From Reasoning to Learning: A Survey on Hypothesis Discovery and Rule Learning with Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Since the advent of Large Language Models (LLMs), efforts have largely focused on improving their instruction-following and deductive reasoning abilities, leaving open the question of whether these models can truly discover new knowledge. In pursuit of artificial general intelligence (AGI), there is a growing need for models that not only execute commands or retrieve information but also learn, reason, and generate new knowledge by formulating novel hypotheses and theories that deepen our understanding of the world. Guided by Peirce's framework of abduction, deduction, and induction, this survey offers a structured lens to examine LLM-based hypothesis discovery. We synthesize existing work in hypothesis generation, application, and validation, identifying both key achievements and critical gaps. By unifying these threads, we illuminate how LLMs might evolve from mere "information executors" into engines of genuine innovation, potentially transforming research, science, and real-world problem solving.

1 Introduction

One major pillar of human intelligence is the capacity to discover hypotheses and learning rules. We call this capability *hypothesis discovery* (or rule learning). Earlier AI systems struggled with it because formal symbolic methods lacked the commonsense background needed for inventive rule formation (Yu et al., 2024a). Recent advances in natural language processing (NLP) have produced LLMs pretrained on extensive text corpora that embed substantial commonsense knowledge. These models now enable tasks that demand rich background knowledge, such as formulating new hypotheses and deriving novel conclusions.

Hypothesis discovery inherently relies on a blend of reasoning that includes abduction, induction, and deduction, each defined differently by various scholars. For instance, Gilbert H. Harman considers induction to be a special case of abduction, describing it as "inference to the best explanation" (IBE) (Harman, 1965; Douven, 2021). However, while this definition is easy to understand, it oversimplifies key aspects of hypothesis discovery. In particular, the notion of the "best" explanation is ambiguous and often requires additional assumptions that vary by context. Moreover, this framework does not fully capture real-world scenarios, where a "best" explanation is rarely reached immediately; rather, we continually experiment, gather new observations, and refine our hypotheses. Based on these considerations, we adopt Charles Peirce's definition of hypothesis discovery and reasoning, which posits that hypothesis discovery begins with forming an explanatory hypothesis to explain observations through **abduction**, proceeds with iteratively apply hypothesis to solve problem or derive new knowledge with **deduction**, and validate hypothesis through **induction** (Frankfurt, 1958; Peirce, 1974; Burks, 1946; Minnameier, 2004) (See explanation in Figure 2).

The rest of the survey is organized as follows. Section 2 presents background knowledge on hypothesis discovery using LLMs, including different forms of reasoning and representations involved in the process. Section 3 examines prior surveys on LLM reasoning and hypothesis discovery, highlighting their narrow emphasis on deductive tasks or application-specific methods. Section 4 reviews methods for forming hypotheses *(Abduction)*. Section 5 then covers approaches for applying these hypotheses *(Deduction)*, and Section 6 focuses on techniques for validating given hypotheses with new observations *(Induction)*. Finally, Section 7 explores the entire hypothesis-discovery cycle by examining the interdependencies among these reasoning steps and

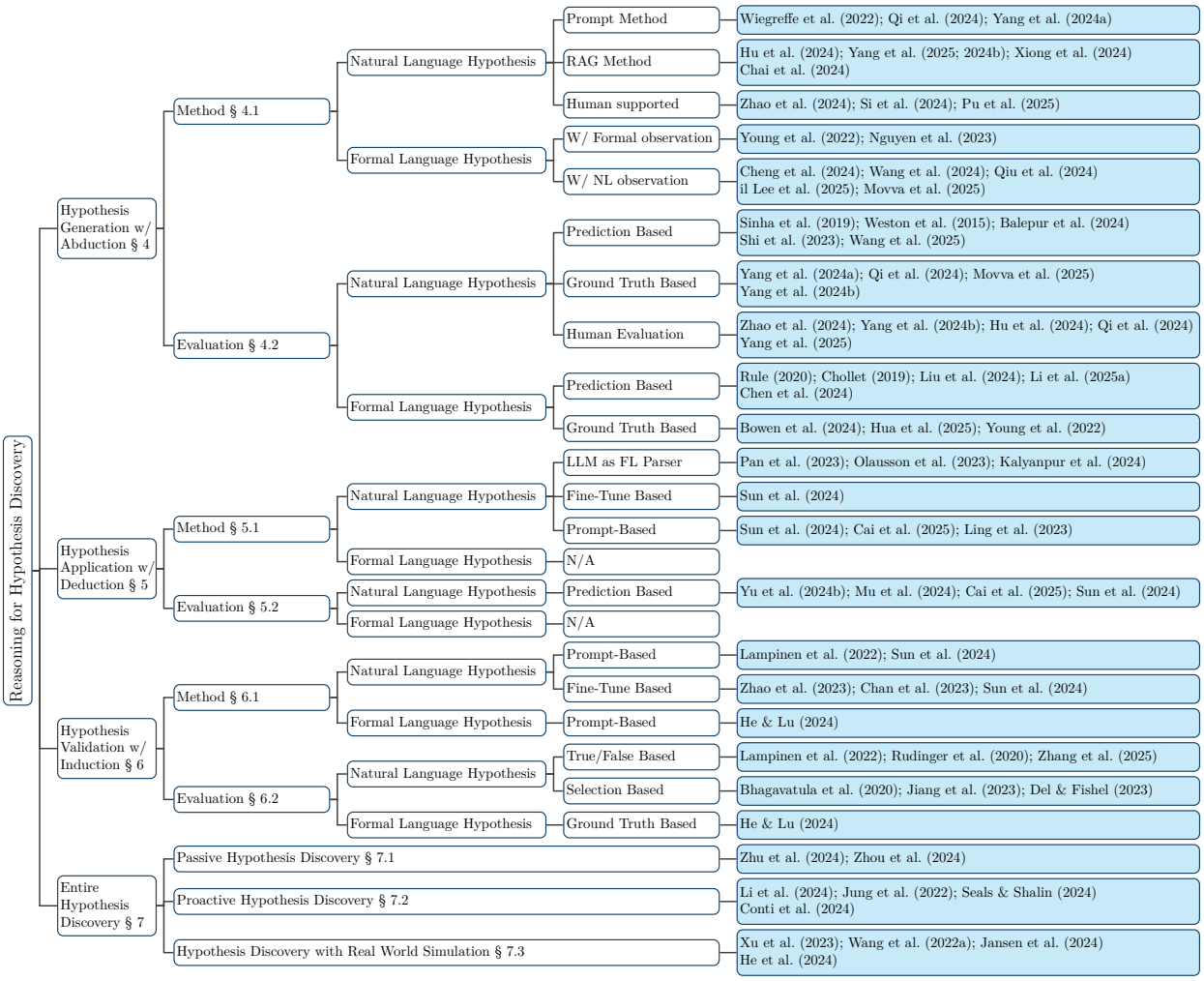

Figure 1: Taxonomy for Hypothesis Discovery with LLMs. Our survey categorizes work into four topics based on Peirce's definition of hypothesis discovery: Generation (creating hypotheses that explain given observations with abduction), Application (deducing new observations from established hypotheses with deduction), Validation (verifying and refining hypotheses against new evidence with induction), and Integrated Hypothesis Discovery (examining the dynamic interdependencies among these components in a continuous, iterative process).

showing how abduction, deduction, and induction can be iteratively used to refine more robust hypotheses. For each stage, we discuss methods, benchmarks, evaluations, and identify limitations and future directions. A high-level taxonomy guiding this survey is shown in Figure 1.

