemoyer, Hajishirzi, and Yih, "Reliable, adaptable, and attributable language models with retrieval," *ArXiv*, 2024.

[8] Kasai, Sakaguchi, Takahashi, Bras, **Asai**, Yu, Radev, Smith, Choi, and Inui, "RealTime QA: What's the answer right now?" In *Neural Information Processing Systems (NeurIPS, Dataset and Benchmark)*, 2022.

[9] **Asai**, Wu, Wang, Sil, and Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *International Conference on Learning Representations (ICLR)*, 2024.

[10] **Asai**, Gardner, and Hajishirzi, "Evidentiality-guided generation for knowledge-intensive nlp tasks," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

[11] **Asai** and Hajishirzi, "Logic-guided data augmentation and regularization for consistent question answering," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[12] **Asai**, Schick, Lewis, Chen, Izacard, Riedel, Hajishirzi, and Yih, "Task-aware retrieval with instructions," in *Findings of ACL*, 2022.

[13] Lin, **Asai**, Li, Oğuz, Lin, Mehdad, Yih, and Chen, "How to train your DRAGON: Diverse augmentation towards generalizable dense retrieval," in *Findings of EMNLP*, 2023.

[14] **Asai**, Hashimoto, Hajishirzi, Socher, and Xiong, "Learning to retrieve reasoning paths over wikipedia graph for question answering," in *International Conference on Learning Representations (ICLR)*, 2020.

[15] Yamada, **Asai**, Shindo, Takeda, and Matsumoto, "LUKE: Deep contextualized entity representations with entity-aware self-attention," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[16] Yamada, **Asai**, Sakuma, Shindo, Takeda, Takefuji, and Matsumoto, "Wikipedia2vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from wikipedia," in *Conference on Empirical Methods in Natural Language Processing (EMNLP, System and Demonstrations)*, 2020.

[17] Shi, Yu, **Asai**, Durrett, Hajishirzi, and Zettlemor, "The 4th workshop on knowledge-augmentation for language models and nlp methods," *NAACL 2025 Workshop*, 2025.

[18] **Asai**, He, Shao, Shi, Singh, Chang, Lo, Soldaini, Feldman, Mike, Wadden, Latzke, Minyang, Ji, Liu, Tong, Wu, Xiong, Zettlemoyer, Weld, Neubig, Downey, Yih, Koh, and Hajishirzi, "OpenScholar: Synthesizing scientific literature with retrieval-augmented lms," *Arxiv*, 2024.

[19] Wang, **Asai**, Yu, Xu, Xie, Neubig, and Fried, "CodeRAG-Bench: Can retrieval augment code generation?" *ArXiv*, 2024.

[20] Yu*, **Asai***, Chatterjee, Hu, and Choi, "Beyond counting datasets: A survey of multilingual dataset construction and necessary resources," in *Findings of EMNLP*, 2022.

[21] **Asai**, Kudugunta, Yu, Blevins, Gonen, Reid, Tsvetkov, Ruder, and Hajishirzi, "BUFFET: Benchmarking large language models for few-shot cross-lingual transfer," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2023.

[22] **Asai**, Yu, Kasai, and Hajishirzi, "One question answering model for many languages with cross-lingual dense passage retrieval," in *Neural Information Processing Systems (NeurIPS)*, 2021.

[23] **Asai**, Kasai, Clark, Lee, Choi, and Hajishirzi, "XOR QA: Cross-lingual open-retrieval question answering," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.

[24] Ogundepo, Gwadabe, Rivera, Clark, Ruder, Adelani, Dossou, Diop, Sikasote, Hacheme, Buzaaba, Ezeani, Mabuya, Osei, Emezue, Kahira, Muhammad, Oladipo, Owodunni, Tonja, Shode, **Asai**, Ajayi, Siro, Arthur, Adeyemi, Ahia, Anuoluwapo, Awosan, Chukwuneke, Opoku, Ayodele, Otiende, Mwase, Sinkala, Rubungo, Ajisafe, Onwuegbuzia, Mbow, Niyomutabazi, Mukonde, Lawan, Ahmad, Alabi, Namukombo, Chinedu, Phiri, Putini, Mngoma, Amuok, Iro, and Adhiambo34, "AfriQA: Cross-lingual open-retrieval question answering for african languages," in *Findings of EMNLP*, 2023.

[25] Shen, **Asai**, Byrne, and Gispert, "xPQA: Cross-lingual product question answering in 12 languages," in *Annual Meeting of the Association for Computational Linguistics (ACL, Industry)*, 2023.

[26] **Asai**, Choi, Clark, Hu, Lee, Kasai, Longpre, Yamada, and Zhang, "The 1st workshop on multilingual information access (mia)," *NAACL 2022 Workshop*, 2022.

[27] **Asai**, Longpre, Kasai, Lee, Zhang, Hu, Yamada, Clark, and Choi, "MIA 2022 shared task: Evaluating cross-lingual open-retrieval question answering for 16 diverse languages," in *Proceedings of the Workshop on Multilingual Information Access (MIA)*, 2022.

[28] Yue, Song, **Asai**, Kim, Nyandwi, Khanuja, Kantharuban, Sutawika, Ramamoorthy, and Neubig, "Pangea: A fully open multilingual multimodal llm for 39 languages," *arXiv*, 2024.