Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29620

Deep Learning and Natural Language Processing Research: Technological Evolution and Frontier Exploration of Hallucination Problems

Nuo Chen

Faculty of Science, Dalhousie University, Halifax, Canada
nuo09883@gmail.com

Abstract. With the virtue of large language models (LLMs) being applied in a growing number of fields, from text generation to medical support and financial analysis, the issue of "hallucination" has gained more and more recognition. For the given instances, "hallucination" can be explained with respect to artificial intelligence outputs that imitate coherent and convincing statements regardless of the underlying fact. Continuity of such lapses not only can undermine credibility in LLMs, but also may catalyze problems in numerous strategies, such as law, health care, or education. This paper provides a critical analysis of the currently available methods of preventing hallucinations in LLMs by outlining the retrieval-augmented generation (RAG) technique, verification frameworks, and planning-based strategies. The paper particularly deals with the TruthX reformulation displayed at ACL 2024, which essentially means redefining the meaning of factualism via representation editing. This dialogue is rounded off by stressing the ongoing problems and future growth routes, while suggesting that multiple methods, human cooperation, and efficient representation control altogether can lay the foundation for many more faithful and traceable language models.

Keywords: Large language model, Hallucination, Retrieval-augmented generation, Human-machine collaboration, Representation editor

1. Introduction

The emergence of artificial intelligence text generators (AIGC) has heralded the new age of intelligent production based on the abundance of good language models (LLMs). Since the activity of machine models like GPT-3, ChatGPT, and PaLM, they have been extensively employed in text generation for learning, coding assistance, medical diagnosis, and public opinion analysis. However, the challenge of model hallucination – producing seemingly fluent and logically well-founded statements devoid of factual data – has been shown to be the main problem fueling the distrust and hampering the use of most LLMs. From the neural machine translation era (NMT), Koehn and Knowles observed that neural approaches equate translation queries to 'fictitious' ones, resulting in generating erroneous translations [1]. Raunak et al. revealed that these unnatural outputs are systematic and determined by the type of preference midstories stored within the model. Thanks to

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ29620

the scaling of model size and the training data growth, the hallucination problem has been more widespread and more frequently goes unnoticed in practice. In modern LLMs, the parameter numbers of which can be translated to tens or even hundreds of billions, hallucinations may pass as being convincing and factually consistent, though, in practice, as soon as they are used within business contexts where real decisions are made, they pose major threats. In the past few years, the research area has shifted its focus from how to identify and understand hallucination in generative models towards systematically developing techniques that could achieve hallucination avoidance. For example, Lin et al. suggested TruthfulQA, which is a benchmark that serves the purpose of evaluating the capacity of language models to avoid the issuance of believable falsehoods [2]. These findings revealed that larger language models among the generation of hallucinations with higher confidence, which explains why construction interventions are requested. In its place, RAG ground outputs on retrieved evidence from external databases, hence capable of curbing factual errors in open-domain Q&A [3]. Other than retrieval-based approaches, the provenance systems and reasoning frameworks have been explored. Self-Consistency (SC) mechanism deals with the robustness of the reasoning chains by *omitting inference paths involving several scenarios and selecting the most aligned outcomes* through reasoning [4]. The authors Dhuliawala et al. presented the Chain-of-Verification (CoVe) method with the goal of extending that research by coming up with a framework that adds an explicit “draft-verify-revise” cycle, as the model checks against itself and the new outputs it generates to assess and fix errors [5]. In addition, these studies combine Buffer of Thoughts and CodePlan [6,7], which restrict the model from extensive generation and query execution until inference reduction takes place. In spite of this progress, this issue remains multifaceted and latent. On the one side, the solely large language model (LLM) data-driven scaling laws can’t ensure factual reliability, mostly because the LLMs often overfit to misleading correlations present in the pretraining corpus. However, this also reduces the speed and performance of the human and machine cooperation. Recently, the current LLM framework, the TruthX [8], introduced the solution of editing machine internal representations, so that they reside in the “truthful space”, which validated that hallucination can be remedied simply by wholesome, straight-view model adjustments rather than through over-complex model topology modification.

