Unveiling Knowledge Boundary of Large Language Models for Trustworthy Information Access

Yang Deng
Singapore Management University
Singapore
ydeng@smu.edu.sg

Moxin Li
National University of Singapore
Singapore
limoxin@u.nus.edu

Liang Pang
Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China
pangliang@ict.ac.cn

Wenxuan Zhang
Singapore University of Technology
and Design
Singapore
wxzhang@sutd.edu.sg

Wai Lam
The Chinese University of Hong Kong
Hong Kong SAR, China
wlam@se.cuhk.edu.hk

Abstract

Large Language Models (LLMs) have emerged as powerful tools for generating content and facilitating information seeking across diverse domains. While their integration into conversational systems opens new avenues for interactive information-seeking experiences, their effectiveness is constrained by their knowledge boundaries—the limits of what they know and their ability to provide reliable, truthful, and contextually appropriate information. Understanding these boundaries is essential for maximizing the utility of LLMs for real-time information seeking while ensuring their reliability and trustworthiness. In this tutorial, we will explore the taxonomy of knowledge boundary in LLMs, addressing their handling of uncertainty, response calibration, and mitigation of unintended behaviors that can arise during interaction with users. We will also present advanced techniques for optimizing LLM behavior in generative information-seeking tasks, ensuring that models align with user expectations of accuracy and transparency. Attendees will gain insights into research trends and practical methods for enhancing the reliability and utility of LLMs for trustworthy information access.

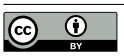
CCS Concepts

• **Computing methodologies** → *Natural language generation*; • **Information systems** → *Users and interactive retrieval*; *Information retrieval query processing*.

Keywords

Trustworthy Information Access, Large Language Model, Knowledge Boundary, Retrieval-augmented Generation

ACM Reference Format:
Yang Deng, Moxin Li, Liang Pang, Wenxuan Zhang, and Wai Lam. 2025. Unveiling Knowledge Boundary of Large Language Models for Trustworthy Information Access. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3726302.3731684



SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3731684

1 Presenters

Yang Deng is an Assistant Professor at Singapore Management University. His research lies in information retrieval (IR) and natural language processing (NLP), especially for trust and reliability in LLMs. He has published over 60 papers on relevant topics at top venues such as SIGIR, WWW, ACL, EMNLP, ICLR, TOIS, TKDE, and serves as Area Chairs. He received the Google Southeast Asia Research Awards in 2024 for his excellent research on trustworthy AI. He has rich experience in organizing tutorials at top conferences, including SIGIR 2024, WWW 2024, and ACL 2023.

Moxin Li is a final-year PhD candidate at National University of Singapore. Her research focuses on IR and NLP, especially for LLM trust and evaluation. She has published over 10 papers at top conferences including SIGIR, WWW, ACL, EMNLP, etc.

Liang Pang is an Associate Professor at the Institute of Computing Technology, Chinese Academy of Sciences. He is renowned for his expertise in trustworthy LLMs and text matching in IR. He has published about 60 papers in top journals and conferences, including SIGIR, WWW, ACL, EMNLP, etc, and received the Best Paper Runner-up of CIKM 2017, the Best Paper Honorable Mention of SIGIR 2024. He has delivered multiple tutorials at SIGIR 2021, WSDM 2021, and KDD 2024.

Wenxuan Zhang is an Assistant Professor at Singapore University of Technology and Design. Before this, he was a research scientist at Alibaba DAMO Academy, Singapore. His primary research areas are NLP and trustworthy AI, with a special aim to advance inclusive NLP, supporting diverse languages and cultures. He has published over 40 papers in top-tier conferences and journals, including SIGIR, WWW, ICLR, NeurIPS, ACL, EMNLP. He also regularly serves on the (senior) program committees of multiple leading conferences and journals. He organized a tutorial at IJCAI 2023.

