Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Health

THGAgents: Traceable Biomedical Hypothesis Generation via Dynamic Causal Reasoning

Mingjian Yang[1] , **Kun Dong**[1] , **Kevin Lim**[2,†] and **Juan Liu**[1,†]
[1]School of Artificial Intelligence, Wuhan University
[2]MetaStar Health Tech
{ymjcou, kevin_dong, liujuan}@whu.edu.cn, 18581020912@163.com

Abstract

While Large Language Models (LLMs) offer promise in scientific discovery, leveraging LLMs to drive biomedical research requires the scientific discovery process to be performed in combination with cutting-edge biomedical research and rigorous mechanistic causal chains. As such, both current Retrieval-augmented generation methods lacking causal reasoning capabilities, and the static traditional knowledge graphs failing to reflect evolving scientific knowledge, present obstacles to utilizing LLMs as scientific discovery tools. In response to these ongoing challenges, we present THGAgents. THGAgents utilize collaborative and dynamically updating agents to build a Traceable Causal Knowledge Graph, which serves as the foundation for the evidence-based knowledge structure. Crucially, we employ an LLM-driven heuristic search algorithm to traverse the complex network, balancing both novelty and rigor to deduce strict, evidence-based mechanistic causal chains. Additionally, THGAgents utilize a generator-critic loop to support hypothesis refinement. In experimental benchmarks across both cancer systems and neuroscience, THGAgents achieved up to a 0.80 hit rate in predicting validated scientific discoveries, providing an almost 9.5% increase in hypothesis quality scores versus current state-of-the-art systems, and decreasing the mechanistic hallucination rate to 1.12%. Our code is available at https://github.com/yangCode-res/THGAgents/.

1 Introduction

In today's biomedical research, the quantity of information available is increasing at an accelerated pace [Bornmann *et al.*, 2021], outpacing researchers' ability to read, comprehend, and assimilate new findings [Bastian *et al.*, 2010]. Consequently, critical cross-paper connections, such as those between newly discovered biomedical pathways and existing medications, are often overlooked, resulting in delayed or missed opportunities for translational research and drug discovery. A compelling demonstration of this latent knowledge problem emerged during the COVID-19 pandemic: BenevolentAI's knowledge graph, integrating millions of PubMed papers, identified baricitinib as a potential treatment by uncovering a hidden antiviral mechanism through AAK1 kinase inhibition [Richardson *et al.*, 2020]. This AI-driven hypothesis was validated through randomized controlled trials [Kalil *et al.*, 2021] and received FDA approval [Marconi *et al.*, 2021], a trajectory from computational inference to clinical validation that would have been impossible through manual literature review alone.

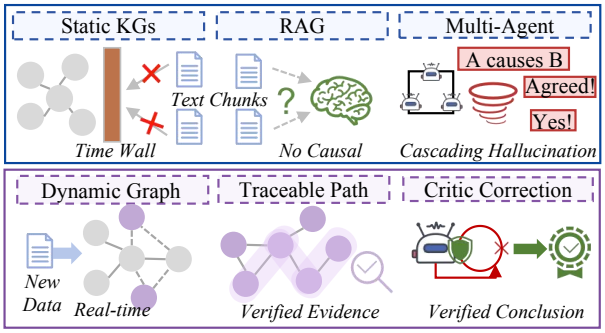

Figure 1: The dynamic, traceable and grounded resolution of THGAgents in comparison to bottlenecks of existing paradigms.

Although Large Language Models (LLMs) are attractive for assisting scientific hypothesis generation [Achiam *et al.*, 2023; Wiggins and Tejani, 2022; Liu *et al.*, 2024a], there is a severe misalignment between how LLMs produce outputs and how scientists develop hypotheses. LLM outputs result from linear token prediction, whereas human scientific reasoning is an iterative process of evaluating multiple candidate mechanisms and restructuring reasoning based on domain knowledge. Consequently, LLMs frequently create false connections and scientifically unfounded hypotheses that appear novel but are hallucinations [Ji *et al.*, 2023; Huang *et al.*, 2025].

To mitigate these issues, current research primarily follows two paradigms: knowledge augmentation via RAG and Knowledge Graphs [Lewis *et al.*, 2020; Hu *et al.*, 2023; Yan *et al.*, 2024], and workflow simulation via Multi-agent sys-

†Corresponding authors: Juan Liu (liujuan@whu.edu.cn), Kevin Lim (18581020912@163.com).

tems [D'Arcy *et al.*, 2024; Baek *et al.*, 2025; Su *et al.*, 2024]. However, as demonstrated in Figure 1, these approaches face critical challenges. Traditional KGs are plagued by latency and superficiality—static structures lagging behind literature growth while oversimplifying complex biological processes. RAG provides external context but relies on semantic matching, lacking mechanistic-level reasoning for causal verification. Multi-Agent frameworks can implement iterative refinement but lack rigorous traceability, without evidence-based structures to anchor their interactions, errors risk cascading amplification.

Our approach relies on two major components: a Traceable Causal Knowledge Graph (TCKG) that can continuously integrate the latest literature, and an auditable reasoning process that uses the TCKG to provide rationale for causal pathways. We develop a scalable method for building TCKGs through collaborative agents that extract, refine, and align causal structures from real-time literature via a coarse-to-fine strategy. To ensure reasoning determinism, we introduce an LLM-driven heuristic search algorithm that navigates the network with a global perspective, scoring candidates across multiple dimensions to deduce evidence-based causal chains. Finally, a generator-critic loop evolves verified pathways into novel hypotheses and experimental designs, ensuring scientific rigor and logical consistency.