This study provides a solid basis for explaining and suppressing the occurrence of hallucinations while focusing on retrieval-based, self-proving, tool usage, and representation editing approaches. However, serious attempts to compare these varied approaches are still limited, and complex interactions between these models both in real life as well as in high stakes areas such as healthcare, finance, or law cap not properly explained yet. Accordingly, we will synthesize the current state of the research, highlight the strong and weak aspects of technical paths, and give an integrated outlook on the developing scenario of hallucination mitigation in large language models.

2. Overview of mainstream technical approaches

2.1. Retrieval-Augmented Generation (RAG)

Starting from the fundamental concept (RAG), information retrieval as well as large language models (LLMs) can be integrated by anchoring the generation procedure in external bases of information. Rather, the model joins the content-drained insert into a knowledge base such as Wikipedia, scientific corpora, or relevant databases prior to generating its answer. Subsequently, the language model runs it in parallel with the query. This mechanism equips in-depth accuracy, interpretability, and control of output.

DOI: 10.54254/2755-2721/2026.TJ29620

The original version of the RAG system employed by Lewis et al. was constructed with two modules – a retriever and a generator [3]. The retriever creates compact, dense vector signatures of the relevant passages and retrieves the top-k relevant excerpts of those passages. The generator, in the most usual case, is a seq-to-seq model that ends up producing the final answer having ingested both the query carried onto it and also the retrieved excerpts. This format opens access to dynamic evidence without retraining or adjustment of parametric knowledge within speculative content.

However, the performance of RAG heavily depends on the retrieval quality and the alignment between retrieved content and model generation. Wrong assumption-making, which causes the generation to proceed on misleading or irrelevant sources, results in the retrieval-caused hallucination. Subsequent works have both been concentrated on this branch and offered enhanced versions of the RAG framework. For instance, in 2021, people observed the development of Fusion-in-Decoder (FiD) by Izacard and Grav, where the integration of multiple retrieved passages is done topically-wise, which leads to an enormous improvement of factual consistency in open-domain QA [9]. Similarly, this retrieval system of two stages, RankRAG, internally incorporates adaptive re-ranking techniques that help to sift low-quality search results upfront just before text generation without losing search recall [10].

The next area that scientists explore is information-based improvements. Such methods apply uniformly on all training data elements and the enhanced training method thus comprises both retrieval and questioning information. Such as the approaches of REALM [11], which is based on the idea of end-to-end retrieval learning during which both the both the retriever and the generator learn in parallel in order to generate more semantically similar information. Additionally, Atlas took this idea one step further to also allow the multi-data set training [12], which ultimately resulted in the techniques to obtain up to date closed-book questions and fashion scientific fact-checking tasks. RARR concentrated on the retrieval-augmented correction [13], which means the framework practically reevaluates its own output through an external evidence in order to reduce its own residual hallucination.

External retrieval-augmented solutions have proven to be indispensable along the way in increasing absolutely the accuracy of practical content such as medical or legal documents. For instance, the information of Singhal et al. points out that RAG combined with PubMed Techniques cuts over 25% of the factual hallucinogenic cases in medical question answering, while AI-Legal systems using mixed real-time technologies mentioned ability to reference facts and precedents to prevent case citation fabrications [14]. The main concept of retrieval-augmented generation is presented in Block Diagram that describes categories for retrieval and generation.

Many issues still remain however. Firstly, systems that include dense retrieval approaches such as DPR and Contriever necessitate large ordering datasets which have to be revised frequently to account for recent materials. Secondly, if the retrieved evidence contains not only vague formulation but also contradicting statements, then there is a chance that RAG produces perceivable but incorrect statements. Moving forward research direction will be focusing on mounting integrated retrieval strategies, estimation of uncertainties, and omitting the mismatches through which the reliability will be improved and the generation of hallucinations will be decreased across multi-step testing environments.

2.2. Reasoning and self-verification

This set of methods addresses hallucination from an angle of model's internal reasoning consistency and develops self-monitoring systems that allow the model to judge its performance on its output and eventually able to refine it. Unlike the retrieval-based methods, which typically use external

DOI: 10.54254/2755-2721/2026.TJ29620

sources of knowledge, self-marking techniques aim to model each inference step of the large language model independently, in order to ensure that answers remain internally consistent and logically correct.