Wai Lam is a Professor at the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, where he is also the Division Head and Department Chairman. His research interests include intelligent information retrieval and text mining. He has authored or co-authored more than 100 papers

in premier conferences and journals (SIGIR, WWW, ACL, EMNLP, etc). He regularly serves as the Senior PC or Area Chair of these conferences. He receives multiple paper awards, including ACL 2021 Outstanding Paper Awards, EACL 2023 Outstanding Paper Awards, and ACL 2024 Area Chair Awards.

2 Motivation

Understanding the knowledge boundaries of large language models (LLMs) is critical for developing safe, reliable, and effective generative information-seeking systems. These boundaries define the scope and limitations of what an LLM knows, ensuring users receive accurate and trustworthy responses rather than fabricated information [9]. By identifying these limits, systems can proactively acknowledge uncertainties, avoid disseminating misinformation, and guide users toward verified sources when queries exceed their expertise. This awareness enhances user trust and reliability, particularly in high-stakes domains like healthcare or law, where errors carry significant consequences. Additionally, studying knowledge boundaries informs model improvements, highlighting gaps in training data or reasoning capabilities, enabling targeted updates for better alignment with real-world needs. It also equips systems to handle ambiguous, emerging, or evolving topics by signaling when information may be outdated or incomplete. Ultimately, mapping these frontiers ensures generative systems operate responsibly while evolving to meet dynamic information demands.

3 Overview and Objectives

Specifically, this tutorial aims to discuss three key research questions (**RQs**) surrounding the knowledge boundary of large language models (LLMs) for trustworthy information assess:

- **RQ1**: *What is the knowledge boundary of LLMs?*
- **RQ2**: *Why do we care about the knowledge boundary of LLMs in generative information-seeking systems?*
- **RQ3**: *How can the knowledge boundary of LLMs be identified?*
- **RQ4**: *How can issues caused by knowledge boundaries be mitigated for trustworthy information access?*

3.1 Taxonomy of Knowledge Boundary (RQ1)

We will first introduce different taxonomy of knowledge boundary in the era of LLMs and then discuss their limitations for shedding light on a more comprehensive taxonomy. A widely-adopted taxonomy of the knowledge for language models (LMs) in the literature [4, 51] is derived from the Uncertainty Matrix [33], also called Known-Unknown Quadrant. The knowledge is categorized based on two factors: (1) whether the model has corresponding parametric knowledge, and (2) whether the model is aware of the first factor. Another taxonomy [50] formally defines three different types of the knowledge for LMs, from the perspective of the model's mastery of knowledge. Our recent survey [24] introduces a more complete and formalized taxonomy of the knowledge for LLMs, including

- ***Outward Knowledge Boundary*** defines the observable knowledge boundary for a specific LLM.
- ***Parametric Knowledge Boundary*** defines the abstract knowledge boundary for a specific LLM.
- ***Universal Knowledge Boundary*** defines the whole set of knowledge known to human.

3.2 Undesired Behaviours (RQ2)

Due to the unawareness of knowledge boundary, LLM-based generative information-seeking systems exhibit several types of undesired behaviors that can compromise the reliability and utility of their outputs for out-of-boundary user queries. In this part, we will introduce the latest empirical analysis and findings of these undesired behaviors:

- **Factuality Hallucination**, *i.e.*, the model output diverges from real-world facts, typically stem from the following causes. 1) Deficiency in domain-specific knowledge: LLMs, primarily trained on broad datasets, often lack detailed knowledge in specialized domains, leading to inaccuracies in domain-specific queries [12, 32]. 2) Outdated knowledge: Without mechanisms to update their internal knowledge, LLMs struggle to adapt to new developments, often resorting to fabricating facts or using outdated responses [20, 30]. 3) Over-confidence in unknown knowledge: LLMs often show overconfidence when addressing topics beyond their knowledge, delivering assertive but incorrect responses [4, 16].
- **Untruthful Responses Misled by Contexts**. Even though LLMs possess the required knowledge for the user query, they often produce untruthful responses when misled by context, which occurs in two forms: *untruthful context*, where the context includes false or misleading information [31], and *irrelevant context*, where extraneous details divert the model from generating precise responses [11, 35].
- **Truthful but Undesired Outputs**. LLMs sometimes produce accurate yet improper responses when handling certain queries, leading to answers misaligned with user expectations, such as random responses to ambiguous queries [15, 54].

