From Retrieval-Augmented Generation (RAG) to Agentic Deep Research

Xin Luna Dong
Meta, Redmond, USA
lunadong@meta.com

Sanat Sharma
Meta, NYC, USA
sanatsharma@meta.com

Kai Sun
Meta, Menlo Park, USA
sunkaicn@meta.com

Jiaqi Wang
Meta, Menlo Park, USA
jqwang@meta.com

Yinglong Xia
Meta, Menlo Park, USA
yxia@meta.com

Xiao Yang
Meta, Menlo Park, USA
xiaoyangfb@meta.com

Scott Wen-tau Yih
Meta, Seattle, USA
scottyih@meta.com

Franklin Zhang
UW CSE, Seattle, USA
frankyz@uw.edu

Abstract

Despite well-known hallucination issues, LLMs have become an increasingly indispensable source of information, and the underlying technology has advanced rapidly to make their answers far more reliable. Early on, *Retrieval-Augmented Generation (RAG)* emerged as the dominant remedy, grounding LLM responses in external knowledge and evolving from simple retrieve-then-read pipelines into modular, graph-enhanced, and agentic systems. More recently, *Agentic Deep Research* has pushed the frontier further, equipping LLMs with autonomous planning, multi-hop investigation, and iterative synthesis to tackle open-ended questions that no single retrieval pass can answer. This tutorial offers an in-depth treatment of modern RAG and Deep Research, grounded in an AI-assisted systematic analysis of 2,000+ recent papers (2020–2026). Attendees will leave with a structured roadmap, evidence-backed practical recommendations, and a clear map of open research opportunities.

1 Motivation and Scope

The way people seek information is being rebuilt around LLMs: users now turn to AI assistants for everything from quick lookups to in-depth investigations. Yet LLMs remain prone to hallucinations and stale knowledge (Tonmoy et al., 2024; Yang et al., 2024a), making trustworthy, current, well-grounded answers one of the field's central engineering challenges.

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs) by retrieving relevant external knowledge at inference time rather than relying solely on parametric memory (Huang et al., 2023; Ji et al., 2023). RAG reduces hallucinations with grounded generation, and offloads long-tail factual knowledge to external corpora, enabling smaller, more efficient models without sacrificing coverage (Gao et al., 2024). However, RAG itself introduces new challenges: noisy retrieval, knowledge conflicts between parametric and retrieved information, and wasted computation from unnecessary retrieval. The past three years have witnessed a burst of research aimed at improving RAG's factuality, robustness, generalizability, and latency.

The next leap is already underway. Products such as *Gemini Deep Research*, *OpenAI Deep Research*, and *Claude Research* go beyond single-pass retrieval, equipping LLMs with *Agentic Deep Research* capabilities—planning multi-step investigations, issuing iterative searches and tool calls, and synthesizing comprehensive reports. Deep Research opens the door to open-ended queries while raising new challenges in planning, long-horizon credit assignment, faithful synthesis, and exploration cost.

This tutorial offers a comprehensive treatment of both. For RAG, we organize the landscape into three paradigms—*Modular Pipeline*, *Graph-Enhanced*, and *Agentic* RAG. For Deep Research, we cover system architectures, planning and tool-use strategies, and evaluation. We also situate both within *context engineering*, *agentic systems*, and *long-term memory*, and discuss multi-modal extensions across text, images, and structured data.

This tutorial is itself an instance of AI-facilitated Deep Research in action. The authors first read ∼150 classical papers to develop an initial taxonomy, then employ a research agent to analyze ∼2,000 additional papers on Factuality, RAG, and Agent (2020–2026) to generate comprehensive summaries[1]. The presenters curate and enrich this synthesis with first-hand practical insights that may not be documented elsewhere. New papers will continue to be incorporated before and after the tutorial to support ongoing learning. This combination of AI-driven synthesis with expert curation in

[1] papers.lunadong.com/area/factuality; papers.lunadong.com/area/rag; papers.lunadong.com/area/agent.

itself is a contribution, setting a replicable model for future tutorials and surveys.

1.1 Tutorial Type

This is a **cutting-edge** tutorial. It synthesizes the rapid advances in RAG from 2020–2026, with particular emphasis on emerging paradigms (graph-enhanced and agentic RAG) and techniques (RL-driven retrieval policies, deep research agents) from 2025–2026 that have not yet been covered in prior tutorials. While we provide sufficient background for newcomers with basic LLM familiarity, the primary focus is on presenting the latest research frontier and open problems.

1.2 Target Audience and Prerequisites

- **Researchers** in NLP, data mining, and knowledge management seeking a structured map of the RAG landscape with clear identification of open problems.
- **Applied scientists and engineers** building production LLM systems that need evidence-backed guidelines for choosing among RAG architectures and retrieval strategies.
- **Graduate students** entering the field who wish for a comprehensive entry point to 2K+ papers organized by topic, paradigm, and timeline.

Prerequisites: Familiarity with basic LLM concepts (pre-training, fine-tuning, prompting). No prior knowledge of RAG or hallucination detection is assumed.

