Toward Reliable Scientific Hypothesis Generation: Evaluating Truthfulness and Hallucination in Large Language Models

Guangzhi Xiong[1] , **Eric Xie**[1] , **Corey Williams**[1] , **Myles Kim**[1] , **Amir Hassan Shariatmadari**[1] , **Sikun Guo**[1] , **Stefan Bekiranov**[1] and **Aidong Zhang**[1]
[1]University of Virginia
{hhu4zu, jrg4wx, cmw6pa, mbt8hz, ahs5ce, qkm6sq, sb3de, aidong}@virginia.edu

Abstract

Large language models (LLMs) have shown significant potential in scientific disciplines such as biomedicine, particularly in hypothesis generation, where they can analyze vast literature, identify patterns, and suggest research directions. However, a key challenge lies in evaluating the truthfulness of generated hypotheses, as verifying their accuracy often requires substantial time and resources. Additionally, the hallucination problem in LLMs can lead to the generation of hypotheses that appear plausible but are ultimately incorrect, undermining their reliability. To facilitate the systematic study of these challenges, we introduce **TruthHypo**, a benchmark for assessing the capabilities of LLMs in generating truthful scientific hypotheses, and **KnowHD**, a knowledge-based hallucination detector to evaluate how well hypotheses are grounded in existing knowledge. Our results show that LLMs struggle to generate truthful hypotheses. By analyzing hallucinations in reasoning steps, we demonstrate that the groundedness scores provided by KnowHD serve as an effective metric for filtering truthful hypotheses from the diverse outputs of LLMs. Human evaluations further validate the utility of KnowHD in identifying truthful hypotheses and accelerating scientific discovery. Our data and source code are available at https://github.com/Teddy-XiongGZ/TruthHypo.

1 Introduction

Large language models (LLMs) have transformed the landscape of artificial intelligence, demonstrating remarkable capabilities across diverse applications, from natural language understanding to creative content generation [Karanikolas *et al.*, 2023; Franceschelli and Musolesi, 2024; Raiaan *et al.*, 2024]. These models, trained on extensive corpora of text, demonstrate an ability to analyze, summarize, and generate human-like text, enabling advancements across diverse domains. Recently, there has been a growing interest in leveraging LLMs for scientific discovery [Zhong *et al.*, 2023; Yang *et al.*, 2023; Kumar *et al.*, 2023; Liu *et al.*, 2024; Baek *et al.*, 2024; Si *et al.*, 2024]. Their capacity to process and synthesize vast amounts of scientific literature positions them as valuable tools in aiding researchers, particularly for tasks such as literature reviews, summarization, and even generating new hypotheses [Qi *et al.*, 2023; Zhou *et al.*, 2024; M. Bran *et al.*, 2024; Wright *et al.*, 2022; Zeng *et al.*, 2023; D'Arcy *et al.*, 2024; Ifargan *et al.*, 2025; Yang *et al.*, 2025].

One particularly promising application of LLMs is their use in scientific hypothesis generation, where they can assist in identifying promising research directions. By analyzing extensive scientific literature, LLMs can uncover gaps in existing knowledge and propose novel hypotheses that may not be immediately apparent to human researchers. For instance, LLMs have been successfully applied to propose novel drug combinations for breast cancer treatment, some of which were later validated in laboratory experiments, showcasing their potential to accelerate biomedical discoveries [Abdel-Rehim *et al.*, 2024].

Despite these advancements, there are substantial challenges that limit the practical utility of LLMs in scientific hypothesis generation. A critical concern is the inability to evaluate the truthfulness of generated hypotheses. While LLMs can generate hypotheses that seem plausible, it remains uncertain whether these hypotheses are valid and grounded in existing knowledge or merely hallucinated and scientifically invalid. This issue is further exacerbated by the well-documented "hallucination" problem, where LLMs confidently produce information that is factually inaccurate or unsupported, posing challenges to their reliability in scientific contexts. While current research has largely focused on improving the novelty and diversity of LLM-generated hypotheses, their truthfulness and grounding in established knowledge remain underexplored [Baek *et al.*, 2024; Hu *et al.*, 2024; Si *et al.*, 2024].

To address these challenges, we introduce TruthHypo[1], a comprehensive benchmark for evaluating the ability of LLMs to generate truthful scientific hypotheses, and KnowHD, a knowledge-based hallucination detection framework designed to assess the groundedness of these hypotheses. TruthHypo, built on a biomedical knowledge graph along with a

[1]An extended version of this paper with Appendix is available at https://arxiv.org/abs/2505.14599.