2 Background

Before LLMs, most AI systems stored knowledge as handcrafted symbols and rules. That format works well for deduction, because most of the problems we need to solve with symbolic AI systems work with limited premises and countable task-specific knowledge; for example, questions in the ProofWriter (Tafjord et al., 2021) and FOLIO (Han et al., 2024) benchmarks are limited to fewer than a hundred premises. However, abduction and induction are different: they call for generating and validating many tentative explanations inspired by vast commonsense or expert domain knowledge (such as weather patterns, social norms, or physics) and for updating beliefs as new observations arrive. Handling these reasoning tasks with symbolic

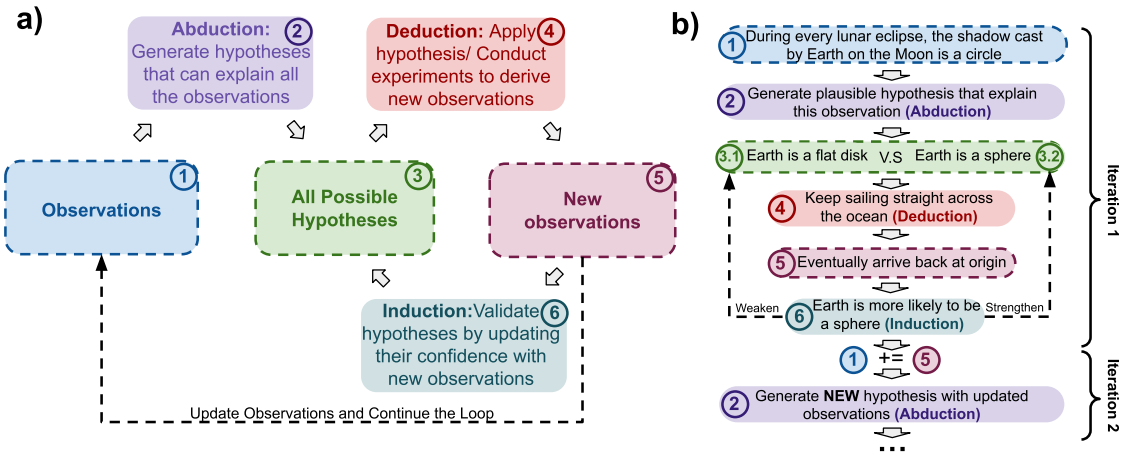

Figure 2: On the left-hand side, **a)** illustrates Peirce's framework for hypothesis discovery through abduction, deduction, and induction. The process begins with abduction, which generates explanatory hypotheses based on an initial set of observations. Deduction is then used to apply these hypotheses and derive predictions. Induction evaluates how well the predicted observations align with actual outcomes, updating the confidence of the hypotheses or rejecting those that are no longer valid. This process is iterative: validated hypotheses may be refined through further rounds of abduction using updated observations, gradually leading to more robust theories. On the right-hand side, **b)** provides a simple example that illustrates this process.

AI meant writing and maintaining vast, interlocking rule bases, an effort so costly that few projects moved beyond toy domains (Yang et al., 2024c). Consequently, the research landscape remained dominated by deductive tasks (Yu et al., 2024a; Liu et al., 2025; Huang & Chang, 2023).

LLMs have transformed this landscape. Trained on vast corpora, they implicitly absorb broad commonsense and domain knowledge, exhibiting strong reasoning abilities on complex, natural-language tasks (Yang et al., 2024c). With a simple text prompt, we can now ask them to carry out abduction or induction and even inspect their intermediate reasoning steps (Li et al., 2024; Jung et al., 2022), exposing the latent information they rely on. This advancement has made it practical to study and deploy **defeasible reasoning** (Yang et al., 2024c). Defeasible reasoning refers to forms of reasoning, such as abduction and induction, that yield probable conclusions that remain open to revision as new evidence emerges. This shift has fueled a wave of NLP research that places such flexible reasoning at the heart of AI progress (Liu et al., 2025; Huang & Chang, 2023).

2.1 Hypothesis Discovery

Hypothesis discovery or Rule learning, the cyclical process of formulating hypotheses, gathering evidence, validating or refuting them, and ultimately establishing robust theories, lies at the heart of scientific progress (Eger et al., 2025). Early humans, for example, hypothesized that the Earth was flat based on everyday observations. Later, Eratosthenes measured shadow angles at different locations, obtaining evidence that suggested the Earth's surface was curved. This evidence challenged the flat Earth hypothesis, and subsequent findings, notably Magellan's circumnavigation, conclusively confirmed the Earth's roundness. Even with today's sophisticated instruments, researchers continue to iterate this loop in new domains, validating and refining theories as new data emerges. Today, there is growing interest in whether LLMs can autonomously generate, apply, and validate hypotheses from natural language represented observations, mirroring this iterative process to achieve interpretable and adaptive hypothesis discovery. Although many studies have explored individual steps of hypothesis discovery, their efforts tend to be scattered across abduction, deduction, and induction, with insufficient attention to how these forms of reasoning interconnect to drive genuinely iterative, hypothesis-driven discovery.

2.2 Reasoning

Reasoning is central to hypothesis discovery. Researchers have historically debated how best to categorize reasoning into clear, operational types. Different frameworks each have strengths and limitations (Harman, 1965; Douven, 2021; Bacon, 1878; Laudan, 1971; Mill, 2024; Stadler, 2011; Popper, 2005; Okoli, 2023). In this survey, we adopt Charles Peirce's definition of reasoning (Peirce, 1974; Burks, 1946; Minnameier, 2004), emphasizing abduction, deduction, and induction as separate but interrelated processes. We choose Peirce's framework for three main reasons. First, clarity: Unlike many other approaches, Peirce explicitly differentiates among the three reasoning types, preventing confusion, such as the common conflation of abduction and induction. Second, practicality: Peirce's model aligns each form of reasoning directly with a distinct phase in the hypothesis discovery cycle—abduction for generating hypotheses, deduction for applying these hypotheses, and induction for validating them. This clear mapping makes his framework particularly suitable for systematically studying the entire process of hypothesis discovery, rather than isolated reasoning components. Finally, granularity: Peirce's framework breaks down the scientific discovery process into well-defined, finer-grained steps, facilitating detailed analysis and enabling more structured evaluation.

Abductive Reasoning is the process of forming explanatory hypotheses to make sense of observed phenomena. It is the only form of reasoning that generates entirely new ideas or explanations (Peirce, 1974; Frankfurt, 1958). Given a set of observations, one uses creative thinking and recalls necessary knowledge to come up with hypotheses that plausibly explain these observations. Importantly, a single set of observations can lead to multiple possible explanations. For instance, if you come home and find the floor wet, you might form several possible explanations: perhaps a pipe leaked, or someone spilled water accidentally. Without additional evidence or testing, you can't know for sure which explanation is correct. This illustrates how abduction helps generate potential explanations, which then must be tested further.