Our main contributions are summarized as follows:

- We propose THGAgents, a framework utilizing collaborative agents to dynamically construct TCKGs from real-time literature, ensuring an up-to-date and traceable knowledge base.
- We develop an LLM-driven heuristic search algorithm to navigate the TCKG, deducing rigorous causal chains and refining hypotheses via a generator-critic loop to ensure scientific validity.
- We construct a retrospective benchmark to evaluate hypothesis generation. Extensive experiments demonstrate that our method significantly improves factual grounding and traceability while reducing mechanistic hallucinations.

2 Problem Formulation

In this section, we provide a formal definition of the biomedical hypothesis generation task as a two-stage approach, where the first stage consists of constructing a query-specific TCKG using unstructured literature and the second stage involves reasoning over that TCKG to produce verifiable hypotheses. As part of the formal definition of the biomedical hypothesis generation task, we will begin by defining the major data structures associated with the task.

Scientific Corpus and Query Let $\mathcal{D} = \{d_1, d_2, \ldots, d_{|\mathcal{D}|}\}$ be the global collection of biomedical literature (e.g., PubMed). $\mathcal{V}_{all}$ is the global vocabulary or set of all biomedical entities (proteins, diseases, chemicals, etc.) that exist within the biomedical domain.

Given a specific research query q, the system first retrieves a relevant subset $\mathcal{D}_q \subset \mathcal{D}$. To further break down the documents contained in the retrieved subset into manageable sections, we segment each document d_i into a series of separate text chunks $\mathcal{C}_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,m}\}$. We represent all text chunks retrieved with query q, as $\mathcal{C}_q$.

Traceable Causal Knowledge Graph We introduce the knowledge structure, a directed graph $\mathcal{G}_q = (\mathcal{V}, \mathcal{E})$ which is grounded in the retrieved text chunks $\mathcal{C}_q$ with respect to the query q as TCKG. TCKG consists of:

1. **Nodes ($\mathcal{V}$):** All biomedical entities such as proteins and phenotypes that were extracted from $\mathcal{C}_q$, mapped to a single terminology system ($\mathcal{V}_{all}$).
2. **Edges ($\mathcal{E}$):** Each edge has a different definition from those of traditional knowledge graphs. Specifically, every edge is a quadruple $e = (u, r, v, \omega) \in \mathcal{E}$. Here, $\omega \subseteq \mathcal{C}_q$ is the provenance of the edge, indicating where the evidence for establishing the relationship exists and that the relationship can be traced back to the original text chunks.
3. **Types of Relationships ($\mathcal{R}$):** Relationships can either be correlations (i.e., u relates to v) or causal relationships (i.e., u causes v). Thus, we used two distinct relationship sub-types: symmetric associations ($\mathcal{R}_{bi}$) and asymmetric causal effects ($\mathcal{R}_{causal}$). A relationship with $r \in \mathcal{R}_{causal}$ implies there is a directed relationship ($u \rightarrow v$) supported by evidence gathered from the extraction confidence, whereas $r \in \mathcal{R}_{bi}$ indicates both u and v are correlated.

Hypothesis Generation Task The goal of our system is to generate a scientific hypothesis H comprising a natural language statement and an experimental plan. We formulate this as a path-constrained generation problem.

First, we aim to construct TCKG $\mathcal{G}_q$ from unstructured chunks $\mathcal{C}_q$:

$$\mathcal{G}_q = f_{\text{construct}}(\mathcal{C}_q, q), \tag{1}$$

where $f_{\text{construct}}$ denotes the collaborative extraction and alignment process from respective agent modules.

Next, the collection of reasoning paths is represented as a set $\Pi = \{\pi_1, \ldots, \pi_k\}$ from $\mathcal{G}_q$. The set contains paths π_j constructed from the collections of entities as well as the causal relationships for each reasoning path from the query-anchored nodes. The purpose of this step is to derive the best hypothesis H^*, with accordance to maximizing the following equation:

$$H^* = \arg\max_{H} \sum_{\pi \in \Pi} P(H \mid \pi, \omega_\pi, q) \cdot P(\pi \mid \mathcal{G}_q, q). \tag{2}$$

Here, the term $P(\pi \mid \mathcal{G}_q, q)$ indicates how logically compelling and innovative a particular reasoning path is, and $P(H \mid \pi, \omega_\pi, q)$ models the generation likelihood of deriving the hypothesized result from the information provided in the particular reasoning path along with its related evidential support ω_π.

3 The THGAgents Framework

3.1 System Overview

We present **THGAgents**, a multi-agent framework for automating the creation of traceable biomedical hypotheses. As