Each poses unique challenges for the effective deployment of LLMs in real-world generative information-seeking systems.

3.3 Knowledge Boundary Identification (RQ3)

Knowledge boundary identification is a crucial step in understanding the limitations of LLMs. We will introduce three key approaches to this challenging problem, including: 1) **Uncertainty Estimation** aims to measure the model uncertainty towards the given knowledge-related queries by token probability-based method [17] and semantic-based method [21, 29]; 2) **Calibration** aims to estimate model confidence on the correctness of their responses by prompting for eliciting confidence [18, 37] or expressing confidence [23, 40], and training additional models as calibrator [34, 38]; 3) **Internal State Probing** aims to fine-tune linear probes on LLM representations to predict LLM answer accuracy [6, 27].

3.4 Out-of-Boundary Query Mitigation (RQ4)

In this part, we will introduce the cutting-edge approaches on mitigating the knowledge gap issue when LLM-based generative information-seeking systems face user queries that exceed different types of knowledge boundaries.

3.4.1 ***Queries Exceeding Outward Knowledge Boundary.*** The answers to this type of queries are sensitive to the form of the user query prompt fed into the LLM. Therefore, although the model possesses corresponding parametric knowledge, sometimes it may make untruthful responses without proper queries or contexts [50].

Query optimization has become a main-stream approach to better elicit this type of knowledge from LLMs for generative information seeking, which can be roughly categorized into *prompt optimization* and *demonstration optimization*. The prompt optimization includes training-free methods such as search-based techniques [42, 56] and adopting LLM as optimizer [47], and training-based methods typically involving reinforcement learning to train prompt optimization modules [14, 55]. The demonstration optimization methods involve various ways to select effective demonstrations such as considering their similarity to the test example and their diversity [26, 31]. The order of demonstrations also largely affect the performance [7].

3.4.2 ***Queries Exceeding Parametric Knowledge Boundary.*** The user queries exceeding parametric knowledge boundary are unable to be answered by the specific LLM, but the query itself is answerable. To mitigate this knowledge gap, researchers investigate different **Retrieval-augmented Generation (RAG)** approaches to supplement necessary knowledge beyond the pre-trained knowledge of LLMs. This enables the model to provide accurate answers to domain-specific or real-time queries that it cannot answer based solely on its pre-trained parameters. We will discuss various approaches, including retriever-enhanced models [36, 45], generator-enhanced models [46, 52], and interaction-enhanced models [44, 48]. These methods are crucial for enhancing the factual accuracy and adaptability of LLMs in specialized or evolving domains.

3.4.3 ***Queries Exceeding Universal Knowledge Boundary.*** This type of query itself is unanswerable, regardless of any prompt or any LLM. Therefore, directly responding to this type of query will result in factuality hallucinations or undesired outputs.

- **Refusal** When faced with user queries involving model-agnostic unknown knowledge, a straight-forward approach is to have LLMs honestly express their knowledge limitations by refusing to answer. We will introduce the latest studies on teaching LLMs to say "I don't know", including prompt-based [4, 43] and fine-tuning-based approaches [16, 53].
- **Asking Clarification Questions** avoids providing direct answers when uncertain. Instead, the LLM gives users an opportunity to further clarify their queries [2, 3]. Recent studies develop various training paradigms to teach LLMs to ask clarification questions, such as in-context learning [15], self-learning [5], reinforcement learning [10], and contrastive learning [8].