Inductive Reasoning is the process of testing whether the hypothesis and deduced consequences really obtain and evaluating to what extent they obtain (Minnameier, 2004; Peirce, 1974). In practice, induction updates a hypothesis's confidence based on new observations, including rejecting it outright, or selects the most convincing candidate from a set of competing hypotheses. Consider the claim "Swans are *(100%)* white," or linguistically, "All swans are white," formed after observing 99 white swans in Texas. Encountering a black swan in New York contradicts that hypothesis. Through induction, we recognize this contradiction and lower our confidence in the original claim, adjusting it to "Swans are *(99%)* white," linguistically expressed as "Almost all swans are white." In this example, although the hypothesis appears "revised," the change is limited to its confidence level; no new explanatory perspective is introduced, and we do not actually form a new hypothesis. By contrast, abduction can lead us to a fresh explanatory hypothesis with new observations, e.g., "Swans' color depends on their habitat," or "All swans in Texas are white," which introduces new ideas and is not a case of induction. Thus, inductive reasoning verifies or refines existing hypotheses (in terms of confidence) based on accumulating evidence.

Deductive Reasoning is the process of logically deriving specific conclusions from general hypotheses or rules. If the initial hypotheses are true, deduction guarantees that the derived conclusions must also be true. For instance, from the general rule "All swans are white" and the observation "This bird is a swan," we logically conclude "This bird must be white". While traditional deductive reasoning tasks, such as instruction-following and standard problem-solving, have been extensively studied with LLMs (Pan et al., 2023; Wei et al., 2022; Liu et al., 2025; Huang & Chang, 2023), deductive reasoning in the context of hypothesis discovery poses unique challenges. Specifically, it emphasizes inferential rule-following, requiring models to consistently apply hypotheses or rules to derive new and potentially unfamiliar conclusions, even when these hypotheses are counterfactual, unfamiliar, or incorrect. For example, when a flawed hypothesis is introduced in an unfamiliar domain, inferential rule-following requires us to strictly derive its predicted consequence, even if that consequence itself is incorrect. By comparing this consequence with experimental data, we can directly assess the hypothesis's validity and guide its revision. Conversely, if the deductive process is unreliable, we may overlook real contradictions and thus retain invalid hypotheses or discard valid ones. Indeed, recent work shows that although LLMs can demonstrate strong deductive performance on in-distribution tasks, they rely heavily on surface-level pattern matching and fail to generalize their inferential rule-following to novel or counterfactual scenarios (Pu et al., 2025; Mirzadeh et al., 2024; Kang et al., 2024; Yan et al., 2025).

There are also other types of reasoning, such as analogical reasoning (Yuan et al., 2023; Jiayang et al., 2023). However, their function in hypothesis discovery is generally covered by abduction and induction. We will include these additional forms when we encounter a relevant case in the following section.

2.3 Rule Representation: Formal Language vs Natural Language

Table 1: Comparison of Natural vs. Formal Language Representations for the Hypothesis "Sam is a dragon". In natural language, commonsense knowledge is implicitly embedded, and derived knowledge relies on extensive commonsense, potentially resulting in different interpretations depending on background knowledge and context. In formal languages (e.g., FOL or code), the knowledge base must be defined explicitly and cannot fully capture all commonsense knowledge, however, the derived conclusions are deterministic and precise.

Representation	**Hypothesis**	**Knowledge Base**	**Derived Knowledge**
Natural Language	English: 'Sam is a dragon"	English commonsense of 'dragon"	Sam is dangerous
	Chinese: 'Sam is a dragon"	Chinese commonsense of 'dragon"	Sam brings good fortune and a bountiful harvest
Formal Language	FOL: Dragon(Sam)	$\forall x(\text{Dragon}(x) \rightarrow \text{Fly}(x))$ ...	Fly(Sam)

	Code: `Sam = Dragon()`	`Class Dragon:`
		`  def fly(self):`
		`  ...`
			`Sam.fly()`

There are many ways to represent hypotheses and rules, which we broadly divide into two categories: **formal languages (FL)** and **natural languages (NL)**. Formal languages, such as first-order logic and programming languages, are systematic and rule-bound. After real-world entities are encoded as explicit literals, precise inference rules yield provably correct and sound conclusions, making these systems well suited to **deductive** reasoning. Yet the encoding process strips away many subtle semantic relationships and commonsense knowledge, limiting the system's ability to handle the creative, defeasible reasoning required for **abduction** and **induction** (McCarthy & Hayes, 1981; Reiter, 1980; Hanks & McDermott, 1987; Liu et al., 2025; Yu et al., 2024a; Huang & Chang, 2023). Natural language preserves those nuances and aligns more closely with human cognition, so it is better suited to **abductive** and **inductive** tasks. However, its meanings are implicit and context-dependent, making it difficult to define a deterministic reasoning pipeline and reducing the reliability of the resulting inferences (See in Table 1). Accordingly, the following sections treat formal-language and natural-language approaches separately, emphasizing how their reasoning methods and evaluation protocols differ.

3 Related Surveys

Most existing work assessing LLM reasoning, both survey syntheses and popular benchmarks such as GSM8K(Cobbe et al., 2021), centres almost exclusively on multi-step deductive tasks, leaving abduction and induction, the engines of hypothesis discovery, largely unexplored. Surveys of the field Yu et al. (2024a); Liu et al. (2025); Huang & Chang (2023) highlight the absence of systematic study and clear analytical frameworks for these modes, while benchmark analyses likewise show that abductive and inductive inference receive limited attention (Plaat et al., 2024; Li et al., 2025b). This imbalance obscures our understanding of whether, and to what extent, LLMs can perform the creative, evidence-based reasoning required for hypothesis-driven discovery.

On the other hand, research in the AI for Science domain takes a distinctly **horizontal, application-driven** approach. This body of work emphasizes practical tasks such as generating research ideas, conducting experiments, and synthesizing reports, often employing domain-specific pipelines tailored to individual scientific fields. However, these studies usually lack a generalizable reasoning framework applicable across different scientific contexts. Furthermore, their evaluation metrics, typically novelty, creativity, or consistency, tend to be subjective, human-centric, and thus difficult to generalize, offering limited theoretical insight into the

underlying reasoning mechanisms involved in scientific discovery (Movva et al., 2025; Alkan et al., 2025; Reddy & Shojaee, 2025; Gridach et al., 2025; Bazgir et al., 2025).

Our survey adopts a **vertical, reasoning-centered perspective** grounded in Peirce's classical framework. It integrates three modes of reasoning into a unified view of hypothesis discovery: **abduction** for hypothesis generation, **deduction** for hypothesis application, and **induction** for hypothesis validation. Unlike prior surveys that emphasize primarily deductive tasks, we concentrate on the entire reasoning process involved in hypothesis discovery, explicitly covering both defeasible reasoning (abduction and induction) and deductive reasoning. By clearly defining each reasoning mode and explaining its role within each stage of the discovery process, we provide a structured basis for designing principled, model-agnostic benchmarks and evaluation tasks. Compared to existing application-oriented surveys, our framework thus offers a more abstract, systematic, and theoretically informed approach to understanding and enhancing the role of LLMs in automated scientific discovery.