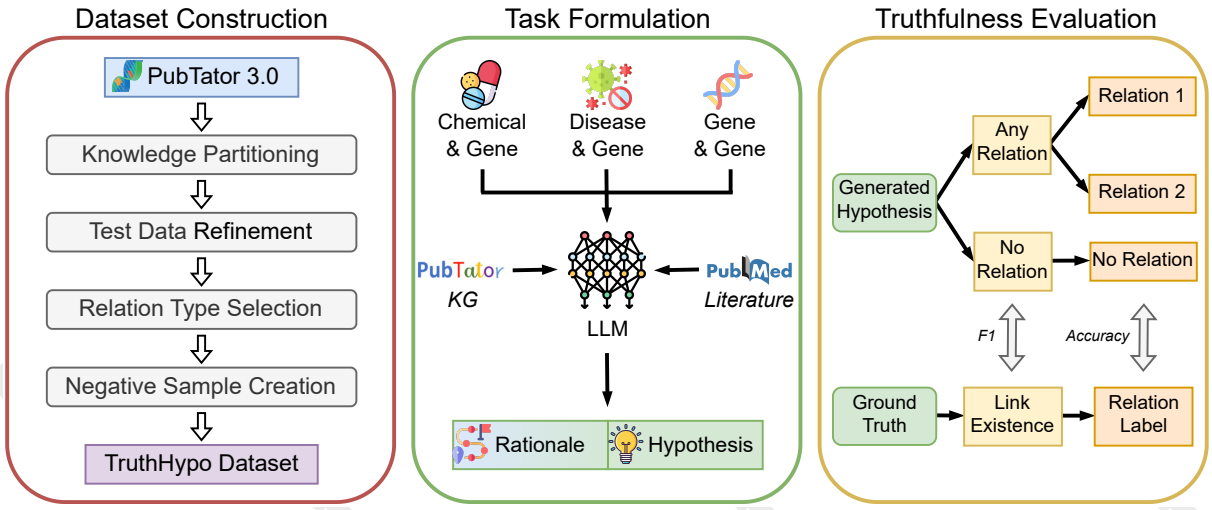

Figure 1: Overview of the TruthHypo benchmark, including dataset construction, task formulation, and truthfulness evaluation.

domain-specific corpus, provides a controlled environment to evaluate how well LLM-generated hypotheses align with established scientific knowledge. KnowHD focuses on analyzing the reasoning processes of LLMs to identify hypotheses that are likely hallucinated or untruthful. Our findings reveal that LLMs face significant challenges in generating truthful hypotheses. By analyzing hallucinations in the reasoning processes behind generated hypotheses, we demonstrate that groundedness scores from KnowHD serve as an effective signal for identifying truthful hypotheses from the diverse outputs of LLMs. Human evaluations on open-ended hypothesis generation tasks further confirm the utility of KnowHD in identifying scientifically valid hypotheses.

Our main contributions are summarized as follows:

- We introduce TruthHypo, a comprehensive benchmark designed to evaluate the ability of LLMs to generate truthful scientific hypotheses.
- We propose KnowHD, a knowledge-based hallucination detection framework that assesses the groundedness of LLM-generated hypotheses and identifies hallucinated claims by analyzing the rationale behind the hypothesis generation.
- We provide an extensive analysis of existing LLMs on TruthHypo, highlighting their limitations and challenges in generating truthful hypotheses.
- Our evaluation further reveals the connection between hallucination and truthfulness of generated hypotheses, showing the effectiveness of using KnowHD to select truthful and grounded hypotheses.

2 Truthful Hypothesis Generation Benchmark

To systematically evaluate the ability of large language models (LLMs) to generate truthful scientific hypotheses, we introduce TruthHypo, a benchmark tailored for biomedical hypothesis generation. TruthHypo is designed to simulate real-world conditions by employing rigorous dataset construction, task formulation, and truthfulness evaluation metrics. An overview of the dataset construction, task formulation, and evaluation framework is depicted in Figure 1.

2.1 Dataset Construction

The dataset for TruthHypo is derived from PubTator 3.0 [Wei *et al.*, 2024], a comprehensive biomedical knowledge graph that includes annotated relations (also called edges) extracted from scientific articles. To simulate the temporal progression of scientific discovery, we partitioned the graph into "seen" and "unseen" subsets based on the publication years of the corresponding articles. Relations in the "seen" subset were extracted from papers published before 2023, identified by PMIDs ≤ 36600000[2]. The "unseen" subset, designed to represent new discoveries, comprises relations extracted from papers published after 2024, identified by PMIDs ≥ 38200000.

To ensure no overlap between the two subsets, we removed the edges in the unseen subset that shared head and tail entities with those in the seen subset. In addition, to maintain quality and validity, only relations discovered by multiple articles in the test data were retained. This filtering process guarantees that the unseen subset exclusively contains knowledge unavailable before 2024, simulating the conditions of future scientific research.

In building the dataset, we focused on three key relation types: "Chemical & Gene", "Disease & Gene", and "Gene & Gene". These relation types were chosen for their complementary nature, detailed annotations, and potential for objective evaluation. To construct comprehensive classification tasks for evaluating different LLMs, we augment the dataset with negative test cases to assess whether LLMs tend to make false-positive predictions on entity pairs that lack a direct relationship in the existing knowledge base. The number of

[2]PMID is the unique identifier of the paper where the edge was extracted.

negative samples (labeled as "no_relation") for each relation type is controlled to align with the average number of instances across other labels of the same relation type. The final dataset has 1209 instances for the "Chemical & Gene" task, 268 instances for the "Disease & Gene" task, and 547 instances for the "Gene & Gene" task. A summary of the dataset statistics is presented in Table 1.

Task	Label	# Instance
Chemical & Gene	positive_correlate	328
	negative_correlate	478
	no_relation	403
Disease & Gene	stimulate	104
	inhibit	75
	no_relation	89
Gene & Gene	positive_correlate	247
	negative_correlate	118
	no_relation	182

Table 1: Statistics of various tasks in the TruthHypo benchmark.

2.2 Task Formulation

The TruthHypo benchmark includes three tasks, corresponding to the selected relation types: "Chemical & Gene", "Disease & Gene", and "Gene & Gene". For each task, the input is a hypothesis generation query with two entities, and the LLM is required to hypothesize the potential relationship between them based on available knowledge and reasoning.