The Wang group is the origin from which Self-Consistency framework based on the Chain-of-Thought paradigm is born and named. Yet, this deduction is not achieved by a unique deterministic rule of reasoning but rather by creating multiple paths of reasoning for the same question using random sampling. Each potential path of inference is realized in the labeled individual lines and the final forecast is completed as the most frequent or the most semantically matching answer is selected. This ensemble-based approach allows removing the individual reasoning errors which are not otherwise identified and having higher performance against low correlations and high discrepancies between different variables under the unstable conditions. Empirical evidence supports Self-Consistency because it minimizes hallucination, fact errors, and inaccuracies in numerical-and-commonsense reasoning.

Averaging out this idea, Dhuliawala et al. later provided Chain-of-Verification (CoVe) framework that has a structured process for verifying answers which would involve exception mechanisms for verification after the fact. In place of treating the human mind as static, the CoVe framework generates various reasoning paths as generated outputs which the model can verify its reasoning claims or problem-solving steps against evidence or its own generated explanations that are truthful or reliable, eliminating imprecise pieces or arguments. The after-generation verification process in the CoVerse model allows exploration of the generate-verify-revise cycle, which reduces false claims, as it represents the true behavior of the self-monitoring. The sampling-based approach contrasts notably to CoVe, as it formalizes this verification process as an ordered chain of sub-questions, allowing one to legislate the relevant corrections with greater clarity and power.

According to the experiments done on the TruthfulQA benchmark, the overall self-consistency framework implementation lowers hallucination frequency in nearly 15% within standard Chain-of-Thought prompting average results. However, retrieval-augmented generation of information is even more beneficial transfer of knowledge is facilitated; this mechanism makes retracement of the evidence easier to do. No matter how high the performance of the model is, the unchanged structural framework of the underlying data model will lead to increased computational expense introduced by multiple tracking passes. This still takes the fact that these efforts are aimed and guided towards developing self-aware, truth-grounded LLM, which are potentially able to contract retrieve information and use them in adaptive reasoning.

2.3. Tool use and planned thinking

Another prominent line of research for mitigating hallucinations conceptualizes large language models (LLMs) not just as passive text generators but instead as programmable reasoning controllers, which can interact with external environments of the reasoning process. In general, these methods may be referred to as dealing with tool use and thinking through planning based on tool use, and their aim for improving the purposeful action of human reasoning is to simplify complex problems by breaking them down into individual executable items and delegating an operation to a specified tool's assistance.

The learning behind this paradigm is based essentially on the premise that hallucinations tend to emerge when internal simulations of activities which exceed the entity's level of knowledge or reasoning, such as calculations, searches, or data retrieval, are attempted. Rather than depending solely on the model's prediction each step of the way, the system enables models to ask external sources for actions that can be checked by others. This includes querying a search API, running

DOI: 10.54254/2755-2721/2026.TJ29620

code, or performing structured data retrieval. The system thus generates explicit grounding and verification on each step, which validate model's decisions and thus control model's generative freedom.

The Buffer of Thoughts (BoT) framework is a at this direction an initial step [7]. Its support helps models saving, recalling, and reviewing intermediate reasoning fragments for turn-by-turn dialogues, meaning it helps as a working memory buffer. This mechanism enables the model to conserve decent reasoning context and hence the outputted incoherence and contradictory statements might be reduced. Experimentally, we have shown that in tasks of complex reasoning and dialog generation that require multiple steps, the consistent internal state of the model leads to a significant decrease in the hallucination rates of the system.

Taking into account the essence of structured reasoning, CodePlan propounds a fusion strategy that changes spoken reasoning to partially executable pseudo-code or Python program [8]. CodePlan essentially facilitates the generation and carrying out of code snippets, giving it the power to determine future scenarios via reliable checks of logical operations, arithmetic computations, and database queries. This strategy reduces not only speculative defeats but also provides excellent fault tracing and correcting options at code base level. In complicated realms like data analysis, financial modeling, and scientific calculations, CodePlan has proved its strength by improving both accuracy and reproduce compared to general strategy like prompts.

The ReAct framework (Reason + Act) is more advanced than the preceding techniques by both including concrete traces of reasoning and tool calls in an interactive process [15]. The model "thinking" method is natural language reasoning, and "act" is programing which is calling outside sources such as search engines or APIs, with each output it is passing into the next reasoning step. This design allows the model to be more efficient in the way it collects missing information, checks its own statements, and dynamically review its own conclusions. For instance, in the cases where the system is faced with the situation of the diagnostic task or legal advice, the system based on ReAct can directly access trusted files, in such way, the hallucination lack of factual mistakes can be enhanced.