4 Hypothesis Generation

Every scientific discovery begins with a set of observations, denoted as $\mathbb{O} = \{o_1, o_2, \ldots, o_n\}$, that we aim to explain. Let h represent the generated explanation or hypothesis. The hypothesis generation task can be defined as generating an h such that:

$$h \models (o_1 \wedge o_2 \wedge \cdots \wedge o_n)$$

This notation means that h logically entails the observations. In other words, assuming h holds, it guarantees that all observations $o_1 \wedge o_2 \wedge \cdots \wedge o_n$ follow. In this survey, we follow Peirce's definitions for reasoning. Accordingly, the primary process used in hypothesis generation is abduction, the method of formulating explanatory hypotheses to account for observed phenomena.

4.1 Method

Despite LLMs' demonstrated prowess in tasks like summarization or code generation, devising robust methods to guide them in hypothesis generation remains an active area of research. Recent work has sought to leverage LLMs' in-context learning and natural language understanding to produce novel or domain-specific hypotheses, spurring the development of new techniques aimed at improving both the quality and applicability of generated hypotheses (Yang et al., 2024c). In this section, we review these methods, spanning approaches that rely solely on prompting, those that integrate external knowledge sources, and those that incorporate human expertise in the loop.

4.1.1 Natural Language Hypothesis Generation with LLMs

Prompt-Based Methods: Due to the lack of large-scale, domain-specific data for hypothesis generation, most abduction approaches rely on prompt-based methods that are easy to deploy and don't require extensive additional data. For instance, when provided with observations expressed in natural language and asked to generate a plausible hypothesis that explains them, both Wiegreffe et al. (2022) and Qi et al. (2024) employ few-shot prompting to guide LLMs in generating hypotheses. Specifically, Wiegreffe et al. (2022) constructs few-shot examples using a triplet format *(question, answer, explanation)*. In solving a task of generating biomedical hypotheses with given observations, Qi et al. (2024) embeds a small set of independent observation-to-hypothesis pairs in the prompt. By showing how each block of biomedical background observations maps to its corresponding hypothesis, the model learns to extract relevant domain cues and generate novel biomedical hypotheses. Their findings indicate that including more examples in the prompt tends to reduce the novelty of the generated hypotheses while increasing their correctness. Furthermore, Yang et al. (2024a) propose a pipeline for hypothesis generation that involves five prompt-based modules: one to generate hypotheses, one to test deductive consistency, one to verify that the hypothesis is not merely a copy of the given context, one to assess its generalizability, and one to determine whether the hypothesis is trivial.

RAG-Based Methods: Labeling massive corpora for pre-training is costly, but assembling a small or medium dataset for Retrieval-Augmented Generation (RAG) is practical, and several studies follow a similar

iterative three-step pattern: (i) retrieve task-specific documents, (ii) let an LLM generate or refine hypotheses, and (iii) iterate with LLM feedback. For instance, after a user supplies a seed paper and asks the LLM to generate a worthwhile hypothesis to pursue in research, Hu et al. (2024) query the Scholar API for related work, then repeatedly generate and critique hypotheses, gradually expanding a web of novel ideas. Yang et al. (2025) apply the same loop to 51 top-tier chemistry papers from 2024: experts first segment each paper into background, inspiration, and hypothesis; an LLM-based multi-agent system (MOOSE-Chem) then retrieves relevant snippets, drafts hypotheses, and scores them for originality. A similar pipeline appears in Yang et al. (2024b), where 50 conference papers are annotated in the same three fields, augmented with thematically similar web documents and 14 survey papers so that the LLM can judge both relevance and novelty.

Two variants enrich the retrieval step with structured or fine-tuned knowledge. Xiong et al. (2024) ground each hypothesis in a domain knowledge graph: entities mentioned during generation are checked against graph relations, ensuring the final claims remain fact-consistent. In contrast, Chai et al. (2024) fine-tune a T5 model (Raffel et al., 2020) on curated scientific abstracts and, during inference, retrieve citation contexts and related data; a novelty-guided loop then re-generates until the candidate is both coherent and inventive, outperforming standard transformer baselines.

Human-in-the-loop Hypothesis Generation with LLM: Recent studies show that combining humans with LLM support yields higher-quality, more novel hypotheses than either party working alone. Zhao et al. (2024) report that a human+LLM pipeline surpasses both human-only and LLM-only baselines. Human annotators first draft hypotheses for uncommon observations; carefully designed prompts then guide the LLM to refine each draft by adding details and improving logical flow. Low-quality hypotheses generated by LLM are filtered with human evaluations and automated metrics such as BERTScore, and the resulting human+LLM collaboration produces the strongest hypotheses. Similarly, Si et al. (2024) involve more than 100 NLP researchers in a three-condition study: LLM-only generation, human-only generation, and LLM generation reranked by humans. Human evaluations rate the human-reranked LLM hypotheses best on Novelty, Excitement, Feasibility, and Effectiveness. Pu et al. (2025) move beyond controlled experiments by introducing IdeaSynth, a copilot-like framework that assists users throughout hypothesis formulation. When a user supplies a high-level hypothesis, IdeaSynth retrieves relevant papers through an API and summarizes key information needed for development. Users interactively edit these summaries with LLM help, adding details and improving clarity. The system then aggregates all refined nodes, employs an LLM to craft a final applicable hypothesis, and supplies suggestions for follow-up experiments.

The quality of natural-language hypothesis generation largely depends on the inherent capabilities of LLMs. Because these models excel at in-context learning, prompt strategies such as Chain-of-Thought (CoT) and Reflexion (Wei et al., 2022; Shinn et al., 2023) can be applied directly to this task. However, unlike computer-vision research, which gained rapid momentum from the ImageNet benchmark, hypothesis generation lacks a comparable, widely recognized task set. The main challenge is therefore the absence of a reliable evaluation task and benchmark for natural-language hypotheses, an issue examined further in Section 4.2.

4.1.2 Formal Language Hypothesis Generation with LLM

One major advantage of formal hypotheses is that once a formal language hypothesis is obtained, we can directly perform inference on it with guarantees of soundness and correctness. Depending on whether observations are represented in formal or natural language, methods for proposing a formal language hypothesis need to be discussed separately.

Formal Language Observations: When observations are encoded in a formal language, dedicated formal language solvers typically yield clear, white-box solutions that outperform language models. Consequently, using an LLM for these tasks is generally not preferred. Nevertheless, a few early studies in the LLM era have explored this approach. For example, Young et al. (2022) trained a transformer model on FOL abduction tasks, demonstrating that the model can generate FOL hypotheses from formal observations. Similarly, Nguyen et al. (2023) fine-tuned state-of-the-art legal transformers on FOL abduction tasks and found that models pre-trained on natural language legal abduction tasks do not show any performance improvements on FOL hypothesis generation problems.