To comprehensively assess LLM performance, we evaluate their ability to generate hypotheses under various knowledge augmentation settings. In the first setting, LLMs rely solely on their parametric knowledge – information encoded in their parameters during pretraining on large corpora. This evaluates the model's intrinsic understanding and reasoning capabilities.

To enhance hypothesis generation, we introduce a second setting in which LLMs are augmented with structured knowledge from the "seen" knowledge graph. In this approach, key entities from the input are mapped to nodes in the graph, and multi-hop link chains connecting these nodes are explored. These chains, representing relevant relationships, are transformed into textual descriptions and provided as context for the model to use during hypothesis generation.

Another setting leverages information from biomedical literature using a retrieval-augmented generation (RAG) pipeline. Relevant documents are retrieved from the PubMed corpus[3] using BM25 [Robertson *et al.*, 2009]. To maintain consistency with the knowledge graph's temporal split, only articles with PMIDs ≤ 36600000 are included in the retrieval. This simulates the process of generating hypotheses based on literature available at a given point in time.

Finally, we consider a combined setting, where both structured knowledge from the graph and unstructured information from retrieved literature are used to support hypothesis generation. This comprehensive approach provides a more holistic context, enabling the model to reason across both sources. The LLM prompt templates we used to combine the external information with the original user instructions can be found in the Appendix.

[3]https://pubmed.ncbi.nlm.nih.gov/

2.3 Evaluation Metrics

To evaluate the quality of generated scientific hypotheses, we employ a set of complementary metrics tailored to different aspects of hypothesis generation. These metrics assess the performance of LLMs in identifying valid connections between entities (link-level evaluation) and predicting specific relations (relation-level evaluation).

For link-level evaluation, we focus on precision, recall, and F1 score. Precision measures the proportion of correctly identified connections among all hypothesized connections, emphasizing the reduction of false positives. Recall evaluates the model's ability to comprehensively identify all valid connections, capturing its sensitivity to true positives. The F1 score, as the harmonic mean of precision and recall, provides a balanced measure of performance, combining both the accuracy of predictions and the coverage of valid connections. These link-level metrics are critical for assessing the LLM's ability to hypothesize plausible relationships between entities, regardless of the specific relation type.

For relation-level evaluation, we employ accuracy to measure how often the generated hypotheses match the correct relation labels in the ground truth. Accuracy captures the overall correctness of hypotheses by considering both the existence of a connection and the predicted relation type. While precision, recall, and F1 focus on identifying potential connections, accuracy provides a finer-grained assessment of the model's capability to generate accurate relation labels.

By combining link-level and relation-level evaluations, the TruthHypo benchmark comprehensively measures the truthfulness of LLM-generated hypotheses, assessing the ability of LLMs to produce scientifically valid outputs.

3 Knowledge-based Hallucination Detection

As discussed earlier, a critical concern regarding the truthfulness of LLM-generated hypotheses is the occurrence of hallucinations, where models generate plausible-sounding but unsupported claims. To address this, we introduce KnowHD, a knowledge-based hallucination detection framework that evaluates the groundedness of LLM-generated hypotheses by analyzing the rationale behind their generation. KnowHD operates using scientific literature, knowledge graphs, or a combination of both as the knowledge base. An overview of the framework is presented in Figure 2.

To evaluate groundedness, each hypothesis and its reasoning chain are first decomposed into a set of atomic claims. This step is critical because hypotheses often consist of compound reasoning steps, some of which may be supported by existing knowledge while others may not. Parsing these into atomic claims allows a more granular evaluation of groundedness and isolates unsupported components. This step is implemented by prompting LLMs with the template shown in the Appendix.

When using scientific literature as the knowledge base, relevant documents for each atomic claim are retrieved from

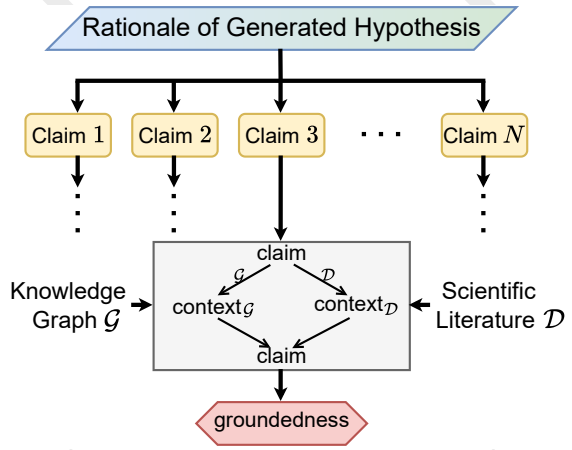

Figure 2: Overview of the KnowHD hallucination detection framework. Hypotheses are parsed into atomic claims, which are then evaluated for groundedness using a knowledge graph, scientific literature or both as knowledge sources.

the PubMed corpus, limited to articles published before 2023 (PMID $\leq$ 36600000). BM25 is employed to rank documents based on their relevance to the claim. To ensure computational efficiency and focus on the most relevant information, only the top-k documents are retained. The context retrieved from the literature corpus $\mathcal{D}$ for a claim p is defined as:

$$\text{context}_\mathcal{D}(p) = \{d_1, d_2, \cdots, d_k | \\ d_i \in \mathcal{D}, \text{BM25}(p; d_i) \geq \tau, \text{rank}(d_i) \leq k\}, \quad (1)$$

where d_i represents a document in the corpus, $\text{BM25}(p; d_i)$ is the relevance score assigned to the document for the claim p. τ is a threshold ensuring relevance, and $\text{rank}(d_i)$ denotes the rank of d_i in the BM25-retrieved list.