2 Related Work

No prior tutorial covers the complete 2023–2026 RAG landscape. Ours introduces a three-paradigm framework (Modular, Graph-Enhanced, Agentic), incorporates 18+ months of advances beyond the most recent tutorial—including RL-driven retrieval policies, GNN-LLM integration, multi-hop reasoning, and autonomous agentic systems—and provides evidence-backed deployment recommendations synthesized from ∼2000 papers.

3 Tentative Schedule (3 Hours)

3.1 Introduction & Factuality Overview (30 min)

We open by demonstrating failure modes of standalone LLMs on knowledge-intensive tasks: outdated facts, entity confusion, confident fabrication, and multi-hop reasoning breakdowns. We provide a high-level overview of the factuality landscape, preview the two-pronged solutions as two sides of the coin: *knowledge internalization and hallucination suppression*—improving what models know and how reliably they express it, and *retrieval-augmented generation*—grounding in external evidence. We show RAG's evolution timeline from simple retrieve-and-read pipelines (2023) through adaptive and graph-enhanced architectures (2024) to autonomous RL-driven agentic systems (2025–2026), based on our systematic analysis of ∼2000 papers. The remainder of the tutorial deep-dives into RAG.

3.2 RAG Deep Dive (60 min)

The next part of the tutorial provides a comprehensive survey of RAG, across *Modular Pipeline RAG* and *Graph-Enhanced RAG*.

3.2.1 Modularized RAG Pipeline (40 min)

We introduce the five-stage modular architecture—triggering, query rewriting, retrieval, post-processing, and answer generation—as independently optimizable components (Gao et al., 2024; Lewis et al., 2020). We trace the evolution from monolithic retrieve-and-read (2023) to modular, swappable pipelines (2024—2025) and show how each stage can be upgraded independently.

1. *RAG Triggering*—When *not* to retrieve: adaptive triggering is the first and most impactful efficiency decision. We analyze the cost—quality trade-off and provide guidelines for choosing a triggering strategy based on deployment constraints.

2. *Query Rewriting*—Transforming user queries for effective retrieval. We describe various strategies and identify which strategies work best for single-hop vs. multi-hop questions.

3. *Retrieval*—The retrieval stage itself: We focus on the surprising finding that retrieval relevance correlates weakly with downstream generation quality—what matters is *informativeness*, not *similarity*.

4. *Post-processing & Context Compression*—What to do with retrieved passages before feeding them to the generator: We describe various methods for reranking, context compression, and position bias mitigation.

5. *Answer Generation & Knowledge Conflicts.*—The final generation stage: We focus on

noise-resilient generation that maintains quality even with imperfect retrieval, and the critical problem of knowledge conflicts:

There are a lot of studies on each component, and we focus on explaining how the most prominent problem is solved. Additionally, we highlight how each component can be extended to answer complex questions, where assembling and reasoning is essential.

3.2.2 Graph-Enhanced RAG (20 min)

Graph-enhanced RAG pipelines construct *Knowledge Graphs (KGs)* from corpora and leverage graph structure to enable structured reasoning. We explain how KG construction and hybrid retrieval combine structured and unstructured knowledge for RAG, how we integrate Graph Neural Networks (GNN) with LLM reasoning, and how we tackle multi-hop questions that require chaining evidence across graph nodes for Complex QA.

3.3 Agentic Deep Research Deep Dive (60 min)

Building on the traditional RAG foundations, we then examine how Agentic Deep Research extends retrieval into autonomous, multi-step investigation.

3.3.1 Agentic RAG (30 min)

Static single-pass retrieval fails on complex multi-hop questions, wastes compute on simple queries, and cannot adapt to varying information needs during generation. Recent Agentic AI trend leads to the Agentic RAG pipelines that iteratively decide whether to retrieve, what to retrieve, and how to integrate external knowledge during language model decoding, treating RAG as a dynamic decision-making process. We describe the paradigm shift from hand-crafted pipelines to learned retrieval policies, and how to scale agentic RAG to reasoning-heavy queries.

3.3.2 Deep Research (30 min)

While agentic RAG enables dynamic retrieval decisions during generation, Deep Research extends this paradigm to autonomous, long-horizon research workflows. Motivated by complex information needs (e.g., literature surveys, competitive analyses) that exceed single-query RAG, deep research agents operate over minutes to hours, iteratively browsing the web, reasoning over gathered evidence, and producing structured long-form reports grounded in cited sources. We describe key techniques at each stage of the deep research pipeline (intent clarification and planning, web exploration, reasoning and synthesis, and report generation) and examine how modern systems execute this pipeline.

3.4 Synthesis & Future Directions (30 min)

We conclude the tutorial with **key takeaways** organized by practitioner profile (researcher, engineer, student), with evidence-backed guidelines for choosing the right RAG architecture. We review the **emerging trends** that shape the future of RAG: RL transforming RAG into autonomous agentic systems with process-level supervision; multimodal RAG extending retrieval and generation to images, video, and tables; the convergence of modular, graph-based, and agentic paradigms into hybrid architectures; domain-specific RAG deployment in healthcare, finance, and legal domains with specialized benchmarks; and long-context trade-offs as context windows scale to millions of tokens. Finally, we discuss **open research opportunities** to inspire more research in this field: unified retrieval-generation architectures, adversarially robust retrieval, long-form and multi-turn RAG for dialogue systems, scalable graph construction for knowledge-intensive domains, compute-optimal inference balancing retrieval cost vs. quality, and efficient agentic RAG for latency-sensitive production deployments.