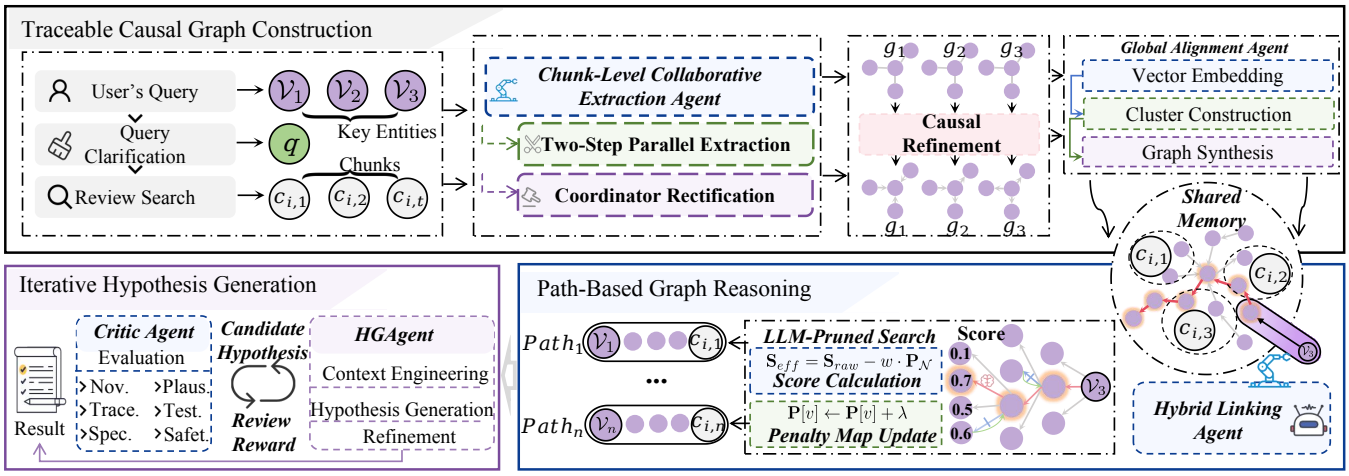

Figure 2: The overall architecture of the THGAgents framework. The workflow consists of three integrated phases: (1) TCKG Construction: A multi-agent team extracts triplets from literature chunks, refines causal directionality, and performs global alignment to build the TCKG. (2) Path-Based Graph Reasoning: Starting from query-anchored entities, the reasoning agent utilizes an LLM-pruned heuristic search—incorporating an effective scoring method and a dynamic penalty map to deduce rigorous mechanistic causal chains. (3) Iterative Hypothesis Generation: Validated paths are synthesized into candidate hypotheses, which undergo adversarial evaluation and refinement via a generator-critic loop to ensure scientific rigor and grounding.

depicted in Figure 2, the framework could be organized into three parts that communicate with one another using a shared memory structure and follow a pipeline architecture. The workflow of THGAgents consists of three phases that run in sequential order: (1) TCKG construction, which involves the conversion of unstructured literature into TCKGs, (2) Path-Based Reasoning, where mechanistic causal chains are discovered and validated, and (3) Iterative Hypothesis Generation, which generates and revises the final hypotheses.

3.2 Traceable Causal Graph Construction

This phase transforms retrieved documents $\mathcal{D}_q$ into TCKG $\mathcal{G}_q$. We break this process down into three hierarchical steps: chunk-level collaborative extraction, causal directionality refinement, and global alignment.

Chunk-Level Collaborative Extraction To resolve the problem of entity-relation misalignment that typically occurs in long texts, we implement a multi-agent team approach where each text chunk $c_i \in \mathcal{C}_q$ is parsed into a local subgraph. *Collaborative Rectification* mechanism:

1. **Two-Step Parallel Extraction:** Rather than performing single-pass operations, both the *Entity Agent* and *Relation Agent* carry out their tasks in parallel in a coarse-to-fine manner. Both agents scan the context to identify the existence of specific entity types (e.g., chemicals) and relation types (e.g., up-regulation). Conditioned on the detected types, they can conduct targeted extraction of actual biomedical terms and predicate types of the interactions. After terms and types are identified, the agents then source the extracted triplets from the source documents and maintain traceability of the subgraphs.
2. **Coordinator Rectification:** A *Coordinator Agent* collects the parallel results and constructs valid triplets. It performs argument assignment of the extracted entities to the head and tail slots of identified relations. Where there may be multiple possible candidates for head or tail entities, the coordinator agent will then utilize an LLM-based adjudication process to identify which entities are the best candidates based on semantic context. Additionally, to completely eliminate probable hallucinated relations, any triplet identified as incomplete (i.e., does not have both head and tail connected to the identified relation) after the adjudication process is discarded from the final subgraph g_i.

Causal Directionality Refinement We perform causal refinement prior to merging each subgraph g_i. The *Causal Critic Agent* evaluates the surrounding semantic context of each extracted triplet in order to measure how much influence an extracted triplet has on one or more entities. The causal critic agent can therefore discriminate between mechanistic correlation and causation. We classify edges into bidirectional associations ($u \leftrightarrow v$) or unidirectional causal effects ($u \rightarrow v$). This process is crucial for filtering out mere co-occurrences that may confuse experimental design.

Global Alignment Agent To unify subgraphs $\{g_i\}$ into $\mathcal{G}_q$, we implement an efficient incremental alignment strategy:

1. **Semantic Encoding:** We utilize SapBERT [Liu *et al.*, 2021] to generate initial embeddings for each entity, which provides superior capture of fine-grained biomedical semantics. For nodes within a local subgraph, their representations are briefly aggregated with immediate neighbors to retain basic relational information.
2. **Incremental Subgraph Fusion:** Subgraphs are integrated into the global graph sequentially. To optimize efficiency, the smaller subgraph acts as an anchor to find potential matches within the existing global structure.
3. **Coarse-to-Fine Adjudication:** For each node in the an-

chor subgraph, a vector-based coarse search retrieves the top-k most similar candidates, followed by an LLM-based fine adjudication for final identity alignment. This reduces LLM invocation overhead and supports real-time TCKG updates.

Eventually, we synthesize the local refined subgraphs into a single unified graph $\mathcal{G}_q$. This space will provide navigation for the following reasoning process.

3.3 Path-Based Graph Reasoning

The next phase of the process involves traversing the causal model of all entities in the graph $\mathcal{G}_q$ to find all legitimate reasoning chains from one or more core entities of a user query q. We introduce a hybrid anchoring-and-search mechanism for seeking the best paths to each entity for analysis.

Hybrid Query Anchoring The first thing the *Reasoning Agent* does when receiving a user query q is to determine the core entities and which graph nodes they should be linked to as $\mathcal{V}$. To provide access to all potential core entities, we use a hybrid linking strategy that incorporates both the idea of lexical distance with semantic similarity.