When using a knowledge graph $\mathcal{G}$ as the knowledge base, the context for a claim is derived from the graph structure. For a claim p, relevant knowledge is extracted as:

$$\text{context}_\mathcal{G}(p) = \{(e_h, r, e_t) \in \mathcal{G} \,|\, \{e_h, e_t\} \subseteq \mathcal{V}(p)\}, \quad (2)$$

where (e_h, r, e_t) represents an edge in the knowledge graph with head entity e_h, tail entity e_t, and relation r. The set $\mathcal{V}(p)$ contains all entities mentioned in the claim p.

The groundedness of a claim is determined based on whether the given context information ($\text{context}_\mathcal{D}$, $\text{context}_\mathcal{G}$, or $\text{context}_\mathcal{D} \cup \text{context}_\mathcal{G}$) can fully support the claim, which is implemented by prompting LLMs to provide a judgment using the template in the Appendix. If the concatenated context collectively entails the claim, it is considered grounded. The overall groundedness of a hypothesis h is computed as:

$$\text{groudedness}(h) = \frac{1}{|\mathcal{C}(h)|} \sum_{p \in \mathcal{C}(h)} \mathbb{1}[\text{context}(p) \models p], \quad (3)$$

where $\mathcal{C}(h)$ represents the set of atomic claims for hypothesis h, and $\mathbb{1}[x \models y]$ returns 1 if x entails y and 0 otherwise. The $\text{context}(p)$ can be $\text{context}_\mathcal{D}(p)$, $\text{context}_\mathcal{G}(p)$, or $\text{context}_\mathcal{D}(p) \cup \text{context}_\mathcal{G}(p)$.

By offering both literature-based and graph-based contexts, KnowHD provides a robust framework for hallucination detection, offering flexibility to adapt to the available knowledge sources. This systematic evaluation of atomic claims enables a detailed assessment of the groundedness of hypotheses, identifying unsupported components and improving the reliability of LLM-generated outputs.

4 Benchmark Analysis on TruthHypo

4.1 Experiment Settings

To assess the ability of existing LLMs to generate truthful scientific hypotheses, we selected a diverse range of models varying in type and size. The Llama-3 family [Dubey *et al.*, 2024] represents open-source LLMs, while the GPT-4 family [Achiam *et al.*, 2023] exemplifies proprietary models. From each family, we evaluated two LLMs of different sizes (Llama-3.1-8B & Llama-3.1-70B, GPT-4o-mini & GPT-4o) to investigate size-related differences in performance. All LLMs were trained on the knowledge available before 2024, preventing recall of the exact knowledge for hypothesis generation. More implementation details are in the Appendix.

The TruthHypo benchmark evaluates LLMs across four distinct settings: (1) parametric knowledge only, (2) parametric knowledge with knowledge graphs (KG), (3) parametric knowledge with literature (Lit.), and (4) parametric knowledge with both KG and literature. These settings allow us to explore the impact of external knowledge sources on hypothesis generation. The F1 and accuracy scores of different models are reported in this section. More detailed results on the precision and recall can be found in the Appendix.

4.2 Comparison of LLMs in Truthful Hypothesis Generation

Table 2 presents the evaluation results for different LLMs and knowledge settings on TruthHypo. Across all tasks, the results indicate that most LLMs struggle to generate truthful scientific hypotheses, with only GPT-4o achieving mean accuracies exceeding 60%. Additionally, we can observe that link-level F1 scores are higher than relation-level accuracy scores, which indicates that LLMs can identify potential connections between entities but often fail to accurately predict the specific relationships.

For models from the same family with different sizes, larger LLMs tend to generate scientific hypotheses more likely to be truthful. This can be attributed to two main factors. First, larger LLMs generally perform better because they can store and leverage more knowledge in their parameters, as shown by the results of parametric knowledge-only setting. Second, LLMs of different sizes have diverse capabilities to process external knowledge for hypothesis generation. For example, GPT-4o-mini shows a modest 1.14% accuracy improvement when augmented with KG and literature, whereas GPT-4o achieves a more substantial 5.14% increase under the same conditions. This suggests that larger LLMs can better utilize additional context to reason about truthful scientific hypotheses. Similar trends are observed when comparing Llama-3.1-8B and Llama-3.1-70B. Interestingly, smaller models, such as Llama-3.1-8B, sometimes experience decreased performance when information from KG and literature is introduced. This degradation may stem from chal-