4 Audience Engagement Strategies

We employ three strategies to encourage participation throughout:

Running example: A single compound question evolves across all sessions, with each technique showing concrete improvement over the previous solution, providing a narrative thread that builds intuition incrementally.

Live polls: At key decision points, attendees predict outcomes before we reveal benchmark results, encouraging active reasoning about trade-offs.

Online companion: The interactive survey we maintain over time will let attendees explore papers, methods, timelines, and benchmarks during and after the tutorial, extending engagement beyond the session.

5 Estimated Audience Size

We estimate an audience of **200–300 attendees**. RAG has become one of the most actively re-

searched areas in NLP, and interest has grown steadily over the past three years. Related tutorials have drawn increasing audiences:

- *Retrieval-Augmented Text Generation*, IJCAI 2022 (∼50 attendees).
- *Retrieval-based Language Models and Applications*, ACL 2023 (∼150 attendees).
- *RAG Meets LLMs*, KDD 2024 (∼100 attendees).
- *RAG for AI-Generated Content*, ACL 2024 (∼200 attendees).

The tutorial team also organized the KDD Cup 2024 (CRAG (Yang et al., 2024a)) and KDD Cup 2025 (CRAG-MM (Wang et al., 2025a)) competitions on RAG, which collectively attracted ∼3K registered participants across ∼600 teams, further demonstrating strong community interest. Our tutorial offers the most comprehensive coverage to date—spanning modular, graph-enhanced, and agentic paradigms with 18+ months of advances beyond the most recent tutorial—and we expect continued growth in audience size.

6 Reading List

The following papers provide essential background and representative coverage across the three RAG paradigms. Asterisked entries (*) are recommended pre-reading.

* Lewis et al. (2020). The original paper introducing RAG for knowledge-intensive NLP tasks.
* Gao et al. (2024). A comprehensive survey of RAG for LLMs, covering architecture taxonomy and evaluation.
* Huang et al. (2023). A survey on hallucination in LLMs, motivating the need for RAG.
- Guu et al. (2020). REALM: retrieval-augmented language model pre-training, an early foundational approach.
- Asai et al. (2023). Self-RAG: learning to retrieve, generate, and critique through self-reflection.
- Jiang et al. (2023b). FLARE: active retrieval augmented generation with adaptive triggering.
- Edge et al. (2024). GraphRAG: a graph-based approach to query-focused summarization.
- Jin et al. (2025). Search-R1: training LLMs to reason and leverage search engines with reinforcement learning.
- Yang et al. (2024a). CRAG: a comprehensive benchmark for end-to-end RAG systems.
- Ni et al. (2025). A recent survey on trustworthy RAG, covering robustness and reliability.

A complete reading list of ∼2000 papers, organized by topic and continuously updated, is available at the tutorial's online companion[1].

7 Societal Impact

Positive impacts: More reliable LLM systems in high-stakes domains; better-informed deployment decisions grounded in evidence from ∼2000 papers and the tutors' real practice; acceleration of research toward robust retrieval-augmented architectures; democratized access through openly available interactive survey materials.

Potential concerns: Improved RAG performance could create false confidence if systems are deployed without rigorous evaluation—we explicitly discuss current limitations. Adversarial robustness remains an open challenge, and no existing defense is universally robust. We emphasize that current RAG techniques significantly reduce but do not eliminate hallucination risk.

8 Ethics Statement

This tutorial covers RAG techniques whose primary goal is to *reduce* hallucinations and improve LLM factual reliability–an inherently ethics-positive objective. Still, several considerations arise. First, RAG reduces but does not eliminate hallucination risk, and improved performance may encourage deployment with insufficient human oversight; we address current limitations throughout and caution against treating RAG as a complete factuality solution. Second, agentic RAG and deep research systems that autonomously browse and synthesize web content pose dual-use concerns: the same capabilities enabling trustworthy information gathering could produce more convincing source-grounded disinformation. We discuss adversarial robustness and responsible deployment as open challenges. Third, retrieval over personal or sensitive data–especially for the wearable device applications covered–raises privacy concerns requiring appropriate access controls and data governance. Finally, this tutorial uses AI-assisted analysis of ∼2000 papers; all AI-generated summaries are curated and verified by the presenters for accuracy and proper attribution.

9 Limitations

This tutorial has no venue-specific limitations: the topic is broadly relevant, not region-specific, and requires only standard presentation equipment.

10 Tutors

We have a strong set of tutors, including early inventors of RAG, active researchers, industry practitioners (with extra insights from wearables devices), and strong system builders. The same team also hosted KDDCups on RAG (CRAG (Yang et al., 2024a), CRAG-MM (Wang et al., 2025a)) in the past two years.