On the whole, tool-planning frameworks and the like convert LLMs from static predictors into dynamic solution makers with the ability to constrain their decision-making procedures to verifiable process. Unfortunately, these approaches draw the additional computational and orchestration burden, but they provide a robust and well-established approach to transparency, interpretivity, and factual reliability among large-scale generative systems. Combination of retrival-based groundings and self-verification procedure with the plans of hybrid reasoning provide a hopeful basis for building systems that can precede human-like performance in both measuring and holding responsible.

3. In-depth analysis of the TruthX method

The TruthX framework, a new approach, achieves this goal by means of representation editing, that is, it allows you to alter the way factual information is being accessed and expressed to LLMs during text generation [6]. In contrast to retrieval-based or post-hoc verification methods, TruthX aims to adjust LLMs' internal representation space accordingly. In this respect, a similar core hypothesis lies behind it. Hallucinations do not necessitate that the model lack factual knowledge, but instead that it have an inability to retrieve relevant information correctly with respect to what is already in its latent memory

Congruent with this goal, TruthX re-reconfigures the interplay between semantics and factual precision within the model's embedding space. Instead of utilizing adjusted training data or new

DOI: 10.54254/2755-2721/2026.TJ29620

sources, it introduces the concept of truthfulness direction that is aligned with the concept of the latent vector orientation that corresponds with factual credibility. While constructing this mission, the system first fully understands the directional analogy by the given response pairs including truthful and hallucinating ones, resulting in the internalization of the gradient of factual alignment. Here this objective is achieved by using an auto-encoder that decomposes which one of the hidden representations reflects the semantic versus the truthfulness category leaving the system with the two subspaces instead of having which yeah the models are being educated into either of them.

Upon inference, the model will implement lightweight transformations, thus encouraging the process to generate against the truthfulness vector without interfering with the goal of the response. Unlike with the 13 large-scale language models, our experiment results show that TruthX leads to an approximate 20% factual realism improvement overall while sustaining linguistic smoothness and diversity throughout. This low-cost, model-agnostic, and straightforward throughout approach is a viable architecture to address the scaling problems of enhancing grounding in generative mechanisms. Playing our hand at the graphic level as opposed to using the power of external sources or rule-based validation, this method is the blueprint for the future work on the building of controllable, reliable even large language models.

4. Experimental results and case studies

RAG also shows efficacy in knowledge-centric query performance. Retrieval-augmented generation proves its strongest merits in open-domain and domain-specific question answering, among the cases when successful evidence has a reliable usage. The main enhancements aren't just as a result of the precision of the answers provided but also due to reductions in unsupported claims and smoother auditor tracing (answers bring up passages). Unlike human authors, although tools are finely tuned to your target audience and products, the performance of small-scale is very much related to where you are and what you have available to you. A knowledge base may have little to date material or even the top-k might be of very random quality; through the models' scope, now they are able to imagine convoluted claims and intertwine irrelevance into the story with a confident voice. For the additional rule of thumb, in which when a re-ranking phase reinforcement is given that is a condition based on the evidence (for example, the forbidden of k tokens that do not rely on the retrieved document), such strengthening of faithfulness is visible. However, the hidden cost of the conservative fallback policy is that there is the risk of low recall.

The aim is to create the reasoning process essentially more reliable and the self-verification approach capable of accurately detecting the inconsistency. Clear occurrences of this are in multi-step commonsense and multi-hop question-answering problems, where the speculations and contradictions are reduced by self-verification pipelines. The Self-Consistency strategy will filter out offshoots stemming from different momentum by aggregating reasoning paths, smoothing interspersed deviations. Verification-of-Chain breaks one string of evolving prospects into two or into verifiable claims, and cough it up by prompting the model to verify it before giving a final revision. Usually, in the case-based settings, such pipelines significantly lower hallucination rates in comparison with single-pass reasoning models with different recall and compute speed due to multiple passes. The better performance in the calibration is another effect we find: models are more self-confident in amending or voicing after the verification stage.