Natural Language Observations: When observations are represented in natural language, traditional symbolic solvers struggle to extract the key information needed for hypothesis generation. With LLMs, however, we can directly generate formal hypotheses. A popular formal language for this purpose is code, as it offers greater flexibility than other symbolic representations like FOL, and LLMs excel at coding.

The simplest variant prompts an LLM with an observation set and asks it to produce executable functions as hypotheses that match the input-output pairs; Cheng et al. (2024) follow this pattern, treating each observation as an *(x, y)* example and evaluating the generated function by execution. Extending this idea, Wang et al. (2024); Qiu et al. (2024) have the LLM create multiple executable hypotheses, run them on the observations, feed the results back to the model, and iterate, discarding weak candidates and refining promising ones until one covers all examples. To encourage diversity, il Lee et al. (2025) first ask the model for a single-word "main concept," then use that concept to steer subsequent code generation, avoiding the similarity of low-temperature outputs and the degeneration of high-temperature sampling while still producing coherent hypotheses.

A complementary line of work probes the model's internal representations. Using sparse autoencoders (SAE) (Bricken et al., 2023), Movva et al. (2025) isolate neurons activated when the LLM predicts the click rate of Twitter posts and discover that neurons associated with "surprise" or "shock" positively influence the score, supporting the hypothesis that surprising or shocking content tends to receive more clicks.

4.2 Evaluation for Hypothesis Generation

Due to LLMs' strong reasoning abilities and natural language interface, many methods have been proposed for hypothesis generation, and numerous ideas based on everyday human reasoning can be adapted for this purpose (Niu et al., 2024). However, a major challenge remains in establishing a grounded and convincing way to evaluate the quality of the generated hypotheses.

4.2.1 Natural Language Hypothesis Evaluation

Although prompting LLMs to generate natural language hypotheses is straightforward, evaluating the quality of these hypotheses is challenging due to the ambiguity inherent in natural language representations. Consequently, a common evaluation method involves either human evaluation or using an LLM to assess the generated hypotheses' validity (Zhao et al., 2024; Yang et al., 2024b; Hu et al., 2024; Qi et al., 2024; Yang et al., 2025). While human evaluation can provide valuable insights without relying on predefined answers, it is inherently subjective, less reproducible, expensive, and sometimes not entirely convincing. Therefore, alternative evaluation strategies are needed.

Implicit Prediction-based Evaluation: Early benchmarks often relied on question-answering (QA) tasks that required the model to implicitly form a hypothesis to answer a question (Sinha et al., 2019; Weston et al., 2015). For example, consider the observation: *"Lily is a swan, Lily is white, Bernhard is green, Gerg is a swan. What color is Greg?"* To answer correctly, one must infer an implicit hypothesis, such as "All swans are white" or "Most swans are white," based on the fact that Lily is both a swan and white. Thus, the correct answer is "white." By verifying whether the model's answer is "white," one can indirectly assess its ability to form an appropriate hypothesis and perform reasoning. Similarly, recent work shows that prompting LLMs to generate an intermediate hypothesis and then using that hypothesis for inference yields higher performance on complex tasks (Balepur et al., 2024; Shi et al., 2023; Wang et al., 2025). However, this approach is problematic: the hypothesis may be formed incorrectly, the subsequent inference could be flawed, and the model might arrive at the correct answer through memorization or random guessing rather than proper abductive reasoning. Therefore, success in these tasks does not directly imply that the model possesses superior abductive capabilities, making them unsuitable for reliably evaluating hypothesis generation.

Ground Truth-based Evaluation: Some studies build benchmarks with labeled hypotheses so that outputs of LLM can be matched directly against references. DEER (Yang et al., 2024a) supplies 1,200 fact–rule pairs, all written in natural language by experts across six topics-zoology, botany, geology, astronomy, history, and physics. Generated hypotheses are compared with the gold rules using token-level mapping metrics

like METEOR (Banerjee & Lavie, 2005). In biomedicine, Qi et al. (2024) curate a benchmark with both seen and unseen samples: the seen split contains 2,700 background–hypothesis pairs collected before January 2023, whereas the unseen split has 200 pairs collected after that date. Outputs are evaluated against the ground truth with BLEU and ROUGE (Papineni et al., 2002; Lin, 2004). On synthetic corpora such as WIKI and BILLS (Pham et al., 2024; Zhong et al., 2024), Movva et al. (2025) treat hypothesis generation as identifying the key features that drive a prediction. The model proposes feature sets, which are judged by how well they match and cover the ground-truth features, thereby quantifying the LLM's ability to isolate causal signals.

Despite these efforts, Yang et al. (2024b) note that reference-based metrics such as BLEU, ROUGE, and METEOR assume a single correct answer and therefore struggle to capture the open-ended nature of hypothesis generation; developing fair, reliable metrics remains an open challenge.

4.2.2 Formal Language Hypothesis Evaluation

Unlike natural language hypotheses, formal hypotheses evaluations are more grounded due to their clarity and unambiguous semantics.

Ground Truth-based Evaluation: Generated formal hypotheses can be evaluated against pre-defined ground truth hypotheses. Unlike natural language evaluation, where ground truth is often written by domain experts and evaluated using token-level metrics like BLEU or ROUGE, formal hypotheses can be evaluated procedurally using solvers. This allows us to verify correctness deterministically. For example, Bowen et al. (2024) designed formal representations for synthetic grouping tasks to evaluate formal language hypothesis generation. Hua et al. (2025) constructed their benchmark based on deterministic regular functions, providing a procedural framework for evaluating formal hypotheses. Similarly, Young et al. (2022) used first-order logic (FOL) representations, where LLMs were tasked with generating FOL hypotheses to explain given facts, and the outputs were evaluated by comparing them against ground truth hypotheses verified by solvers.

Prediction-based Evaluation: Since inference on formal hypotheses is deterministic, a common evaluation method is to test whether the generated hypothesis produces correct outcomes on held-out examples. For instance, Rule (2020) propose the *list function task*, where LLMs generate a hypothesis function from observed (x, y) pairs, and evaluation is based on how well the hypothesis predicts hidden pairs. Similarly, Chollet (2019) introduces the Abstract Reasoning Corpus (ARC), where tasks involve transforming input grids of colored cells into output grids. The generated function is executed on test inputs, and correctness is determined by exact matches with the target output grids, including grid dimensions. Liu et al. (2024) further propose a benchmark consisting of arithmetic calculations, color token mapping, and Kalamang vocabulary tasks, all evaluated in the same way. Additionally, Li et al. (2025a) construct diverse application scenarios, including list transformations, real-world problems, code generation, and string transformations, where the generated hypothesis is executed on both seen and test observations and the final score aggregates performance across both sets. In a more realistic setting, Chen et al. (2024) extract 102 tasks from 44 peer-reviewed publications, unifying the target output for every task into a self-contained Python program file, accompanied by a set of test cases validated by human experts. LLMs are then asked to read the paper and reproduce the tasks in code, and the generated code is directly evaluated on the prepared test cases.