10.1 In-Person Presenters

Xin Luna Dong (Email: lunadong@meta.com. Phone: 1-201-650-3494) **(Corresponding Tutor)** is a Principal Scientist at Meta Wearables AI, where she leads the RAG and Agentic AI efforts for building trustworthy and personalized assistants on wearable devices. Previously, she spent over a decade advancing knowledge graph technology, including the Amazon Product Graph and the Google Knowledge Graph. She is co-author of Machine Knowledge: Creation and Curation of Comprehensive Knowledge Bases and Big Data Integration. She was named an ACM Fellow and an IEEE Fellow for "significant contributions to knowledge graph construction and data integration", awarded the VLDB Women in Database Research Award and VLDB Early Career Research Contribution Award, and invited as an ACM Distinguished Speaker. She has presented many tutorials in the past at KDD, WebConf, WSDM, ACL, VLDB, SIGMOD, etc. She serves in the PVLDB advisory committee, was a member of the VLDB endowment, a PC co-chair for KDD'2022 ADS track, WSDM'2022, VLDB'2021, and Sigmod'2018.

Sanat Sharma is a Research Engineer at Meta Reality Labs, working on LLMs, RAG, post training, and NLP. Sanat has contributed to multiple conference papers in venues like ICLR, CVPR, SIGIR, AAAI, ICDM, ECIR. Sanat was an organizer for the KDD 2025 Cup.

Jiaqi Wang is a Machine Learning Engineer at Meta Reality Labs. Her recent work focuses on RAG and Multimodal LLMs. She built the CRAG-MM benchmark and organized the associated KDD Cup competition and workshop. Prior to Meta, she developed personalized ranking models for e-commerce recommendation systems.

Yinglong Xia is an Applied Research Scientist at Meta Recommendation Systems (MRS), where he focuses on large-scale recommendation systems with knowledge-driven AI technology for personalization. Prior to that, he was a Chief Architect on Enterprise AI at Futurewei / Huawei US, and a Research Staff Member and a Postdoc at IBM T.J. Watson Research Center. He published 100+ technical papers and 40+ patents. He is active in professional activities, such as serving as an AE for IEEE TBD/TKDE, an ADS area chair at KDD 2025 and 2026, an industry co-chair at CIKM 2024 and WSDM 2026, etc.

Xiao Yang is a Research Scientist at Meta Reality Labs. Her recent research focuses on RAG, factuality, and agentic AI. Prior to Meta, Xiao focused on Natural Language Understanding at Siri, Apple Inc. Xiao published multiple papers in major AI venues like NeurIPS, EMNLP, NAACL. She also served as the organizer of 2024 and 2025 KDD Cups.

10.2 Contributors

Kai Sun is a Research Scientist at Meta Reality Labs. His recent research focuses on RAG, LLM factuality, and knowledge retrieval, particularly for applications on wearable devices. He served as an organizer for the 2024 and 2025 KDD Cups, both focused on RAG, and has been an area chair for major NLP venues since 2022.

Scott Wen-tau Yih is a Research Scientist at Meta Fundamental AI Research (FAIR) and an affiliate professor at University of Washington. Prior to this, he held notable positions as Principal Research Scientist at the Allen Institute for Artificial Intelligence (2017-2019) and Senior Researcher at Microsoft Research (2005-2017), after receiving his PhD from the University of Illinois at Urbana-Champaign in 2005. Dr. Yih has published numerous influential papers, including WikiQA, DPR, and RAG, which have garnered significant recognition. He received the best paper award from CoNLL'11 and an outstanding paper award from ACL'15. Additionally, he has served in various leadership roles within the NLP and ML communities, including program co-chairs (CEAS'09, CoNLL'14, EMNLP'21) and senior area chairs for NLP (ACL, NAACL, EMNLP, EACL) and ML (ICLR, NeurIPS) conferences. In 2024, Dr. Yih was selected as an ACL Fellow for "significant contributions to information extraction and question answering, neural retrieval and retrieval-augmented generation."

Franklin Zhang is a student at UW CSE, responsible for the infrastructure and maintenance of the Online AI Companion, and actively shapes the tutorial content with fresh learning perspectives.

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, David Stutz, Andr´as Gy¨orgy, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesv´ari, Ali Taylan Cemgil, and Nenad Tomasev. 2024. Mitigating llm hallucinations via conformal abstention. *arXiv preprint*.

Uri Alon, Frank F. Xu, Junxian He, Sudipta Sengupta, Dan Roth, and Graham Neubig. 2022. Neuro-symbolic language modeling with automaton-augmented retrieval. In *ICML*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint*.

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke S. Zettlemoyer, Hanna Hajishirzi, and Wen-tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. *arXiv preprint arXiv:2403.03187*.

Jiaxin Bai and Yangqiu Song. 2025. Autoschemakg: Autonomous kg construction through dynamic schema induction from web-scae corpora. *arXiv preprint*.

Vincent-Pierre Berges, BArlas Oguz, Daniel Haziza, Wen-tau Yih, Luke Zettlemoyer, and Gargi Ghosh. 2024. Memory layers at scale. *arXiv preprint*.