For each query keyword k and each node v in TCKG, we compute a composite score:

$$S_{\text{link}}(k, v) = \alpha \cdot \text{Sim}_{\text{lex}}(k, v) + (1-\alpha) \cdot \text{Sim}_{\text{sem}}(\mathbf{h}_k, \mathbf{h}_v), \quad (3)$$

where Sim_{lex} denotes the lexical overlap and Sim_{sem} denotes the cosine similarity between their embeddings. Subsequently, the LLM performs an adjudication process on these candidates to resolve ambiguities and verify their alignment with the query context. The finalized nodes form the anchor set $\mathcal{V}_q$. These verified nodes serve as the entry points for subsequent graph reasoning.

LLM-Pruned Heuristic Search As formalized in Algorithm 1, we perform a multi-hop search starting from the identified anchor set $\mathcal{V}_q$. To efficiently navigate TCKG and prevent logical drift, we introduce a dynamic scoring method.

At each expansion step, the reasoning agent performs batch adjudication on candidate neighbors. To each transition, the agent evaluates the next step by synthesizing the user query q, the preceding reasoning chain π_{curr}, and the supporting evidence chunk ω_e. A raw validity score is assigned based on three critical scientific dimensions: plausibility, novelty and testability. Crucially, to learn from dead ends and avoid repetitive failures, we calculate an effective score (S_{eff}) by subtracting a dynamic penalty term $P[v]$ from raw LLM score.

The search greedily expands the branch with the highest S_{eff}. If the best candidate fails to meet the validity threshold, the branch is pruned, and the penalty $P[v_{next}]$ is incremented to inhibit future traversal. This ensures that the system self-corrects during exploration, filtering out irrelevant connections while focusing resources on the most scientifically meaningful chains.

3.4 Iterative Hypothesis Generation

As a final step in the iterative hypothesis-generation process, all of the valid hypotheses from the previous steps are combined into one coherent scientific proposal. For this phase, we use the generator-critic loop, which reflects the iterative process of synthesis and refinement.

Algorithm 1 LLM-Pruned Heuristic Search

Input: Graph TCKG, Query q, Depth d, Weights $\{w, \lambda, \tau\}$
Output: Set of Validated Paths Π^*

Initialize: $\Pi^* \leftarrow \emptyset$, Initial nodes $\mathcal{V}_q \leftarrow \text{Anchor}(q)$
Initialize penalties $P[v] \leftarrow 0, \forall v \in \mathcal{V}$
for all $v_{start} \in \mathcal{V}_q$ **do**
  SEARCH($\pi_{curr} = [v_{start}]$)
end for
return Π^*

Function SEARCH(π_{curr})
if $|\pi_{curr}| = d$ **then**
  $\Pi^* \leftarrow \Pi^* \cup \{\pi_{curr}\}$ {Goal reached}
  return
end if
$\mathcal{N} \leftarrow \text{GetNeighbors}(\text{Last}(\pi_{curr})) \setminus \pi_{curr}$
$\mathbf{S}_{raw} \leftarrow \text{LLM}_{judge}(\mathcal{N} \mid q, \pi_{curr})$ {LLM Scoring}
$\mathbf{S}_{eff} \leftarrow \mathbf{S}_{raw} - w \cdot \mathbf{P}_{\mathcal{N}}$ {Apply Penalty}
if $\max(\mathbf{S}_{eff}) > \tau$ **then**
  Sort $\mathcal{N}_{valid}$ by $\mathbf{S}_{eff}$ descending {Prioritize best candidates}
  for all $v_{next} \in \mathcal{N}_{valid}$ **do**
    SEARCH($\pi_{curr} \cup \{v_{next}\}$)
  end for
else
  $P[v] \leftarrow P[v] + \lambda, \forall v \in \pi_{curr}$ {Backtrack Penalty}
end if
End Function

Evidence-Aware Synthesis The *Writer Agent* acts as the generator, integrating the structured path information Π^* with their source evidence chunks ω. By performing contextual integration, the agent generates a comprehensive output package containing: (i) a candidate hypothesis describing the mechanism; (ii) an estimated confidence score; and (iii) a detailed experimental plan for verification. This step ensures the generation is strictly conditioned on retrieved facts, modeled as $H_{cand} \sim P_{\text{gen}}(\cdot|\Pi^*, \omega)$.

Critic-Driven Refinement To ensure scientific rigor, we establish an adversarial feedback loop. A Critic Agent evaluates H_{cand} against dimensions such as groundedness and novelty. If the quality score falls below a threshold, the Critic provides textual feedback, prompting the Writer to regenerate the hypothesis. This iterative optimization continues until the hypothesis is deemed logically sound and supported by the evidence, yielding the final refined hypothesis H_{final}.

4 Experiments

4.1 Benchmark

We assess THGAgents from a temporal perspective by creating a temporal split benchmark that emulates forward-leaning discovery. To ensure that we align the model's knowledge boundaries with those of many models (e.g., GPT-4o), we established a cutoff date of October 2023. We selected 60 queries from "research gaps" we identified in 30 PubMed reviews (January 2021 - September 2023). To strictly prevent data leakage, our system's retrieval scope was restricted

exclusively to literature published prior to this cutoff, ensuring the model deduces insights without accessing future evidence. Ground truth was established using empirical studies published after October 2023 that resolved these specific gaps.

In order to assess generalizability across disciplines, the benchmark evaluates two of the most complex and rapidly evolving research areas: Cancer Immunotherapy (focusing on immune evasion) and Neuroscience (focusing on the gut-microbiome-brain axis). Both areas present excellent environments for validating automated hypothesis generation due to their mechanistic complexity and rapid pace of discovery.