4.3 Discussion and Future Directions in Hypothesis Generation

There exists a significant gap between formal and natural language approaches to hypothesis generation. In natural language hypothesis generation, observations typically originate from recent research papers, and generated hypotheses can potentially inspire novel research ideas with tangible real-world impacts (Eger et al., 2025). However, rigorous and reliable evaluation methods for such hypotheses remain underdeveloped. Token-based metrics, such as BLEU or ROUGE, do not effectively capture the qualitative aspects of open-ended hypothesis generation (Yang et al., 2024b). Meanwhile, alternative approaches involving human or LLM-based evaluations are costly, subjective, and prone to inconsistencies.

Conversely, formal language hypothesis generation benefits from grounded, objective evaluation methods. Nevertheless, existing formal tasks often involve simplified or artificial scenarios that fail to reflect the com-

plexity and nuance inherent in real-world applications. Consequently, the field faces a trade-off: formal representations facilitate robust evaluation but risk omitting critical real-world nuances, while natural language representations capture real-world complexity yet lack rigorous evaluation mechanisms.

To address this challenge, future research in hypothesis generation could focus on two key directions. Firstly, there is an urgent need to develop novel evaluation methodologies tailored specifically for natural language hypothesis generation. Current implicit prediction-based evaluations suffer from inherent limitations, and ground truth-based evaluations remain inadequate due to reliance on token-level similarity metrics. Alternative evaluation strategies, potentially involving multi-dimensional human assessments, structured feedback mechanisms, or hybrid evaluation frameworks integrating automated and expert evaluations, merit exploration. Secondly, bridging the gap between formal and natural language hypothesis generation is crucial. Leveraging code as an intermediate representation offers a promising path forward, combining evaluative rigor with expressive capability. However, existing code-based hypothesis generation benchmarks tend to focus on oversimplified problems that lack relevance to practical scenarios. Thus, developing realistic, code-based hypothesis-generation tasks grounded in established research papers, real-world datasets, and open-source repositories presents a compelling and valuable direction for future research (Chen et al., 2024).

5 Hypothesis Application

Given a hypothesis h, hypothesis application is defined as the derivation of a new observation o_{new} such that:

$$h \models o_{\text{new}}$$

In some cases, the hypothesis may depend on a context c, so that h can be viewed as a function of c. In this context-dependent formulation, hypothesis application is defined as deriving a new observation o_{new} such that:

$$h(c) = o_{\text{new}}$$

In our work, we follow Peirce's definitions for reasoning. Accordingly, the primary process used in hypothesis application is deduction, the method of deriving necessary consequences from a given hypothesis.

Notably, when a hypothesis is expressed in a formal language, directly applying it with a deterministic solver yields a correct and sound prediction. Therefore, there is little motivation to leverage LLMs for deductive reasoning on formal hypotheses. This section, consequently, focuses on the natural language hypothesis application and evaluation.

5.1 Method

LLM as Formal Language Parser: Since formal symbolic solvers yield sound and correct predictions, Pan et al. (2023); Olausson et al. (2023); Kalyanpur et al. (2024) treat LLMs as formal language parsers, using them to translate natural language hypotheses into formal representations like FOL and code before applying a formal inference procedure. This translation significantly improves deductive correctness. However, these methods have primarily been evaluated on benchmarks such as ProofWriter (Tafjord et al., 2021) and FOLIO (Han et al., 2024), where the questions are already closely aligned with formal language. For example, given the input *"Fact1: Eric is young, Fact2: Dave is white, Rule 10: if someone is young and not kind then they are big"*, translating this into FOL is relatively straightforward. It remains unclear whether LLMs can reliably parse more complex, everyday natural language into formal representations.

Fine-Tuning-Based Method: Fine-tuning is a common approach to improve model performance when corresponding training data is available. Sun et al. (2024) proposed a synthetic "StringGame" task in which ground truth hypotheses and answers are provided. Leveraging the CoT approach, a LLM is prompted to generate multiple candidate hypothesis application trajectories along with their results. By comparing these results with the ground truth, the trajectories that produce correct outcomes are identified as correct and stored for fine-tuning. The resulting fine-tuned model then demonstrates improved performance in both hypothesis application and instruction following.

Prompt-Based Method: Although CoT prompting has improved performance on multi-hop question answering tasks, Sun et al. (2024) found that it does not directly enhance performance in hypothesis application. Therefore, new prompting methods have been designed specifically for this purpose. Inspired by mathematical induction, Cai et al. (2025) propose quantifying the difficulty of a question so that the LLM can solve it incrementally, from simpler versions to more complex ones, ultimately arriving at the correct answer. In another approach, Ling et al. (2023) design a pipeline that supervise the correctness of each reasoning step during hypothesis application. First, the LLM indexes all premises; then it is asked to label the minimal set of premises required to derive new facts. This pipeline generates multiple candidate hypothesis application trajectories, and by having the LLM vote on each step, the most convincing deductive trajectory is selected.

5.2 Evaluation for Hypothesis Application

Although many benchmarks and evaluation methods exist for general deductive reasoning, such as question answering and mathematical tasks like GSM-8k (Cobbe et al., 2021), these question types do not explicitly test the formation of new facts based on given hypotheses or rules. Evaluating the correctness of a natural-language deductive trajectory is challenging because annotated reasoning paths for hypothesis application are scarce, and the same result can follow from different reasoning paths. As a result, most evaluations use prediction-based checks. We assume that, given a correct hypothesis and a known ground-truth result, a valid deduction will reproduce that result. By comparing the model's deduced outcome with the ground truth, we can judge whether its deduction is correct. For example, take the hypothesis *"Coin flips are independent and identically distributed (i.i.d.) with a 50 percent chance of heads."* When asked, *"After three consecutive heads, what is the probability of a tail on the fourth flip?,"* a flawed model might claim the chance of a tail has increased. In fact, under the i.i.d. assumption, the probability remains 50 percent. Supplying the correct hypothesis and comparing the model's answer to the true result lets us evaluate whether its deductive reasoning is valid.

Building on this idea, Yu et al. (2024b) create the TURTLEBENCH benchmark, inspired by the "Turtle Soup" game in which players deduce a story's hidden explanation by asking yes/no questions; in TURTLEBENCH, the LLM instead answers human-annotated questions with "True," "False," or "Not Relevant," across 1,532 high-quality question-answer pairs sourced from an online platform to test whether it can fully follow a story and provide accurate answers. Similarly, Mu et al. (2024) introduced the RULES benchmark, comprising 14 rule-following scenarios, each paired with concise test cases and programmatic evaluation functions that objectively assess adherence to specified rules. In addition, Cai et al. (2025) presented the Holiday Puzzle benchmark, which features multiple holiday schedule scenarios ranging from simple single-week planning to multi-phase arrangements and complex date arithmetic tasks, again using test cases and evaluation functions to verify correct computation of extra holiday rest days under provided rules. Moreover, Sun et al. (2024) constructed RuleBench to evaluate not only whether models can produce correct answers based on factual rules but also their ability to apply counterfactual rules, designed to yield incorrect outcomes, and experiments show that while LLMs achieve near-perfect accuracy on factual rules, their performance drops dramatically under counterfactual rules, revealing a significant gap in counterfactual rule-following capability.