Sebastian Borgeaud. 2022. Retro: Improving language models by retrieving from trillions of tokens. *arXiv preprint*.

Sindhuja Chaduvula, Ahmed Y. Radwan, Azib Farooq, Yani Ioannou, and Shaina Raza. 2026. Reducing hallucinations in llms via factuality-aware preference learning. *arXiv preprint*.

Manish Chandra, Debasis Ganguly, and I. Ounis. 2026. Lure-rag: Lightweight utility-driven reranking for efficient rag. *arXiv preprint arXiv:2601.19535*.

Haotian Chen, Qingqing Long, Siyu Pu, Xiao Luo, Wei Ju, Meng Xiao, Yuanchun Zhou, Jianghua Zhao, and Xuezhi Wang. 2026. Deepera: A deep evidence reranking agent for scientific rag qa. *arXiv preprint arXiv:2601.16478*.

Lida Chen, Zujie Liang, Xintao Wang, Jiaqing Liang, Yanghua Xiao, Feng Wei, Jinglei Chen, Zhenghong Hao, Bing Han, and Wei Wang. 2025a. Coke: Teaching large language models to express knowledge boundary from their own signals. In *KnowFM workshop*.

Mingda Chen, Yang LI, Karthik Padthe, Rulin Shao, Alicia Sun, Luke Zettlemoyer, Gargi Ghosh, and Wen-tau Yih. 2025b. Ewe: Improving factuality w. explicit working memory. In *ACL*.

Mingyang Chen, Linzhuang Sun, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. 2025c. Research: Learning to reason with seach for llms via rl. In *NeurIPS*.

Yiqun Chen, Lingyong Yan, Weiwei Sun, Xinyu Ma, Yi Zhang, Shuaiqiang Wang, Dawei Yin, Yiming Yang, and Jiaxin Mao. 2025d. Co-marl: Improving retrieval-augmented generation through multi-agent reinforcement learning. In *NeurIPS*.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? In *ICML*.

Cheng-Han Chiang and Hung-yi Lee. 2024. Merging facts, crafting fallacies: Evaluating the contradictory nature of aggregated factual claims in long-form generations. *arXiv preprint*.

Daniel Christoph, Max Ploner, Patrick Haller, and Alan Akbik. 2025. Learning to study for better factual recall. *arXiv preprint*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in llms. In *ICLR*.

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. 2024. I don't know: Explicit modeling of uncertainty with an [idk] token. In *NeurIPS*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Cove: Chain-of-verification reduces hallucination in llms. *arXiv preprint*.

Hanxing Ding, Shuchang Tao, Liang Pang, Zihao Wei, Liwei Chen, Kun Xu, Huawei Shen, and Xueqi Cheng. 2025. On the diminishing returns of complex robust rag training in the era of powerful llms. *SigIR-AP*.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024. Claim verification in the age of llms: A survey. *arXiv preprint*.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form llms. In *ACL*.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *NAACL*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Seven Truitt, and Jonothan Larson. 2024. From local to global: A

graph rag approach to query-focused summarization. *arXiv preprint*.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, Shizhu Liu, Pingchuan Tian, Yuping Wang, and Yuxuan Wang. 2023. Halo: Estimation and reduction of hallucinations in open-source weak llms. *arXiv preprint*.

Wenqi Fan, Yujuang Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards ra llm. *arXiv preprint*.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024a. Maat: Enhancing noise robustness of retrieval-augmented lms with adaptive adversarial training. In *Annual Meeting of the Association for Computational Linguistics*.

Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald. 2024b. Trace the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. *arXiv preprint*.

Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald. 2025. Kirag: Knowledge-driven iterative retriever for enhancing retrieval-augmented generation. *arXiv preprint*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. In *ACL*.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2025. Ragbench: Explainable benchmark for rag systems. *arXiv preprint*.

L. Gao, X. Ma, J. Lin, Callan, and J. 2023. Precise zero-shot dense retrieval without relevance labels. In *In ACL*.

Y. Gao and Y. Xiong. 2024. Retrieval-augmented generation for ai-generated content: A survey and tutorial. In *Tutorial at ACL, Bangkok, Thailand*.

Y. Gao, Y. Xiong, and X. Gao. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *EMNLP*.

Yuntao Gui and James Cheng. 2025. Search-r3: Unifying reasoning and embedding generation in llms. *arXiv preprint*.

Aman Gupta and Vinamra Benara. 2024. Rag vs fine-tuning: pipelines, tradeoffs, and a case study on agriculture. *arXiv preprint*.

Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. 2024. A comprehensive survey of rag: Evolution, current landscape, and future directions. *arXiv preprint*.

Bernal Jimenez Guti´errez´, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for llms. In *NeurIPS*.

Bernal Jimenez Guti´errez´, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. Hipporag2: From rag to memory: Non-parametric continual learning for llms. *arXiv preprint*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint*.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *EMNLP*.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2024. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? In *North American Chapter of the Association for Computational Linguistics*.

Yucheng Hu and Yuxing Lu. 2024. Rag and rau: A survey on ra lm in natural language processing. *arXiv preprint*.