In the complex procedures delivered by the tool usage and plan-based methods, the fundamental goal is to prepare the output for machine execution. For example, arithmetic, data retrieval, or when programmatic transformations are involved, tools (like search engines, calculators, and their executions) complimentarily augment planning-based methods that offer the best standards. Through

DOI: 10.54254/2755-2721/2026.TJ29620

converting a natural language plan into an executable set of objectives, the detection of fatal bugs becomes possible, and the failure mode sees clear signals (for instance, programming errors instead of fictitious quantities). This approach is sensitive to orchestration quality (scripted-procedural calls, error traversal) but given the right signature, it will suppress fabrications and lead to accurate metrics and computations.

TruthX will correct semantic with its method. Representation editing complements representation and strengthens the reasoning by creating the truthfulness vector even though the arguments are slightly vague. In ablations where retrieval returns bordering-referenced passages, including the vector of TruthX truthfulness, the model tends to abstain from deeper commitment to what is not yet fact, instead leading to tighter language constructions or neutral phrases. Such method is less intrusive during the process of inference and well composable with RAG (pre-generation grounding) and CoVe (post-generation checking).

Domain case studies:

Medicine: Reinforcing PubMed and clinical guideline retrieval together with knowledge-centric RAG enhances the specificity of diagnostic guidance alongside the reduction in diagnostic recommendation not based on evidence. Model misbehavior is usually seen in outdated guides such as more recent dosing or drug synonyms. Existence of CoVe layer in this pipeline pushes the model to crosscheck the date of the guideline; the inclusion of a TruthX layer will discourage the display of certain when the evidence is conflicting.

Law: Hybrid RAG that relies on statute and precedent indices drastically decrease the instances of fabricated citations in judicial case law and statute searching. A verification passes in which the model is instructed to cite clause numbers as well as check the locations of their references helps prevent the cross-jurisdictional flow with these kinds of references. Systems that are orally positioned absent this fit and/or the fit is put to the second plate when only ancillary material is gathered.

Finance: For ratio analysis and KPI extraction, a multi-tasking module is being used which generates executable code to interpret unstructured filling and compute metrics that are suitable to structural filing versus free-text arithmetic functions more commonly practiced. Compared to narrative parts, system brings the gap to attention when the computed values are in disagreement by hyperlinking the calculated values with statement sections as a substitute to average them in prose – i.e., accentuating the difference between possible hallucination views conceptually and in practice.

5. Challenges and future outlook

Hallucination eradication revises offer appreciable outcomes; however, the issue continues serving as a two-headed hydra being persistent and complex. Evaluation measures confined by the narrow benchmarks are still the common practices of today, which are incapable of reflecting the changes, complexities, and variety of the real-world information in the remain. Consequently, such tests emphasize starters, dull the duration ability of realistic knowledge, and reveal models' adjustment exhaustion to novel information. On the basic matters, don't simply raise parameters scales if knowing already that some of them are less credible. The bigger models could be the cause of the very fact that they have such creativity in generating vague or unreliable statements, which could be the reason as for the trade-off between the confidence of generative and the accuracy of facts.

Not only does hallucination involve technical aspects but it raises ethical and social concerns too. Possible mistrust in automated systems by the public is one of the effects of the propagation of fake news by language models across public discourse. The impact of wrong information in the working

DOI: 10.54254/2755-2721/2026.TJ29620

environment can be enormous. For instance, legal and fiscal consequences may be the implications of inaccurate outputs. This indicates the need for establishing proper transparency frameworks.

Future research shall therefore turn to involving all mechanisms in action while initially making a trial of combining different methods instead of targeting single-method strategies. Here, we propose hybrid pipelines that are enhanced alongside retrieval grounding, representation editing, and self-verification – culminating a framework called ReAct with TruthX and CoVe – which provides different balance of accuracy and efficiency. In addition to the above mixed systems, the creation of dynamic, officially used certifying system permits models to be tested against a shifting scale of factual knowledge will add value to the whole process. By using light-touch techniques that change model run time behavior rather than fully retrain it, and also encourage people to join the process of coding, releasing information about the model, they will be controllable, transparent, and foster social trust. As clarify, the big deal about large language models is a combination of algorithmic innovation and ethical supervision of deployments.