5.3 Discussion and Future Directions in Hypothesis Application

While traditional deductive reasoning tasks (e.g., question answering, problem solving) in LLMs have been widely studied, the capability for hypothesis application remains significantly underexplored. According to Sun et al. (2024), hypothesis application involves inferential rule-following, requiring models to consistently apply given hypotheses to derive novel knowledge in unfamiliar domains. Robust hypothesis application is critical to hypothesis discovery, as hypotheses must generalize to scenarios with unseen observations. However, existing LLMs frequently struggle to extend hypotheses beyond familiar contexts, thus limiting the evaluation of hypothesis generation.

Future research could therefore focus on rigorously evaluating LLMs' hypothesis application, both factual and counterfactual, in novel scenarios. Developing benchmarks explicitly designed for hypothesis-driven

inference in unfamiliar domains could reveal important insights into model adaptability and generalization. Additionally, current evaluations of hypothesis application mainly rely on outcome-based correctness, comparing predicted results to ground truth given correct hypotheses. However, incorrect reasoning may still lead to correct predictions in natural-language contexts. Although Ling et al. (2023) propose improving hypothesis application by intervening in reasoning trajectories, a large-scale benchmark specifically designed to evaluate trajectory-based hypothesis application remains absent.

6 Hypothesis Validation

According to Peirce, induction validates a hypothesis by updating its confidence when new evidence appears. However, in studies that focus exclusively on induction, tasks are typically one-off: a hypothesis (or set of hypotheses) and a collection of observations are provided, and there is no iterative updating of confidence. A simplified framework for hypothesis validation treats it as a multiple-choice problem: given observations $\mathbb{O} = \{o_1, o_2, \ldots, o_n\}$ and a set of hypothesis $\mathbb{H} = \{h_1, h_2, \ldots, h_m\}$, the model selects the most possible hypothesis. In simpler scenarios, where only one hypothesis is provided, the model determines whether the hypothesis correctly explains the observations. In the next section, when combined with deduction and abduction, induction can subsequently be used to iteratively update the confidence in the hypothesis.

Natural language representations add significant complexity to induction. In formal language settings, all necessary information is explicitly provided, and reasoning follows rigorous, well-defined steps. In contrast, validating a natural language hypothesis often requires commonsense knowledge and interpretation of nuanced language. For example, consider the observations "Neil wanted to see the mountains of Asia" and "Neil loved being so close to the mountains in Nepal," with candidate hypotheses "Neil booked a trip online" and "Neil took a trip to see the Rocky Mountains instead." Here, the nuanced meaning of the term "instead" and the geographic relationships require careful analysis and may lead to different conclusions. Indeed, Zhang et al. (2020) reports that, when verifying their dataset where five annotators judged the plausibility of handwritten hypotheses, disagreements occurred in 62.34% of 1,365 explanations, underscoring the challenge of natural language hypothesis validation.

6.1 Method

6.1.1 Formal Language Hypothesis Validation

He & Lu (2024) introduce the CauseJudger framework, which leverages LLMs at every stage to validate candidate hypotheses. First, an LLM transforms the natural language inputs into an FOL-based representation by integrating each candidate hypothesis into the premises. Next, an LLM filters out irrelevant premises and rules. Finally, another LLM performs forward reasoning to decide which hypothesis explains the observations.

6.1.2 Natural Language Hypothesis Validation

Prompt-Based Method: Lampinen et al. (2022); Sun et al. (2024) employ a few-shot prompting approach for hypothesis validation. In this method, case triplets, consisting of an observation, a hypothesis, and its corresponding validity, are provided to the model, which then answers a hypothesis validation question. Although this approach improves performance, Sun et al. (2024) reports that the performance boost is limited. Their experiments further indicate that fine-tuning outperforms few-shot prompting.

Fine-Tuning-Based Method: Since hypothesis validation essentially constitutes a classification problem, many Natural Language Inference (NLI) datasets can be adapted into hypothesis validation tasks. Consequently, fine-tuning is a popular method in this context. For example, Zhao et al. (2023); Chan et al. (2023); Sun et al. (2024) fine-tune models to select the correct hypothesis from a set of hypotheses based on new observations.

6.2 Evaluation for Hypothesis Validation

6.2.1 Formal Language Evaluation

Along with the CauseJudger framework, He & Lu (2024) also proposed the CauseLogics dataset. Based on the required formal reasoning depth, the dataset is divided into four difficulty levels for hypothesis validation tasks, with 50,000 samples per level. Each hypothesis is assigned a binary ground-truth label indicating whether it correctly explains the observations.

6.2.2 Natural language Evaluation

Binary-Classification-Based Evaluation: Lampinen et al. (2022) chose a subset of 40 tasks from the crowd-sourced benchmark BIG-bench (bench authors, 2023) and constructed their own benchmark specifically for hypothesis validation. Each data sample consists of an observation, its corresponding hypothesis, and a ground truth label indicating whether the hypothesis truly explains the observation.

Hypothesis validation using natural language is inherently challenging because the implicit information and required common-sense background are not explicitly stated. This often leads different individuals to draw different conclusions when validating a hypothesis based solely on recalled information. Rudinger et al. (2020) mitigate this issue by adopting a different strategy. Instead of asking annotators to directly judge whether an observation explains a hypothesis, they ask the model to determine if a given observation weakens or strengthens the hypothesis. Specifically, they sample observation–hypothesis pairs from existing datasets and then manually craft two types of sentences: one that acts as a "strengthener" (increasing the likelihood of the hypothesis) and one that acts as a "weakener" (decreasing the likelihood of the hypothesis). Their validation process showed that the strengthening and weakening effects are consistent across different annotators. During evaluation, the model is required to decide whether a new observation strengthens or weakens the hypothesis. This approach aligns with the paper's goal of modeling defeasible inference by leveraging explicit contextual updates rather than relying on potentially variable human interpretations of implicit information. Furthermore, Zhang et al. (2025) extended this task to include visual observations. In their extension, given a visual observation and a natural language hypothesis, an LLM is tasked to determine whether the provided sentence serves as a strengthener or a weakener.

Multiple-Choice-Based Evaluation

Bhagavatula et al. (2020) introduce the ART benchmark, comprising roughly 20k narrative contexts where each sample includes two time-ordered observations, one depicting a story's start (o_1) and the other its outcome (o_2), alongside two hypotheses: a plausible explanation (h^+) and a less plausible one (h^-), challenging models to choose the best explanatory hypothesis and enabling adaptation to hypothesis-generation tasks evaluated against ground-truth explanations. Similarly, Jiang et al. (2023) present the BRAINTEASER benchmark of about 1.1k lateral-thinking puzzles, each offering a question with multiple-choice answers, one that defies commonsense and several conventional distractors, in both sentence (narrative) and word (meaning-alteration) formats to test creative reasoning, with additional semantic and context reconstruction variants assessing reasoning consistency and robustness across formulations. Moreover, Del & Fishel (2023) introduced the True Detective benchmark for deep hypothesis validation, featuring 191 long-form detective puzzles ($\approx$ 1200 words each) from the "5 Minute Mystery" platform, where models (and humans) select the correct explanation from 4–5 options, human accuracy averages 47%, top solvers exceed 80%, and each puzzle includes golden chain-of-thought explanations detailing the reasoning steps that lead to the correct answer.