L. Huang, W. Yu, and W. Ma. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Gautier Izacard and Edouard Grave. 2021. Fid: Leveraging passage retrieval with generative models for open domain question answering. In *EACL*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C. Park. 2024. Adaptive-rag: Learning to adapt ra llm through question complexity. In *North American Chapter of the Association for Computational Linguistics*.

Z. Ji, N. Lee, and R. Frieske. 2023. Survey of hallucination in natural language generation. In *ACM Computing Surveys, 55(12):1–38*.

Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, Han, and Jiawei. 2025. Deepretrieval: Hacking real search engines and retrievers w. llms via rl. In *COLM*.

Z. Jiang, F. Xu, and L. Gao. 2023a. Active retrieval augmented generation. In *In EMNLP*.

Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Flare: Active retrieval augmented generation. In *EMNLP*.

Bowen Jin, hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O. Arik, Dong Wang, Hamed Azmani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engies with rl. In *COLM*.

Saurav Kadavath, Tom Conerly, and ... Jared Kaplan. 2022. Language models (mostly) know what they know. *arXiv preprint*.

Nikhil Kandpal, haikang Deng, Adam Orberts, Eric Wallace, and Colin Raffel. 2023. Llms struggle to learn long-tail knowledge. In *ICML*.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah GOldblum, and Andrew Gordon Wilson. 2024. Llms must be taught to know what they don't know. In *NeurIPS*.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. knn-lm: Generalization through memorization: Nearest neighbor language models. In *ICLR*.

Jiyeon Kim, Hyunji Lee, Hyowon Cho, Joel Jang, Hyeonbin Hwang, Seungpil Won, Youbin Ahn, Dohaeng Lee, and Minjoon Seo. 2025. Knowledge entropy decay during lm pre-training hinders new knowledge acqusition. In *ICLR*.

Abdullatif Koksal, Renat Aksitov, and Chung-Ching Chang. 2023. Har: Hallucination augmented recitations for language models. *arXiv preprint*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in nlg. In *ICLR*.

Michael Lan, Philip Torr, Austin Meek, David Krueger, and Fazl Barez. 2024. Sparse autoencoders reveal universal feature spaces for hallucination detection. *arXiv preprint*.

Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. 2023. Conli: Chain of natural language inference for reducing llm ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*.

Yongqi Leng, Yikun Lei, Xikai Liu, Meizhi Zhong, Bojian Xiong, Yurong Zhang, Yan Gao, Yao Hu, and Deyi Xiong. 2025. Decision-execution separation for retrieval-augmented generation as mdp. *arXiv preprint*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktaschel, Sebastian Riedel, and Douwe Kiela. 2020. Rag: Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeuIPS*.

Bo Li, Tian Tian, Zhenghua Xu, Hao Cheng, Shikun Zhang, and Wei Ye. 2025a. Entropy-trend constraint for proactive retrieval triggering. *arXiv preprint*.

Bo Li, Tian Tian, Zhenghua Xu, Hao Cheng, Shikun Zhang, and Wei Ye. 2025b. Etc: Modeling uncertainty trends for timely retrieval in dynamic rag. *arXiv preprint*.

Jiaqi Li, Yixuan Tang, and Yi Yang. 2025c. Ustuning: Know the unknown: An uncertainty-sensitive method for llm instruction tuning. In *ACL*.

Kenneth Li, Oam Patel, Fernanda Viegas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention (iti): Eliciting truthful answers from a lm. In *NeurIPS*.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025d. [c] search-o1: Agentical search-enhanced large reasoning models. *arXiv preprint*.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2024a. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. In *International Conference on Learning Representations*.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiouzhu Mei, and Michael Bendersky. 2024b. Self-route: Rag or long-context llms? a comprehensive study and hybrid approach. In *EMNLP*.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for llms. *arXiv preprint*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint*.

Xiaoqiang Lin, Aritra Ghosh, K. H. Low, Anshumali Shrivastava, and Vijai Mohan. 2025. Refrag: Rethinking rag based decoding. *arXiv preprint arXiv:2509.01092*.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. 2024a. Think-on-graph 2.0: Deep and faithful llm reasoning w. knowledge-guided rag. In *International Conference on Learning Representations*.

Zexiong Ma, Shengnan An, Zeqi Lin, Yanzhen Zou, Jian-Guang Lou, and Bing Xie. 2024b. Depac: Enhancing retrieval-augmented generation with dehallucinating parallel context extension. In *Long-Context Fnd workshop*.

E. Magesh and D. Kvasyuk. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *Stanford HAI Working Paper*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust lms: Investigating effectiveness of parametric and non-parametric memories. In *ACL*.

K. Meng, D. Bau, A. Andonian, Belinkov, and Y. 2022. Locating and editing factual associations in gpt. In *In NeurIPS*.

Dezhuang Miao, Yibin Du, Xiang Li, Xiaoming Zhang, Jiahe Li, Bo Zhang, Bingyu Yan, Lian Zhang, and Litian Zhang. 2025. R-cot: Reverse cot and causal path verification: A modular plugin for aligning llms with kgs. In *CIKM*.