Knowledge	LLM	Chemical & Gene		Disease & Gene		Gene & Gene		Average	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc
Parametric [Wei *et al.*, 2022]	Llama-3.1-8B	80.16	42.43	79.37	41.04	79.19	46.07	66.90	43.23
	Llama-3.1-70B	81.36	52.44	83.29	54.48	76.66	49.91	71.54	52.03
	GPT-4o-mini	83.31	61.29	81.84	59.33	79.32	53.02	75.49	58.79
	GPT-4o	80.74	66.17	75.38	54.85	71.56	55.58	73.17	61.81
Parametric + KG [Baek *et al.*, 2024]	Llama-3.1-8B	81.37	40.61	79.59	48.13	79.61	48.45	70.65	43.73
	Llama-3.1-70B	87.85	62.86	67.62	52.24	78.29	58.14	79.10	60.18
	GPT-4o-mini	86.42	57.65	74.17	55.60	81.65	62.34	79.40	58.65
	GPT-4o	88.66	63.85	79.50	56.72	82.73	61.06	81.62	62.15
Parametric + Lit. [Lewis *et al.*, 2020]	Llama-3.1-8B	80.78	46.07	80.46	43.28	79.91	42.60	68.58	44.76
	Llama-3.1-70B	82.56	56.74	84.16	52.99	79.18	51.55	73.37	54.84
	GPT-4o-mini	85.28	59.80	**85.71**	53.73	81.50	51.19	77.08	56.67
	GPT-4o	79.52	65.92	75.84	55.97	64.69	51.92	71.84	60.82
Parametric + KG + Literature	Llama-3.1-8B	75.98	36.48	77.58	41.42	79.19	45.70	65.37	39.62
	Llama-3.1-70B	84.80	59.31	77.64	56.34	81.24	55.76	77.37	57.95
	GPT-4o-mini	88.34	60.96	84.47	58.21	84.17	58.50	81.42	59.93
	GPT-4o	**89.71**	**69.31**	82.86	**62.31**	**85.91**	**63.99**	**83.55**	**66.95**

Table 2: Performance comparison of different LLMs on the TruthHypo benchmark across various knowledge settings. The metrics reported are link-level F1 and relation-level accuracy (Acc) for each task (Chemical & Gene, Disease & Gene, Gene & Gene), as well as their averages. "Param." denotes parametric knowledge, while "KG" and "Lit." refer to knowledge graphs and literature, respectively. All scores are percentages (%).

lenges in effectively integrating internal and external information, which can disrupt the model's reasoning processes.

Performance differences are also observed across the three relation types: "Chemical & Gene", "Disease & Gene" and "Gene & Gene". Notably, all larger models, including GPT-4o, GPT-4o-mini, and Llama-3.1-70B, tend to perform better on "Chemical & Gene" tasks than on the other two types. This trend suggests that the "Chemical & Gene" task may be more aligned with the pre-trained knowledge or reasoning capabilities of these models. In contrast, the smaller Llama-3.1-8B shows a more inconsistent pattern, with performance varying across tasks and settings, likely reflecting its more limited parametric capacity and reasoning abilities. These variations in performance across relation types may be attributed to differences in training data distributions or the complexity of the relation types themselves. The relatively stronger performance on the "Chemical & Gene" task highlights potential domain-specific biases or strengths in the LLMs, offering insights into their suitability for targeted applications in real-world scientific discovery.

4.3 Hallucination Detection on LLM-generated Hypotheses

To assess the groundedness of the generated hypotheses, we evaluated their rationales using KnowHD under various knowledge settings. KnowHD measures how well a hypothesis is supported by structured knowledge (KG), unstructured knowledge (literature), or both combined. The groundedness evaluation results for hypotheses generated by GPT-4o-mini are presented in Table 3.

The results demonstrate distinct contributions of KG and literature to grounding hypotheses. For example, KnowHD with the literature as the support knowledge base can verify 76.30% claims in the rationales of literature-augmented 'Chemical & Gene" hypotheses. However, the hallucination detector can hardly verify the rationale generated based on adding KG information to parametric knowledge with only 51.08% of the claims being grounded. Combining KG and literature yields the highest groundedness scores, effectively leveraging the complementary strengths of both sources to identify grounded claims and detect hallucinations.

Task	Knowledge	KnowHD		
		KG	Lit.	KG + Lit.
Chemical & Gene	Parametric	44.77	67.34	74.49
	+ KG	49.93	51.08	73.03
	+ Lit.	47.19	76.30	83.20
	+ KG + Lit.	50.57	65.25	78.90
Disease & Gene	Parametric	45.44	71.56	78.91
	+ KG	57.07	60.70	79.81
	+ Lit.	49.34	78.65	85.32
	+ KG + Lit.	51.11	75.26	86.68
Gene & Gene	Parametric	42.94	67.81	76.16
	+ KG	58.07	56.41	79.64
	+ Lit.	44.49	76.43	84.48
	+ KG + Lit.	54.03	67.96	82.87

Table 3: KnownHD (KG, Lit., and KG + Lit.) groudedness scores of hypotheses generated by GPT-4o-mini under different knowledge settings. All scores are percentages (%).

To further explore the relationship between hallucination and truthfulness, Figure 3 examines mean accuracy as a function of groundedness scores. Hypotheses were grouped based on their groundedness scores, and the average accuracy for each group was calculated. The figure reveals a positive cor-

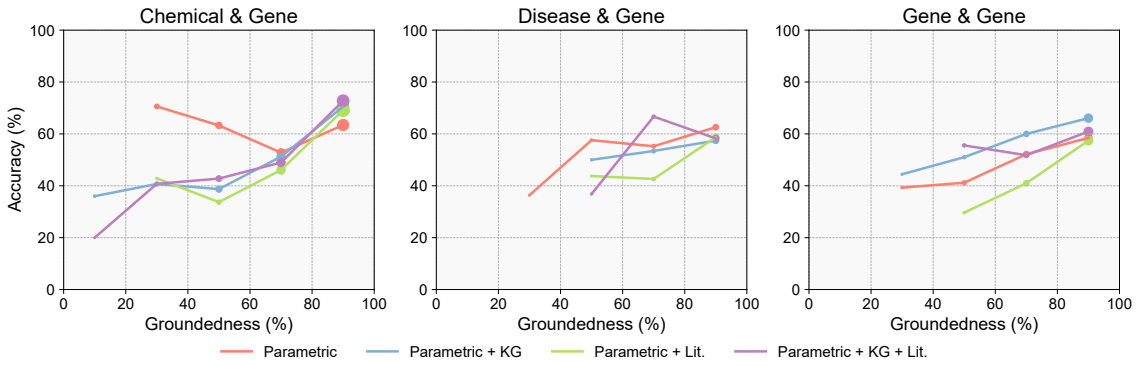

Figure 3: Mean accuracy corresponding to different levels of groundedness. Hypotheses are grouped based on their groundedness scores provided by KnowHD (KG + Literature). Only groups with no less than 10 hypotheses are shown in the plots. The dot size reflects the number of samples in each level of groundedness.