3.5 Open Challenges and Prospects

In the last part, we will discuss the open challenges in investigating the knowledge boundary of LLMs and potential future directions.

- **Interpretability of Knowledge Boundary** One of the key challenges lies in making the knowledge boundary of LLMs interpretable to users. Developing methods for localizing [27] or explaining [28] the boundary in human-readable ways is essential to improve trust and user interaction with LLMs.
- **Generalization of Knowledge Boundary** While a model might be able to identify its knowledge gaps in a specific domain, generalizing this capability across different languages [1] and application domains [22] remains difficult. Future research need to develop more adaptive and scalable techniques that can generalize knowledge boundary detection or mitigation across domains.
- **Utilization of Knowledge Boundary** Estimating and informing an LLM of its knowledge boundary should not be the final step. Instead, recognizing the knowledge gap can be further leveraged to enhance the LLM's utilities in areas where the knowledge is lacking, such as reducing costs in RAG [49] and facilitating self-improvement [41].

4 Other Information

Relevance to IR. The integration of LLMs is a trending topic across various IR applications, including but not limited to search engine, recommender systems, and conversational systems. LLM-based conversational systems (*e.g.*, ChatGPT) have revolutionized our daily information seeking paradigms. However, the trust and reliability is a key issue for applying LLMs into real-world IR applications, especially for information-seeking systems. Several tutorials about the information seeking systems have been given in related top-tier conferences, including but not limited to 1) WWW 2024 / SIGIR 2024 - *Recent Advances in Generative Information Retrieval* [39], 2) SIGIR 2022 / WWW 2023 - *Conversational Information Seeking: Theory and Application* [13], and 3) SIGIR 2020 - *Recent Advances in Conversational Information Retrieval* [19]. However, these tutorials mainly introduce advanced designs for building information seeking systems. In our tutorial, we aim to elaborate a comprehensive introduction to cutting-edge research on the knowledge boundary of LLMs and shed light on building LLM-based generative information-seeking systems for trustworthy information access.

Detailed Schedule. The following summarizes the detailed schedule of the tutorial:

(1) Introduction and Motivations [30 min]
(3) Taxonomy of Knowledge Boundary [30 min]
(4) Knowledge Boundary Identification [30 min]
(5) Out-of-Boundary Query Mitigation [60 min]
(6) Open Challenges and Beyond [30 min]

Supporting Materials. (1) Slides will be made publicly available; (2) The tutorial is accompanied with two comprehensive surveys [24, 25] on this topic; and (3) A github repo will be maintained for annotated references.

Acknowledgment

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (No. MSS24C004, No. MSS24C012).

References

[1] Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat, and Monojit Choudhury. 2022. On the Calibration of Massively Multilingual Language Models. In *EMNLP 2022*.
[2] Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and Evaluating Open-Domain Dialogue Corpora with Clarifying Questions. In *EMNLP 2021*.
[3] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking Clarifying Questions in Open-Domain Information-Seeking Conversations. In *SIGIR 2019*. ACM, 475–484.
[4] Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. 2024. Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with Large Language Models. In *Findings of the ACL, ACL 2024*.
[5] Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2024. STaR-GATE: Teaching Language Models to Ask Clarifying Questions. In *COLM 2024*.

[6] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. In *Findings of the ACL: EMNLP 2023*.

[7] Liang CHEN, Li Shen, Yang Deng, Xiaoyan Zhao, Bin Liang, and Kam-Fai Wong. 2025. PEARL: Towards Permutation-Resilient LLMs. In *ICLR 2025*.

[8] Maximillian Chen, Ruoxi Sun, Sercan Ö. Arik, and Tomas Pfister. 2025. Learning to Clarify: Multi-turn Conversations with Action-Based Contrastive Self-Training. In *ICLR 2025*.

[9] Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. 2023. Hallucination Detection: Robustly Discerning Reliable Answers in Large Language Models. In *CIKM 2023*.