Sabrina J. Mielke, Anthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *arXiv preprint*.

Dehai Min, Kailin Zhang, Tongtong Wu, and Lu Cheng. 2025. Quantifying uncertainty from the pre-training corpus for retrieval-augmented generation. *arXiv preprint*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *COLM*.

N. Muennighoff, H. Su, and L. Wang. 2024. Generative representational instruction tuning. In *In ICLR*.

D. Muhlgay, O. Ram, and I. Magar. 2024. Generating benchmarks for factuality evaluation of language models. In *In EACL*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint*.

Bo Ni, Zheyuan Liu, Leyao Wang, ... Luna Dong, Yinglong Xia, ... Wenqi Fan, Erik Blasch, Yu Wang, Meng Jiang, and Tyler Derr. 2025. Towards trustworthy rag for llms: A survey. *arXiv preprint*.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. When do llms need ra? mitigating llms' overconfidence helps ra. In *ACL*.

Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2023. Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint*.

Xu Pan, Ely Hahami, Zechen Zhang, and Haim Sompolinsky. 2025. Mega: Memorization and knowledge injection in gated llms. *arXiv preprint*.

Core Francisco Park, Zechen Zhang, and Hidenori Tanaka. 2025. Sys-2-ft: New news: System-2 fine-tuning for robust integration of new knowledge. *arXiv preprint*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving llms with external knowledge and automated feedback. *arXiv preprint*.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. 2024. Graph rag: A survey. *arXiv preprint*.

F. Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktaschel, and Sebastian Riedel. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *NAACL*.

Yifu Qiu, Varun Embar, Shay B. Cohen, and Benjamin Han. 2023a. Tweak: Think while you write: Hypothesis verfication promotes faithful knowledge-to-text generation. *arXiv preprint*.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. 2023b. Detecting and mitigating hallucinations in multilingual summarization. *arXiv preprint*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context ralm: In-context retrieval-augmented language models. *TACL*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S. M. Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in llms – an extensive definition, quantification, and prescriptive remediations. In *EMNLP*.

Evgeniia Razumovskaia, Ivan Vulic, Pavle Markovic, Tomasz Cichy, Qian Zheng, Tsung-Hsien Wen, Budzianowski, and Pawel. 2023. Dial beinfo for faithfulness: Improving factuality of information-seeking dialogue via behavioural fine-tuning. *arXiv preprint*.

Revanth Gangi Reddy and Chengxiang Zhai. 2024. Charmbana: Progressive responses with real-time internet searchfor knowledge-powered conversations. In *WSDM*.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2025. Investigating the factual knowledge boundary of llms with retrieval augmentation. In *Coling*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? by roberts et al. In *EMNLP*.

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint*.

H. Shi and X. Qiu. 2024. Adaptive context-aware decoding for faithful rag. *arXiv preprint*.

Weijia Shi, Xiaochuang Han, M. Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and S. Yih. 2023a. Why lms hallucinate? In *North American Chapter of the Association for Computational Linguistics*.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2024. In-context pretraining: Language modeling beyond document boundaries. In *ICLR*.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint*.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Seeker: Language models that seek for knowledge: Modular search generation for dialogue and prompt completion. In *EMNLP*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *EMNLP*.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. *arXiv preprint*.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic rag based on the information needs of llms. In *ACL*.

J. Sun, C. Xu, and L. Tang. 2024a. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *In ICLR*.

Jiashuo Sun, Chengjin Xun, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. 2024b. Think-on-graph: Deep and responsible reasoning of llm on kg. In *ICLR*.

Kai Sun, Y. Xu, Hanwen Zha, Yue Liu, and Xinhsuai Dong. 2023a. Head-to-tail: How knowledgeable are large language models? a.k.a. will llms replace knowledge graphs? In *North American Chapter of the Association for Computational Linguistics*.

Yushi Sun, Kai Sun, Yifan Ethan Xu, Xiao Yang, Xin Luna Dong, Nan Tang, and Lei Chen. 2025. Kerag: Knowledge-enhanced retrieval-augmented generation for advanced question answering. In *EMNLP*.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Dnny Zhou. 2023b. Recite: Recitation-augmented language models. In *ICLR*.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Evaluation of chatgpt as a question answering system for answering complex questions. *arXiv preprint*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D. Manning, and Chelsea Finn. 2023a. Fine-tuning language models for factuality. In *NeurIPS workshop*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023b. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint*.

Zhiliang Tian, Wei Bi, Xiaopeng Li, and N. Zhang. 2019. Learning to abstract for memory-augmented conversational response generation. In *Annual Meeting of the Association for Computational Linguistics*.

S.M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinjia Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. "a comprehensive survey of hallucintion mitigation techniques in llms". *arXiv preprint*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Ircot: Iterleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint*.

Lynn Vonderhaar, Daniel A. Machado, and Omar Ochoa. 2025. Rag for llms: A survey. In *International Conference on Artificial Intelligence Testing*.

Tu Vu, Mohit Ivyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, quoc Le, and Thang Luong. 2023. Freshllms: Refreshing llms with search engine augmentation. *arXiv preprint*.