relation between groundedness scores and hypothesis truthfulness. As groundedness scores increase, the likelihood of the hypothesis being truthful also increases. For example, GPT-4o-mini achieves a mean accuracy of 60.96% on "Chemical & Gene" tasks under the combined KG + Literature setting, but this rises to 72.77% for hypotheses with groundedness scores above 80%. These findings underscore the potential of KnowHD to identify hypotheses with a higher probability of being truthful, particularly in contexts enriched with external knowledge.

4.4 Improving Generation of Truthful Hypotheses with KnowHD

To validate the utility of KnowHD on enhancing hypothesis generation, we prompted LLMs to generate five candidate hypotheses for each input and selected the one with the highest groundedness score as the final output. This approach was compared to two baselines: the greedy search method, where the hypothesis is generated using greedy next-token selection by the LLM, and the self-consistency method [Wang *et al.*, 2022], which selects hypotheses based on majority voting across multiple predictions.

As shown in Figure 4, groundedness-based hypothesis selection generally outperforms both the greedy search and majority-voting methods across most knowledge settings. In the parametric knowledge-only setting, the majority-voting method achieves slightly higher accuracy (61.86%) compared to groundedness-based selection (59.83%). However, as external knowledge is introduced, groundedness-based selection demonstrates consistent improvements over both baselines. For example, in the combined parametric + KG + Literature setting, GPT-4o-mini achieves an average accuracy of 63.44% when groundedness-based selection is used, approaching the performance of the larger GPT-4o model.

These results highlight the effectiveness of groundedness scores in scenarios where external knowledge is incorporated, as they help identify hypotheses that are more likely to be truthful. By detecting hallucinations in reasoning steps and focusing on grounded hypotheses, KnowHD provides a robust mechanism for enhancing the reliability and truthfulness of LLM-generated scientific hypotheses.

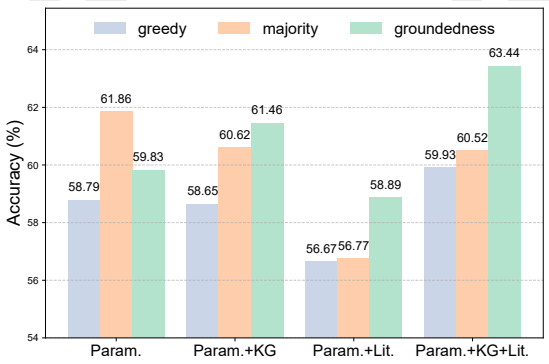

Figure 4: Accuracy improvements of GPT-4o-mini using KnowHD (KG + Lit.) groundedness scores for hypothesis selection. "Param.", "KG" and "Lit." denote parametric knowledge, knowledge graphs, and literature, respectively.

5 Human Study on Open-ended Tasks

To further assess the generalizability of KnowHD's effectiveness in selecting truthful hypotheses, we conducted experiments on open-ended hypothesis generation tasks. These tasks were designed to evaluate whether KnowHD could reliably identify hypotheses with a higher likelihood of truthfulness across broader and less structured generation scenarios. For this analysis, we utilized the publicly available hypothesis generation dataset introduced by [Qi *et al.*, 2024], which involves generating free-form hypotheses based on given background information. We selected GPT-4o-mini as the tested LLM and enhanced its hypothesis generation process by incorporating external knowledge from scientific literature and knowledge graphs (KG). The model was prompted to generate five distinct scientific hypotheses for each input. These hypotheses were then evaluated by KnowHD, which assessed their groundedness based on their alignment with both structured (KG) and unstructured (literature) knowledge sources.

To analyze the relationship between groundedness scores and hypothesis truthfulness, we filtered generated hypotheses to create pairs with contrasting groundedness levels. For each input, we identified one hypothesis with the highest ground-

Group	Groudedness	GPT-4o	Human
highly-grounded	83.94	61.11	59.26
lowly-grounded	40.61	38.89	40.74
p-value	7.84×10^{-11}	1.05×10^{-2}	2.71×10^{-2}

Table 4: Results of analysis on open-ended hypothesis generation tasks. "GPT" and "Human" denote the selection ratios by GPT-4o and human experts, respectively. All scores are percentages (%). p-values were calculated using Wilcoxon signed-rank test and Z-test.

edness score and another with the lowest. We retained pairs where the higher groundedness score was at 30% greater than the lower score. This filtering resulted in 54 pairs of hypotheses with significant differences in groundedness levels. To validate KnowHD's effectiveness, we involved two domain experts to annotate each pair (80% agreement), selecting the hypothesis they deemed more likely truthful based on the given information. Additionally, GPT-4o was prompted to analyze the same pairs and provide its judgment. Results of this annotation study, summarized in Table 4, report the selection ratio for each group, defined as the proportion of hypotheses in each group identified as more truthful.