[10] Yue Chen, Chen Huang, Yang Deng, Wenqiang Lei, Dingnan Jin, Jia Liu, and Tat-Seng Chua. 2024. STYLE: Improving Domain Transferability of Asking Clarification Questions in Large Language Model Powered Conversational Agents. In *Findings of the ACL, ACL 2024*.

[11] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *SIGIR 2024*. ACM, 719–729.

[12] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (06 2024), 64–93. https://doi.org/10.1093/jla/laae003

[13] Jeffrey Dalton, Sophie Fischer, Paul Owoicho, Filip Radlinski, Federico Rossetto, Johanne R. Trippas, and Hamed Zamani. 2022. Conversational Information Seeking: Theory and Application. In *SIGIR 2022*.

[14] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing Discrete Text Prompts with Reinforcement Learning. In *EMNLP 2022*.

[15] Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023. Prompting and Evaluating Large Language Models for Proactive Dialogues: Clarification, Target-guided, and Non-collaboration. In *Findings of the ACL: EMNLP 2023*.

[16] Yang Deng, Yong Zhao, Moxin Li, See-Kiong Ng, and Tat-Seng Chua. 2024. Don't Just Say "I don't know"! Self-aligning Large Language Models for Responding to Unknown Questions with Explanations. In *EMNLP 2024*.

[17] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting Attention to Relevance: Towards the Uncertainty Estimation of Large Language Models. In *ACL 2024*.

[18] 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 2024*.

[19] Jianfeng Gao, Chenyan Xiong, and Paul Bennett. 2020. Recent Advances in Conversational Information Retrieval. In *SIGIR 2020*.

[20] Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. RealTime QA: What's the Answer Right Now?. In *NeurIPS 2023*.

[21] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *ICLR 2023*.

[22] Moxin Li, Wenjie Wang, Fuli Feng, Yixin Cao, Jizhi Zhang, and Tat-Seng Chua. 2023. Robust Prompt Optimization for Large Language Models Against Distribution Shifts. In *EMNLP 2023*.

[23] Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024. Think Twice Before Trusting: Self-Detection for Large Language Models through Comprehensive Answer Reflection. In *Findings of ACL: EMNLP 2024*.

[24] Moxin Li, Yong Zhao, Yang Deng, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, and Tat-Seng Chua. 2024. Knowledge Boundary of Large Language Models: A Survey. *CoRR* abs/2412.12472 (2024).

[25] Siheng Li, Cheng Yang, Taiqiang Wu, Chufan Shi, Yuji Zhang, Xinyu Zhu, Zesen Cheng, Deng Cai, Mo Yu, Lemao Liu, Jie Zhou, Yujiu Yang, Ngai Wong, Xixin Wu, and Wai Lam. 2024. A Survey on the Honesty of Large Language Models. arXiv:2409.18786 [cs.CL]

[26] Xiaonan Li and Xipeng Qiu. 2023. Finding Support Examples for In-Context Learning. In *Findings of the ACL: EMNLP 2023*.

[27] Junteng Liu, Shiqi Chen, Yu Cheng, and Junxian He. 2024. On the Universal Truthfulness Hyperplane Inside LLMs. In *EMNLP 2024*.

[28] Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. 2023. Cognitive Dissonance: Why Do Language Model Outputs Disagree with Internal Representations of Truthfulness?. In *EMNLP 2023*.

[29] Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel Language Entropy: Fine-grained Uncertainty Quantification for LLMs from Semantic Similarities. *arXiv preprint arXiv:2405.20003* (2024).

[30] Yasumasa Onoe, Michael J. Q. Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity Cloze By Date: What LMs Know About Unseen Entities. In *Findings of the ACL: NAACL 2022*.

[31] Andrew Parry, Debasis Ganguly, and Manish Chandra. 2024. "In-Context Learning" or: How I learned to stop worrying and love "Applied Information Retrieval". In *SIGIR 2024*. ACM, 14–25.