To minimize the impact of evaluator bias and to achieve the best level of rigor in our evaluation, our evaluation process uses the independent expert committee model. Instead of using aggregate scoring methods that utilize averaging and voting, which can create potential bias in evaluators, we have taken evaluations from numerous frontier LLMs (including, but not limited to, GPT-5, Gemini-3 Pro and DeepSeek-V3.2 [Singh *et al.*, 2025; Google, 2025; Liu *et al.*, 2025]) and treat them as separate independent expert judges. Furthermore, to validate the reliability of LLM-based evaluation, we conducted a human evaluation study with four independent domain experts (Section 4.6), demonstrating strong agreement with the LLM judges.

Our system generates a set of three candidate hypotheses $H = \{h_1, h_2, h_3\}$ for each query and evaluates them against the ground truth T and biomedical knowledge $\mathcal{K}$ from independent judges $j \in \mathcal{J}$.

Hit Rate (HR): Measures the degree to which H logically aligns with ground truth T. HR of H is computed as:

$$HR_j = \max_{h \in H}(Score_j(h \approx T)) \in [0, 1]. \tag{4}$$

Mechanistic Hallucination Rate (MHR): Quantifies the average number of mechanistic errors across H. Let $\mathcal{M}_h$ represent the set of N atomic mechanisms that make up hypothesis h, and $\mathbb{I}(\cdot)$ represent an indicator function that returns a value of 1 if mechanism m contradicts the established biomedical knowledge base $\mathcal{K}$. MHR of H is computed as:

$$MHR_j = \frac{1}{|H|} \sum_{h \in H} (\frac{1}{N} \sum_{m \in \mathcal{M}_h} \mathbb{I}(m \perp \mathcal{K})). \tag{5}$$

Quality Score (QS): Every judge will utilize a 5-point Likert scale across six dimensions $\mathcal{D}$ (Novelty, Plausibility, Grounding, Testability, Specificity, and Safety and Ethics). QS of H is computed as:

$$QS_j = \frac{1}{|H|} \sum_{h \in H} (\frac{1}{6} \sum_{d \in \mathcal{D}} (Score_j(d)). \tag{6}$$

Finally, we will provide independent scores from judges in order to demonstrate that the superior performance of the system is recognized consistently across different expert models.

4.2 Experimental Setup

Baseline THGAgents' effectiveness will be tested against four typical baseline comparisons. Parametric [Wei *et al.*, 2022] relies exclusively on the LLM's pre-trained parametric knowledge without accessing external databases. NaiveRAG [Gao *et al.*, 2023] retrieves the top-k semantically similar text chunks via vector embeddings to provide unstructured external context. LightRAG [Guo *et al.*, 2024] combines local structural context of the entities in addition to the textual context retrieved by the RAG method. Finally, ResearchAgent [Baek *et al.*, 2025] utilizes a multi-agent pipeline to generate and refine ideas using an iterative peer-review process starting from a key source or paper.

Parameters Settings In building the graph structure, we included chunks that were 1200 tokens or less in size and split them at the sentence level. The cosine similarity with a threshold of 0.5 and Top-K = 5 were used for matching entities. Using the entity linking methodology, we selected the three most relevant candidate nodes for each entity and set α = 0.5. The path-based reasoning phase was conducted with a depth limit d of five paths and a penalty weight w of 0.5, with three different hypotheses generated for independent verification by the committee of experts.

4.3 Main Results

Table 1 presents a comprehensive performance comparison between THGAgents and various baselines across two distinct biomedical domains. The HR of THGAgents demonstrates the potential for generating hypotheses aligned with later validated discoveries in extremely mechanistically complex fields of cancer immunotherapy and neuroscience involving multi-system interactions. Our method gained the highest HR in evaluations by all judges. For instance, on DeepSeek's evaluation of THGAgents, an unprecedented HR of 0.80 from Group 0 was achieved, which far outperformed the state-of-the-art ResearchAgent. This enables the THGAgent framework to show useful support in the studied domains and its capability to handle rapidly changing streams of information from the most current literature. Additionally, it produces scientific evolutions of theory that extend the boundaries of information provided in train data. Therefore, THGAgent produces verifiable hypotheses that closely align with what is possible for later empirical findings.

Regarding QS, THGAgents also established a significant advantage. Unlike baselines that often generate general content lacking logical support, our LLM-based heuristic path search simulates the stepwise decision-making process of human scientists. This method provides a means to balance novelty and plausibility within complex knowledge networks, ultimately generating high-quality and verifiable hypotheses. Quantitatively, compared to the best-performing baseline, THGAgents achieved an approximate 9.5% improvement in average quality score. It performed particularly well in specificity and grounding. For instance, it reaches a grounding score of 4.23 in Group 0 evaluated by Gemini, far exceeding Parametric's 3.82. Notably, the novelty scores are relatively conservative. This stems from the strict temporal cutoff. Consequently, judge models with updated knowledge often perceive our historical-data-driven insights as established facts rather than novel hypotheses. Despite differences in architecture among the four judge models, they reached a consensus on the superiority of THGAgents, which also aligns