The results demonstrate a significant relationship between groundedness scores and the perceived truthfulness of hypotheses. Hypotheses with higher groundedness scores were consistently more likely to be selected as truthful by both human experts and GPT-4o, as indicated by the substantial differences in selection ratios. These findings highlight the utility of KnowHD in distinguishing truthful hypotheses, even in unstructured, open-ended generation tasks. By effectively leveraging groundedness as a criterion, KnowHD provides a robust mechanism for improving the reliability of LLM-generated hypotheses, reinforcing its potential for facilitating real-world scientific discovery processes.

6 Related Work

6.1 Scientific Hypothesis Generation

The use of LLMs for scientific hypothesis generation is a rapidly growing field, leveraging the ability of these models to process and synthesize vast amounts of scientific literature [Qi *et al.*, 2023; Yang *et al.*, 2023; Zhou *et al.*, 2024; Ciucă *et al.*, 2023; Park *et al.*, 2024; Skarlinski *et al.*, 2024; Radensky *et al.*, 2024]. LLMs have been applied in identifying research gaps and generating novel hypotheses, with notable successes in areas such as drug discovery, where generated hypotheses have led to experimentally validated drug combinations [Abdel-Rehim *et al.*, 2024]. Despite these advancements, most existing studies emphasize the novelty and diversity of hypotheses without addressing the critical aspect of truthfulness [Qi *et al.*, 2024; Baek *et al.*, 2024; Wang *et al.*, 2023; Hu *et al.*, 2024; Li *et al.*, 2024; Si *et al.*, 2024]. The prevalent hallucination problem exacerbates this issue, as LLMs often generate hypotheses that appear plausible but lack factual support [Huang *et al.*, 2023]. This gap motivates the development of TruthHypo, a benchmark explicitly designed to assess the ability of LLMs to generate truthful and grounded scientific hypotheses.

6.2 Knowledge Graph Reasoning

Knowledge graph reasoning involves inferring missing facts or relationships within a knowledge graph, with tasks such as link prediction, entity classification, and relation extraction being extensively studied [Lin *et al.*, 2015; Ji *et al.*, 2021; Shu *et al.*, 2024]. Traditional link prediction focuses on predicting edges between entities based on graph structure. These tasks primarily target structured graph completion, emphasizing pattern detection rather than creative reasoning [Liu *et al.*, 2023; Wu *et al.*, 2023; Gu and Krenn, 2024]. TruthHypo introduces a novel benchmark that centers on LLM-driven scientific hypothesis generation, leveraging LLMs' ability to flexibly integrate external knowledge through contextual inputs. Unlike static graph reasoning, TruthHypo evaluates how well LLMs generate grounded and truthful hypotheses. This shift highlights the growing role of LLMs in scientific discovery and bridges the gap between symbolic graph reasoning and natural language creativity.

6.3 Retrieval-augmented Generation

Retrieval-augmented generation (RAG) has emerged as a powerful approach for improving the factual accuracy and relevance of LLM outputs by integrating external knowledge during the generation process. This technique has been applied with literature retrieval, as demonstrated by [Lewis *et al.*, 2020], to dynamically incorporate up-to-date information into model outputs. Retrieval-augmented generation methods enhance the ability of LLMs to ground their outputs in external knowledge, making them particularly valuable in tasks requiring factual accuracy, such as scientific text generation [Lála *et al.*, 2023; Munikoti *et al.*, 2023]. In addition to literature retrieval, retrieval-augmented generation using knowledge graphs has gained attention for its potential to provide structured, domain-specific knowledge during text generation [Peng *et al.*, 2024; Ma *et al.*, 2024; Wang *et al.*, 2025]. TruthHypo builds on this paradigm by integrating both literature and knowledge graph retrieval to provide a robust evaluation of LLMs' ability to generate truthful scientific hypotheses. This dual approach enables a comprehensive analysis of the role of external knowledge in mitigating hallucinations and ensuring the groundedness of generated hypotheses.

7 Conclusion

We presented TruthHypo, a benchmark for evaluating the ability of LLMs to generate truthful scientific hypotheses, and KnowHD, a framework for detecting hallucinations by assessing groundedness in reasoning. Through extensive evaluation, we highlighted the limitations of existing LLMs and demonstrated that selecting highly grounded hypotheses improves truthfulness. These contributions offer valuable insights for improving the reliability and utility of LLMs in scientific discovery.

Acknowledgements

This work is supported in part by the US National Science Foundation under grants 2217071, 2213700, 2106913, 2008208, and NIH grant 1R01LM014012.

References

[Abdel-Rehim *et al.*, 2024] Abbi Abdel-Rehim, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J Collins, Elizabeth Bourne, Gareth W Fearnley, Emma Tate, Holly X Smith, Larisa N Soldatova, et al. Scientific hypothesis generation by a large language model: Laboratory validation in breast cancer treatment. *arXiv preprint arXiv:2405.12258*, 2024.

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Baek *et al.*, 2024] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*, 2024.

[Ciucă *et al.*, 2023] Ioana Ciucă, Yuan-Sen Ting, Sandor Kruk, and Kartheik Iyer. Harnessing the power of adversarial prompting and large language models for robust hypothesis generation in astronomy. *arXiv preprint arXiv:2306.11648*, 2023.

[D'Arcy *et al.*, 2024] Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*, 2024.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Franceschelli and Musolesi, 2024] Giorgio Franceschelli and Mirco Musolesi. On the creativity of large language models. *AI & SOCIETY*, pages 1–11, 2024.