[32] Weicong Qin, Zelin Cao, Weijie Yu, Zihua Si, Sirui Chen, and Jun Xu. 2024. Explicitly Integrating Judgment Prediction with Legal Document Retrieval: A Law-Guided Generative Approach. In *SIGIR 2024*. ACM, 2210–2220.

[33] Donald Rumsfeld. 2002. Defense. gov News Transcript: DoD News Briefing–Secretary Rumsfeld and Gen. Myers, United States Department of Defense (defense. gov)". *February* 12 (2002), 11.

[34] Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory Wornell, and Soumya Ghosh. 2024. Thermometer: Towards Universal Calibration for Large Language Models. *arXiv preprint arXiv:2403.08819* (2024).

[35] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *ICML 2023*.

[36] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2024. REPLUG: Retrieval-Augmented Black-Box Language Models. In *NAACL 2024*.

[37] Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2024. Llamas Know What GPTs Don't Show: Surrogate Models for Selective Classification.

[38] Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2024. LACIE: Listener-Aware Finetuning for Confidence Calibration in Large Language Models. *arXiv preprint arXiv:2405.21028* (2024).

[39] Yubao Tang, Ruqing Zhang, Weiwei Sun, Jiafeng Guo, and Maarten de Rijke. 2024. Recent Advances in Generative Information Retrieval. In *WWW 2024*.

[40] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *EMNLP 2023*.

[41] Jianing Wang, Yang Zhou, Xiaocheng Zhang, Mengjiao Bao, and Peng Yan. 2024. Self-Evolutionary Large Language Models through Uncertainty-Enhanced Preference Optimization. *arXiv preprint arXiv:2409.11212* (2024).

[42] Xinyuan Wang, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric Xing, and Zhiting Hu. 2024. PromptAgent: Strategic Planning with Language Models Enables Expert-level Prompt Optimization. In *ICLR 2024*.

[43] Bingbing Wen, Bill Howe, and Lucy Lu Wang. 2024. Characterizing LLM Abstention Behavior in Science QA with Context Perturbations. *arXiv preprint arXiv:2404.12452* (2024).

[44] Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2024. Search-in-the-Chain: Interactively Enhancing Large Language Models with Search for Knowledge-intensive Tasks. In *WWW 2024*.

[45] Shicheng Xu, Liang Pang, Jun Xu, Huawei Shen, and Xueqi Cheng. 2024. List-aware Reranking-Truncation Joint Model for Search and Retrieval-augmented Generation. In *WWW 2024*.

[46] Shicheng Xu, Liang Pang, Mo Yu, Fandong Meng, Huawei Shen, Xueqi Cheng, and Jie Zhou. 2024. Unsupervised Information Refinement Training of Large Language Models for Retrieval-Augmented Generation. In *ACL 2024*.

[47] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large Language Models as Optimizers. In *ICLR 2024*.

[48] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *ICLR 2023*.

[49] Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. 2024. SeaKR: Self-aware Knowledge Retrieval for Adaptive Retrieval Augmented Generation. *arXiv preprint arXiv:2406.19215* (2024).

[50] Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. 2024. Benchmarking Knowledge Boundary for Large Language Models: A Different Perspective on Model Evaluation. In *ACL 2024*.

[51] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do Large Language Models Know What They Don't Know?. In *Findings of the ACL: ACL 2023*.

[52] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In *ICLR 2024*.

[53] Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-Tuning: Instructing Large Language Models to Say 'I Don't Know'. In *NAACL 2024*.

[54] Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024. CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models. In *ACL 2024*.

[55] Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E. Gonzalez. 2023. TEMPERA: Test-Time Prompt Editing via Reinforcement Learning. In *ICLR 2023*.

[56] Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large Language Models are Human-Level Prompt Engineers. In *ICLR 2023*.