Method	Group 0								Group 1							
	HR	QS	Nov.	Plaus.	Grnd.	Test.	Spec.	Safe.	HR	QS	Nov.	Plaus.	Grnd.	Test.	Spec.	Safe.
Judge: Human Experts (n=4)																
Parametric	0.52	3.42	2.50	3.58	3.17	4.08	2.67	4.52	0.61	3.38	2.33	3.67	3.04	4.00	2.58	4.67
NaiveRAG	0.54	3.40	2.42	3.63	3.08	4.04	2.58	4.67	0.64	3.34	2.26	3.71	2.96	3.92	2.50	4.71
LightRAG	0.56	3.48	2.46	3.58	3.26	4.08	2.76	4.71	0.66	3.36	2.29	3.63	3.04	3.96	2.54	4.67
ResearchAgent	0.55	3.60	2.76	3.50	3.58	4.17	3.17	4.42	0.67	3.67	2.67	3.54	3.67	4.26	3.42	4.46
Ours	**0.63**	**3.94**	**2.92**	**3.76**	**3.92**	**4.38**	**3.71**	**4.96**	**0.71**	**3.98**	**2.88**	**3.79**	**4.04**	**4.42**	**3.88**	**4.88**
Judge: Gemini3-Pro-Preview																
Parametric	0.67	3.77	2.21	3.92	3.82	4.64	3.11	**4.89**	0.74	3.83	2.08	4.20	3.71	4.77	3.30	4.93
NaiveRAG	0.70	3.76	2.16	3.97	3.81	4.68	3.07	4.84	0.80	3.77	2.03	**4.24**	3.60	4.64	3.14	**4.97**
LightRAG	0.67	3.80	2.12	3.93	3.89	**4.71**	3.26	4.87	**0.82**	3.74	2.02	4.18	3.59	4.58	3.11	4.96
ResearchAgent	0.64	3.80	2.34	3.71	4.02	4.68	3.44	4.62	0.79	4.00	2.34	3.90	4.09	4.73	4.02	4.89
Ours	**0.75**	**3.96**	**2.30**	**4.04**	**4.23**	4.67	**3.84**	4.69	**0.82**	**4.27**	**2.57**	4.01	**4.39**	**4.96**	**4.74**	4.93
Judge: GPT-5																
Parametric	0.49	3.17	2.40	3.06	2.87	3.72	2.47	**4.50**	0.62	3.12	2.27	3.20	2.60	3.61	2.41	4.60
NaiveRAG	0.50	3.16	2.38	**3.09**	2.82	3.77	2.48	4.44	0.65	3.11	2.24	**3.27**	2.59	3.58	2.36	**4.63**
LightRAG	0.54	3.16	2.34	3.03	2.86	3.71	2.59	4.43	0.67	3.08	2.24	3.26	2.51	3.59	2.29	4.59
ResearchAgent	0.53	3.13	2.63	2.89	2.90	3.63	2.69	4.03	0.70	3.29	2.58	3.10	2.91	3.72	3.13	4.28
Ours	**0.58**	**3.47**	**2.90**	2.98	**3.03**	**3.87**	**3.67**	4.38	**0.70**	**3.54**	**2.78**	3.17	**3.12**	**3.89**	**3.67**	4.59
Judge: DeepSeek-V3.2																
Parametric	0.58	3.63	2.43	3.92	3.63	4.28	2.84	4.67	0.58	3.66	2.43	3.92	3.63	4.28	2.84	**4.88**
NaiveRAG	0.63	3.58	2.33	3.93	3.57	4.20	2.73	**4.69**	0.64	3.49	2.18	3.98	3.22	4.09	2.66	4.81
LightRAG	0.57	3.60	2.32	3.92	3.59	4.22	2.92	4.63	0.59	3.48	2.22	3.94	3.29	4.08	2.54	4.80
ResearchAgent	0.53	3.84	2.80	3.82	3.94	4.62	3.48	4.37	0.64	4.02	2.61	3.96	4.13	4.96	4.04	4.39
Ours	**0.80**	**4.39**	**3.99**	**4.01**	**4.68**	**4.92**	**4.24**	4.49	**0.65**	**4.32**	**3.34**	**3.99**	**4.57**	**4.98**	**4.26**	4.78

Table 1: Quantitative results of THGAgents and baselines across two domains. Scores are reported for Group 0 (Neuroscience) and Group 1 (Cancer Immunotherapy) as assessed by the independent expert committee. Bold values indicate the best results per metric.

with the preferences observed in our human evaluation. Despite differences in architecture among the four judge models, they reached a consensus on the superiority of THGAgents. This consensus is further validated by human expert evaluation (Section 4.6), with an average ranking correlation of Kendall's $\tau = 0.80$.

4.4 Ablation Studies

To evaluate how well the essential components of the THGAgents perform, we conducted an ablation test as shown in Table 2. For the THG Agent$_{l_}$ comparison, we replaced the LLM computation of S_{eff} for goal-directed path selection with a random walk method. This resulted in a dramatic reduction in HR. Without being aided by S_{eff} and having no penalty for choosing dead-end paths, random walk lacks both goal orientation and a productive means of traversing the intricate causal graph. Therefore, it fails to create original hypotheses that are logically related. In the THG Agent$_{c_}$ comparison, we eliminated the Causal Critic Agent, which determines causality versus correlation. Consequently, this ablation treated the graph as a standard undirected knowledge graph. As a result, this variant exhibited an increased MHR, which supports the hypothesis that treating biological correlations as co-occurrence causes the model to confuse correlation and causality and to develop likely plausible yet scientifically causality-reversed hypotheses. For the THG Agent$_{r_}$ comparison, we removed the adversarial feedback mechanism, which resulted in lower QS. While the initial generation of hypotheses achieved the intended concept, there was a deficiency in logical depth and evidence that would have been provided via adversarial feedback. Thus, our study supports the notion that iterative optimization is instrumental in the evolution of hypotheses.