[Gu and Krenn, 2024] Xuemei Gu and Mario Krenn. Forecasting high-impact research topics via machine learning on evolving knowledge graphs. *arXiv preprint arXiv:2402.08640*, 2024.

[Hu *et al.*, 2024] Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*, 2024.

[Huang *et al.*, 2023] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

[Ifargan *et al.*, 2025] Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. Autonomous llm-driven research—from data to human-verifiable research papers. *NEJM AI*, 2(1):AIoa2400555, 2025.

[Ji *et al.*, 2021] Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S Yu Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE transactions on neural networks and learning systems*, 33(2):494–514, 2021.

[Karanikolas *et al.*, 2023] Nikitas Karanikolas, Eirini Manga, Nikoletta Samaridi, Eleni Tousidou, and Michael Vassilakopoulos. Large language models versus natural language understanding and generation. In *Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics*, pages 278–290, 2023.

[Kumar *et al.*, 2023] Varun Kumar, Leonard Gleyzer, Adar Kahana, Khemraj Shukla, and George Em Karniadakis. Mycrunchgpt: A chatgpt assisted framework for scientific machine learning. *arXiv preprint arXiv:2306.15551*, 2023.

[Lála *et al.*, 2023] Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[Li *et al.*, 2024] Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, et al. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *arXiv preprint arXiv:2410.13185*, 2024.

[Lin *et al.*, 2015] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[Liu *et al.*, 2023] Xingyu Liu, Juan Chen, and Quan Wen. A survey on graph classification and link prediction based on gnn. *arXiv preprint arXiv:2307.00865*, 2023.

[Liu *et al.*, 2024] Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Conversational drug editing using retrieval and domain feedback. In *The Twelfth International Conference on Learning Representations*, 2024.

[M. Bran *et al.*, 2024] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11, 2024.

[Ma *et al.*, 2024] Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, and Jian Guo. Think-on-graph 2.0: Deep and interpretable large language model reasoning with knowledge graph-guided retrieval. *arXiv e-prints*, pages arXiv–2407, 2024.

[Munikoti *et al.*, 2023] Sai Munikoti, Anurag Acharya, Sridevi Wagle, and Sameera Horawalavithana. Evaluating the effectiveness of retrieval-augmented large language models in scientific document reasoning. *arXiv preprint arXiv:2311.04348*, 2023.

[Park *et al.*, 2024] Yang Jeong Park, Daniel Kaplan, Zhichu Ren, Chia-Wei Hsu, Changhao Li, Haowei Xu, Sipei Li, and Ju Li. Can chatgpt be used to generate scientific hypotheses? *Journal of Materiomics*, 10(3):578–584, 2024.

[Peng *et al.*, 2024] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*, 2024.

[Qi *et al.*, 2023] Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. Large language models are zero shot hypothesis proposers. *arXiv preprint arXiv:2311.05965*, 2023.

[Qi *et al.*, 2024] Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. Large language models as biomedical hypothesis generators: A comprehensive evaluation. In *First Conference on Language Modeling*, 2024.

[Radensky *et al.*, 2024] Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S Weld. Scideator: Human-llm scientific idea generation grounded in research-paper facet recombination. *arXiv preprint arXiv:2409.14634*, 2024.

[Raiaan *et al.*, 2024] Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 2024.

[Robertson *et al.*, 2009] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[Shu *et al.*, 2024] Dong Shu, Tianle Chen, Mingyu Jin, Chong Zhang, Mengnan Du, and Yongfeng Zhang. Knowledge graph large language model (kg-llm) for link prediction. *arXiv preprint arXiv:2403.07311*, 2024.

[Si *et al.*, 2024] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint arXiv:2409.04109*, 2024.

[Skarlinski *et al.*, 2024] Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

[Wang *et al.*, 2022] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[Wang *et al.*, 2023] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. *arXiv preprint arXiv:2305.14259*, 2023.

[Wang *et al.*, 2025] Shijie Wang, Wenqi Fan, Yue Feng, Xinyu Ma, Shuaiqiang Wang, and Dawei Yin. Knowledge graph retrieval-augmented generation for llm-based recommendation. *arXiv preprint arXiv:2501.02226*, 2025.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[Wei *et al.*, 2024] Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, Robert Leaman, Shubo Tian, Ling Luo, Qiao Jin, Zhizheng Wang, Qingyu Chen, and Zhiyong Lu. Pubtator 3.0: an ai-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*, page gkae235, 2024.

[Wright *et al.*, 2022] Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. Generating scientific claims for zero-shot scientific fact checking. *arXiv preprint arXiv:2203.12990*, 2022.

[Wu *et al.*, 2023] Chaokai Wu, Yansong Wang, and Tao Jia. Dynamic link prediction using graph representation learning with enhanced structure and temporal information. In *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 279–284. IEEE, 2023.

[Yang *et al.*, 2023] Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*, 2023.

[Yang *et al.*, 2025] Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. Large language models for rediscovering unseen chemistry scientific hypotheses. In *2nd AI4Research Workshop: Towards a Knowledge-grounded Scientific Research Lifecycle*, 2025.

[Zeng *et al.*, 2023] Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. Meta-review generation with checklist-guided iterative introspection. *arXiv preprint arXiv:2305.14647*, 2023.

[Zhong *et al.*, 2023] Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. *Advances in Neural Information Processing Systems*, 36:40204–40237, 2023.

[Zhou *et al.*, 2024] Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. *arXiv preprint arXiv:2404.04326*, 2024.