6. Conclusion

This study considered all the current mainstream methods used for hallucination control of large language models, including: retrieval-augmented generation (RAG), reasoning-centred self-verification, tool- and planning-based pipelines, and representation editing presented by using TruthX. Even within the most diverse frameworks, the outcome is ultimately the following one: hallucination is not a one-method-one-solution problem, that is, there are several methods of treatment. Nevertheless, this method of narration by the use of the RAG not only leads to more knowledge-centric responses when the surface is very recent and covers many issues but also the model neutralizes jumps of reasoning and belief in ungrounded claims by different means like self-verification which makes reasoning that is grounded in facts much stronger. Each family puts forward their strengths – good coverage, consistency, performance, or semantic calibration – and their efficiency and effectiveness are supplementary rather than mutually exclusive.

Hence, a set of design principles is gradually suggested that can enable practitioners to develop powerful claims efficiently on the account of the above findings. The generation loop should be started with evidence, progressing applies the condition of decoding on retrieved spans and keeping forex trading to abstain whenever the help is inadequate. Loop back after result generation should be break answer into checkable claims and enabling anyhow revision rather than writing the answer again. In the event its entailed real one mode of computation or structured queries systems should be coded to script or invoke tool calls so that truth is realized in computing processes by assigning truth a lower level of styling. The representation of sentiments with model-agnostic representations and affordable intervention sampling will lead to truthful phrasing with slightly more cost implications. However, these strategies must also be sanctioned with provable-at-side Incidents/ for example, verbal withdraw Policy and Visible Providentiality of Expert witness due to the path taken to find out trustworthy quality or messages it could not access.

So, behind what is this-connected ecosystem level, stagnant, single-lingual metrics are not enough data quality judgment. Multilayer, turbulent, and geographically-aligned test beds are essential for assessing currency, predictability, and adaptability to distribution shift. Both governance and deployment protocols are of equal importance to the algorithms, hence, formulation of biasable corpuses for retrievals, tracking models for changes and regular revision, and of human-in-the-loop evaluation for critical decision-making processes. In this regard, the recently promising path of neural networking in architecture, where the internal process of reasons forsimply incorporating, representation editing, the external retrieval ground, and signal from the other

DOI: 10.54254/2755-2721/2026.TJ29620

important modalities is highly valuable. In an actual sense, construct data models that are pretty much verifiable, traceable, and actionable not only represents a performance target but at this deep level is the fundamental principle for trustworthy AI.

References

[1] Koehn, P., and Knowles, R. (2017). Six Challenges for Neural Machine Translation. In Proceedings of the NMT Workshop.
[2] Lin, S., Hilton, J., Evans, O. (2021). TruthfulQA: Measuring How Models Mimic Human Falsehoods. ACL.
[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS.
[4] Wang, X., Wei, J., Schuurmans, D., Le, Q., et al. (2023). Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. ICLR.
[5] Dhuliawala, S., Alayrac, J.-B., et al. (2024). Chain-of-Verification Reduces Hallucination in Large Language Models. Findings of ACL.
[6] Yang, L., Yu, Z. C., Zhang, T. J. et al. (2024). Buffer of Thoughts: Thought-Augmented Reasoning with Large Language Models. NeurIPS.
[7] Wen, J. X. et al. (2025). CodePlan: Unlocking Reasoning Potential in Large Language Models by Scaling Code-form Planning. ICLR.
[8] Zhang, S. L., Yu, T., Feng, Y. (2024). TruthX: Alleviating Hallucinations by Editing Large Language Models in Truthful Space. ACL.
[9] Izacard, G., and Grave, E. (2021). Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering (FiD). ICLR.
[10] Glass, M., Shen, S., et al. (2022). RankRAG: Improved Retrieval-Augmented Generation with Re-ranking Mechanisms. NeurIPS.
[11] Guu, K., Lee, K., Tung, Z., et al. (2020). REALM: Retrieval-Augmented Language Model Pre-Training. ACL.
[12] Izacard, G., et al. (2022). Atlas: Few-shot Learning with Retrieval-Augmented Language Models. arXiv preprint.
[13] Gao, L., et al. (2023). RARR: Retrieval-Augmented Refinement for Reducing Hallucination in Large Language Models. ACL.
[14] Singhal, K., Tu, T., et al. (2023). Large Language Models Encode Clinical Knowledge. Nature.
[15] Yao, S., Zhao, J., et al. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. ICLR.