Method	Group 0			Group 1		
	HR	QS	MHR	HR	QS	MHR
THGAgents$_{l_}$	0.64	4.14	1.90	0.60	4.24	2.25
THGAgents$_{c_}$	0.69	4.17	3.80	0.63	4.19	2.10
THGAgents$_{r_}$	0.75	4.04	2.06	0.63	4.10	2.45
Ours	**0.80**	**4.39**	**1.12**	**0.65**	**4.32**	**1.84**

Table 2: Ablation study results across three variants. HR and QS are evaluated by the DeepSeek-V3.2 judge, while MHR is assessed by Gemini-3-Flash-Preview.

4.5 Mechanistic Hallucination Analysis

Figure 3 demonstrates that THGAgents achieves the best balance between TM and MHR in comparison to the performance of the Parametric model, RAGs, and agentic systems. The highest observed TM for THGAgents demonstrates that this framework can create accurate, fully mechanistic, coherent hypotheses while simultaneously producing an MHR of approximately 2.7%, which is the lowest of all. ResearchAgent produces an impressive number of mechanisms at a high TM. Nonetheless, as a general-purpose system, the Re-

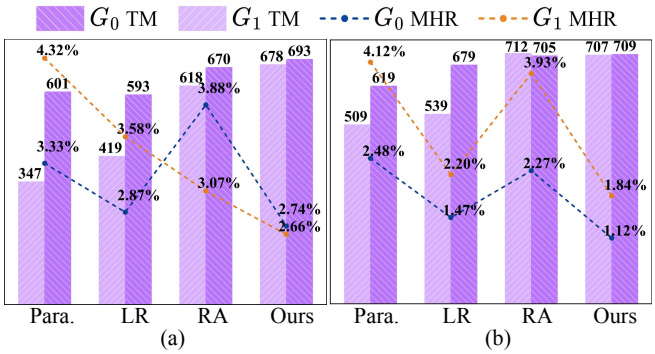

Figure 3: Evaluation of MHR and TM across two groups. The results are based on evaluations by (a) GPT-5 and (b) Gemini-3-Flash-Preview. Para., LR, and RA denote Parametric, LightRAG, and ResearchAgent, respectively. TM denotes the total count of detected mechanisms and MHR denotes the mechanistic hallucination rate.

searchAgent does not strictly control for the occurrence of biomedical hallucinations. As evidence, the ResearchAgent produced a MHR of approximately 3.9%, which highlights how the implementation of specific domain causal constraints helps to maintain the scientific rigor of the hypothesis. However, due to the limitations of structural identification, the use of RAGs often leads to broader, more general hypotheses when compared to the previous methods of constructing causal mechanisms. The vast increase in THGAgents' ability to produce hypotheses that have low or no hallucinations results from our systematic verification process during TCKG construction phase and primarily the use of a coordinator agent to reject any triplet without an explicit basis in the source text. In addition, a more rigorous causal relationship evaluation of the graph construction is done via the causal critic agent, which distinguishes causation from correlation in the causal graphs produced.

4.6 Human Evaluation

To validate the reliability of LLM-based evaluation, we conducted a human study with four domain experts (two in cancer immunotherapy, two in neuroscience). Experts evaluated a stratified sample of 30 hypotheses across the six quality dimensions, blinded to method identities. As shown in Table 1, human evaluations closely mirror LLM assessments, with THGAgents ranked first across all dimensions. The average Kendall's τ between human and LLM rankings is 0.80, indicating strong agreement.

5 Related Work

5.1 LLM-Driven Scientific Hypothesis Generation

Recent advancements indicate that LLM-based hypothesis generation has evolved from standalone generative methods to knowledge-augmented approaches, and further to automated multi-agent systems. Early studies focused on producing and harnessing the creative abilities of LLMs. Findings from the LLM format examined the ability of LLMs to create new hypotheses by recombining patterns they learned during the pre-training phase incrementally [Zhou *et al.*, 2024; Qi *et al.*, 2023; Ciucă *et al.*, 2023]. While these models can produce seemingly plausible scientific narratives, they suffer from a validity gap, often yielding hallucinated mechanistic linkages [Xiong *et al.*, 2025; Si *et al.*, 2024] rather than actionable scientific leads. To bridge this validity gap, researchers have introduced external knowledge constraints. RAG [Li *et al.*, 2024] integrates contextual evidence from source corpora, while structured approaches such as Causal Knowledge Graphs [Tong *et al.*, 2024] enforce logical consistency. However, their reliance on static knowledge snapshots inevitably results in knowledge lag within rapidly evolving scientific domains. More recently, multi-agent frameworks have attempted to automate the research workflow [Baek *et al.*, 2025; Liu *et al.*, 2024b; Ifargan *et al.*, 2025; Hu *et al.*, 2024] by simulating human collaboration to iteratively refine ideas. Nevertheless, such iterative processes risk triggering cascading hallucinations. In this context, existing approaches generally fail to provide outlined causal mechanisms in specific evidentiary sentences. Thus, the underlying reasoning chain is inadequately specified, requiring more specificity.

5.2 Knowledge Graph Reasoning

KG reasoning aims to infer implicit relations by discovering interpretable paths that connect query entities. Early work primarily relied on random walk-based strategies [Lao and Cohen, 2010; Mazumder and Liu, 2017; Zhu *et al.*, 2023]. While effective, these approaches do not explicitly account for semantic alignment with the target triple. To address these limitations, reinforcement learning-based methods cast KG reasoning as a sequential decision-making problem [Xiong *et al.*, 2017; Das *et al.*, 2017; Hou *et al.*, 2021], navigating graphs with query embeddings. However, these models may struggle when no finite-length path exists between the query and target entities. Graph-CoT [Jin *et al.*, 2024] leverages LLM to conduct step-wise reasoning on graphs, implicitly constraining the search process by selecting coherent intermediate steps. ToG [Sun *et al.*, 2023] further formulates graph reasoning as an interactive exploration process, where the LLM proposes and evaluates candidate transitions to steer the search toward semantically relevant paths. Nevertheless, prior LLM-based methods often lack a global perspective. They fail to formulate multi-step KG reasoning as a controlled search process integrated with query-aware anchoring and evidence-grounded path validation.