Jiaqi Wang, Xiao Yang, Kai Sun, Parth Suresh, Sanat Sharma, Adam Czyzewski, Derek Andersen, Surya Appini, Arkav Banerjee, Sajal Choudhary, Shervin Ghasemlou, Ziqiang Guan, Akil Iyer, Haidar Khan, Lingkun Kong, Roy Luo, Tiffany Ma, Zhen Qiao,

David Tran, and 22 others. 2025a. Crag-mm: Multi-modal multi-turn comprehensive rag benchmark. In *arXiv*.

Ke Wang, Yiming Qin, Nikolaos Dimitriadis, Alessandro Favero, and Pascal Frossard. 2025b. Memoir: Lifelong model editing with minimal overwrite and informed retention for llms. *arXiv preprint arXiv:2506.07899*.

X. Wang, J. Wei, and D. Schuurmans. 2023a. Self-consistency improves chain of thought reasoning in language models. In *In ICLR*.

Yu Wang, Yifan Gao, Xiusi Chen, Haoming Jiang, Shiyang LI, Jingfeng Yang, Qingyu Yin, Zheng Li, Xian LI, Bing Yin, JIngbo Shang, and Julian McAuley. 2024. Memoryllm: Towards self-updatable llms. In *ICML*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023b. Filco: Learning to filter context for rag. *arXiv preprint*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024a. Simpleqa: Measuring short-form factuality in llms. *arXiv preprint*.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024b. Safe: Long-form factuality in llms. *arXiv preprint*.

Zhepei Wei, Wei-lin Chen, and Yu Meng. 2025. Instructrag: Instructing retrieval augmented generation via self-synthesized rationals. In *ICLR*.

Di Wu, Jia-Chen Gu, Kai-Wei Chang, and Nanyun Peng. 2025. Self-routing rag: Binding selective retrieval with knowledge verbalization. *arXiv preprint arXiv:2504.01018*.

Yiqing Xie, Wenxuan Zhou, Pradyot Prakash, Di Jin, Yuning Mao, Quintin Fettes, Arya Talebzadeh, Sinong Wang, Han Fang, Carolyn Rosé, Daniel Fried, and Hejia Zhang. 2025. Fence: Improving model factuality w. fine-grained critic-based evaluator. In *ACL*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*.

Hao-Xiang Xu, Jun-Yu Ma, Ziqi Peng, Yuhao Sun, Zhen-Hua Ling, and Jia-Chen Gu. 2026. Multiplicative orthogonal sequential editing for scalable knowledge updates. *arXiv preprint*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Crag: Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, and 8 others. 2024a. Crag–comprehensive rag benchmark. In *NeurIPS*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024b. Alignment for honesty. In *NeurIPS*.

S. Yao, D. Yu, and J. Zhao. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *In NeurIPS*.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. 2024. Characterizing truthfulness in llm geneations w. local intrinsic dimension. *arXiv preprint*.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024a. Evaluation of rag: A survey. *arXiv preprint*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024b. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *EMNLP*.

Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. 2022. A survey of knowledge-enhanced text generation. *arXiv preprint*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng JI, and Tong Zhang. 2024a. R-tuning: Instructint llms to say "i don't know". In *NAACL*.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023. Sac3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *EMNLP*.

Jie Zhang, Bo Tang, Wanzi Shao, Wenqiang Wei, Jihao Zhao, Jianqing Zhu, Zhiyu Li, Wen Xi, Zehao Lin, Feiyu Xiong, and Yanchao Tan. 2025. Tadarag: Task adaptive rag via on-the-fly kg construction. *arXiv preprint*.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, M. Zaharia, Ion Stoica, and Joseph Gonzalez. 2024b. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.

Penghao Zhao, Hailin Zhang, Jie Jiang, and Bin Cui. 2024. Rag for ai-generated content: A survey. *arXiv preprint*.

Yi Zhao, Yanxiang He, Xiaotong Zhang, and Weidong Wen. 2025. Combining retrieved with generated contexts via a listwise reranker for open-domain question answering. *Neurocomputing*.

H. Zheng, S. Mishra, and X. Chen. 2024. Take a step back: Evoking reasoning via abstraction in large language models. In *In ICLR*.

Zexuan Zhong, Tao Lei, and Danqi Chen. 2022. Trime: Training with in-batch memory. In *EMNLP*.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *EMNLP*.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S. Yu. 2024. Trustworthiness in rag systems: A survey. *arXiv preprint*.

Zijian Zhou, Ao Qu, Zhaoxuan Wu, Sunghwan Kim, Alok Prakash, Daniela Rus, Jinhua Zhao, Bryan Kian Hsiang Low, and Paul Pu Liang. 2025. Mem1: Learning to synergize memory and reasoning for efficient long-horizon agents. *arXiv preprint*.

Yijie Zhu, Haojie Zhou, Wanting Hong, Tailin Liu, and Ning Wang. 2025. Reap: Enhancing rag with recursive evaluation and adaptive planning for multi-hop qa. *arXiv preprint*.