6 Conclusion

We developed the THGAgents framework, which enables rigorous and evidence-based hypothesis generation through the dynamic creation of TCKGs combined with heuristic searching techniques. The results of experiments confirm the capability of THGAgents to reduce hallucinations and improve the quality of generated hypotheses, thus demonstrating significant potential for advancing biomedical scientific discovery. In future work, we will extend our benchmark to more biomedical domains and incorporate stronger structural constraints to further improve the reliability of the framework.

Ethics Statement

This framework was developed to support researchers rather than clinicians, using only publicly available literature and no patient records. We acknowledge potential dual-use risks and publication bias, and recommend that all outputs be experimentally validated through responsible audits.

Acknowledgments

The work was supported by the Intelligent Computing Center of the National Cybersecurity Talent and Innovation Base, Wuhan and the Translational Medicine and Interdisciplinary Research Joint Fund of Zhongnan Hospital of Wuhan University (ZNJC202226).

Contribution Statement

Mingjian Yang and Kun Dong contributed equally to this work.

References

[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.*, 2025] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, 2025.

[Bastian *et al.*, 2010] Hilda Bastian, Paul Glasziou, and Iain Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9):e1000326, 2010.

[Bornmann *et al.*, 2021] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15, 2021.

[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.

[Das *et al.*, 2017] Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. Go for a walk and arrive at the answer: Reasoning over paths in knowledge bases using reinforcement learning. *arXiv preprint arXiv:1711.05851*, 2017.

[Gao *et al.*, 2023] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

[Google, 2025] Google. Gemini 3 pro - best for complex tasks and bringing creative concepts to life. https://deepmind.google/models/gemini/pro/, 2025.

[Guo *et al.*, 2024] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024.

[Hou *et al.*, 2021] Zhongni Hou, Xiaolong Jin, Zixuan Li, and Long Bai. Rule-aware reinforcement learning for knowledge graph reasoning. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 4687–4692, 2021.

[Hu *et al.*, 2023] Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1413–1430, 2023.

[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.*, 2025] 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. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[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.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[Jin *et al.*, 2024] Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, et al. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. *arXiv preprint arXiv:2404.07103*, 2024.

[Kalil *et al.*, 2021] Andre C Kalil, Thomas F Patterson, Aneesh K Mehta, et al. Baricitinib plus remdesivir for hospitalized adults with covid-19. *New England Journal of Medicine*, 384(9):795–807, 2021.

[Lao and Cohen, 2010] Ni Lao and William W Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine learning*, 81(1):53–67, 2010.

[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.

[Liu *et al.*, 2021] Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. Learning domain-specialised representations for cross-lingual biomedical entity linking. *arXiv preprint arXiv:2105.14398*, 2021.

[Liu *et al.*, 2024a] Haoyang Liu, Yijiang Li, Jinglin Jian, Yuxuan Cheng, Jianrong Lu, Shuyi Guo, Jinglei Zhu, Mianchen Zhang, Miantong Zhang, and Haohan Wang. Toward a team of ai-made scientists for scientific discovery from gene expression data. *arXiv preprint arXiv:2402.12391*, 2024.

[Liu *et al.*, 2024b] 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.

[Liu *et al.*, 2025] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

[Marconi *et al.*, 2021] Vincent C Marconi, Athimalaipet V Ramanan, Sujatro de Bono, et al. Efficacy and safety of baricitinib for the treatment of hospitalised adults with covid-19 (cov-barrier): a randomised, double-blind, parallel-group, placebo-controlled phase 3 trial. *The Lancet Respiratory Medicine*, 9(12):1407–1418, 2021.

[Mazumder and Liu, 2017] Sahisnu Mazumder and Bing Liu. Context-aware path ranking for knowledge base completion. *arXiv preprint arXiv:1712.07745*, 2017.

[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.

[Richardson *et al.*, 2020] Peter Richardson, Ivan Griffin, Catherine Tucker, Dan Smith, Olly Oechsle, Anne Phelan, Michael Rawling, Edward Savory, and Justin Stebbing. Baricitinib as potential treatment for 2019-ncov acute respiratory disease. *The Lancet*, 395(10223):e30–e31, 2020.

[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.

[Singh *et al.*, 2025] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[Su *et al.*, 2024] Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jinzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. 2024.

[Sun *et al.*, 2023] Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*, 2023.

[Tong *et al.*, 2024] Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. Automating psychological hypothesis generation with ai: when large language models meet causal graph. *Humanities and Social Sciences Communications*, 11(1):1–14, 2024.

[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.

[Wiggins and Tejani, 2022] Walter F Wiggins and Ali S Tejani. On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence*, 4(4):e220119, 2022.

[Xiong *et al.*, 2017] Wenhan Xiong, Thien Hoang, and William Yang Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. *arXiv preprint arXiv:1707.06690*, 2017.

[Xiong *et al.*, 2025] Guangzhi Xiong, Eric Xie, Corey Williams, Myles Kim, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 7849–7857. ijcai.org, 2025.

[Yan *et al.*, 2024] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. 2024.

[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.

[Zhu *et al.*, 2023] Zhaocheng Zhu, Xinyu Yuan, Michael Galkin, Louis-Pascal Xhonneux, Ming Zhang, Maxime Gazeau, and Jian Tang. A* net: A scalable path-based reasoning approach for knowledge graphs. *Advances in Neural Information Processing Systems*, 36:59323–59336, 2023.