6.3 Discussion and Future Directions in Hypothesis Validation

Previous literature often conflates hypothesis generation and hypothesis validation, primarily due to ambiguity inherent in the IBE paradigm. Within IBE-based approaches, hypothesis validation typically appears as an implicit intermediate step, where selecting the "best" hypothesis is frequently based on unclear or subjective criteria without dedicated, independent evaluation. However, adopting Peirce's explicit distinction

between abduction, deduction, and induction clearly separates validation from generation, underscoring the need for dedicated research on validating hypotheses against newly observed evidence.

Current validation methodologies predominantly adopt end-to-end metrics that only assess final correctness, neglecting the reasoning processes and commonsense knowledge required to validate hypotheses in realistic settings. The subjective nature of natural language, coupled with different interpretations of observations, highlights the necessity for richer evaluative frameworks. Future benchmarks should incorporate detailed intermediate Chain-of-Thought data, capturing explicit reasoning steps humans take when validating hypotheses, such as recalling relevant commonsense knowledge and performing nuanced inference. Evaluations should then emphasize consistency between the reasoning process and available commonsense context rather than relying solely on superficial similarity to reference answers. Such benchmarks would greatly enhance our understanding of hypothesis validation and better reflect the complexities of human-like reasoning.

7 Hypothesis Discovery

Although many works introduced in the previous sections propose methods and evaluation metrics, they mainly focus on individual phases of **Hypothesis Discovery**—Hypothesis generation *(Abduction 4)*, Hypothesis application *(Deduction 5)*, and Hypothesis validation *(Induction 6)*. However, in real-life **Hypothesis Discovery**, these reasoning stages are not independent and must be treated holistically. Initially, we form hypotheses based on limited observations using abduction, which subsequently informs the application of these hypotheses through deduction, enabling the collection of further evidence. Concurrently, induction continuously evaluates and resolves inconsistencies arising between newly obtained observations and earlier hypotheses. This iterative interplay means that each hypothesis formulated, action taken, observation gathered, and inconsistency identified dynamically shapes and reshapes our evolving understanding, influencing subsequent reasoning steps and contributing to diverse interpretations of the world. Treating any single reasoning phase in isolation oversimplifies hypothesis discovery. For example, although Bowen et al. (2024) evaluated every reasoning, they handled each step separately and thus failed to assess the true rule-learning capability of LLMs. Consequently, integrating abduction, deduction, and induction into a unified learning loop remains both challenging and largely understudied, yet it is the ultimate goal for constructing end-to-end agents capable of scientific discovery.

Despite a few studies that acknowledge the interdependence among reasoning types and allow models to refine hypotheses iteratively, they still overlook two decisive aspects of real-world hypothesis discovery. First, most benchmarks remain static and passive: they hand agents a fixed set of observations deemed sufficient to reach the correct hypothesis, whereas real-life hypothesis discovery requires actively seeking additional evidence. Second, even in settings that allow proactive information gathering, the granularity of the action space is still too coarse: agents fetch observations via one-shot "recall" or "web-search" commands, whereas real scientists must strategically plan and carry out precisely staged experiments—often designing specialized equipment at each step. Recognizing these limitations, we categorize existing hypothesis-discovery research into three classes (see Fig. 3).

7.1 Passive Hypothesis Discovery

In this type of study, LLMs generate, apply, and validate hypotheses iteratively. However, the observations are provided by a fixed dataset. The LLM does not need to worry about which observations it will receive. Instead, it simply reasons based on the given data, passively receiving and processing the information provided.

Zhu et al. (2024) proposed the Hypotheses-to-Theories (HtT) Framework to generate formal hypotheses (e.g., *"if A then B"*) by leveraging existing benchmarks (Sinha et al., 2019; Wang et al., 2022b; Rule, 2020). In HtT, LLMs generate a hypothesis and propose learned rules to solve each question. When a new question is received, the model first formulates a preliminary hypothesis based on the context. It then proposes candidate rules that might lead to the correct answer. These candidate rules are applied to the problem and verified against the ground truth. Rules that consistently yield correct predictions are retained and

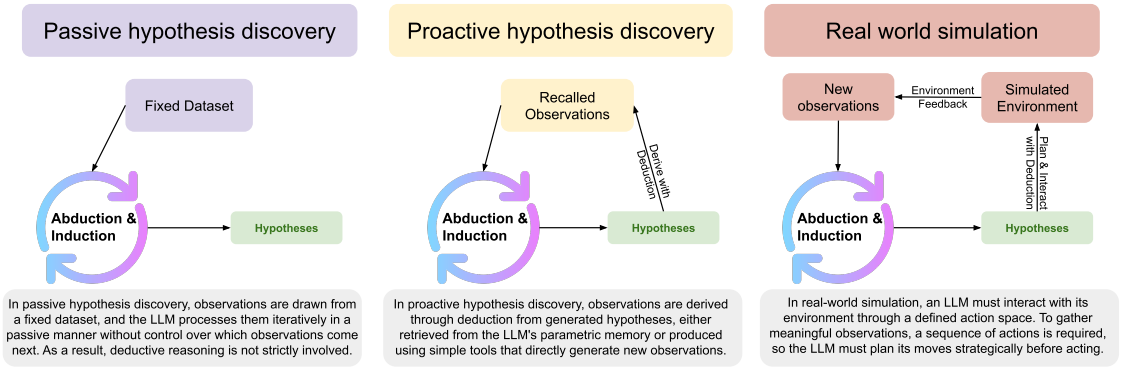

Figure 3: Differences and similarities among different types of hypothesis discovery tasks

added to the rule library, while ineffective ones are discarded. Iteratively, after processing all questions in the benchmark, the LLM builds a rule library containing effective rules for solving the questions.

Zhou et al. (2024) proposed the HypoGeniC framework. Unlike HtT, HypoGeniC is evaluated on more realistic datasets such as Shoe Sales, Deceptive Reviews (Granhag & Vrij, 2005), Headline Popularity (Matias et al., 2021), and Tweet Popularity (Tan et al., 2014). Due to the complexity of real-world data, the generated hypotheses are more nuanced and expressed in natural language. Similar to HtT, HypoGeniC begins by generating a set of candidate hypotheses from a small number of examples. As new observations are processed, each hypothesis is used to make predictions and is assigned a reward based on its accuracy. The system dynamically updates the confidence of each hypothesis; those that consistently perform poorly are removed from the hypothesis bank. New hypotheses are generated from examples that existing hypotheses fail to explain, allowing the model to refine and expand its understanding over time.

Both HypoGeniC and HtT simplify hypothesis discovery by relying on benchmark questions that include ground-truth answers. This configuration allows an external algorithm, not the LLMs themselves, to validate generated hypotheses and update their confidence based on the correctness of predictions. In real-world scenarios, where no ground-truth answers are available, these frameworks become inapplicable and would require substantial adaptation.

7.2 Proactive Hypothesis Discovery

In real-life hypothesis discovery, we do not start with a predefined set of observations that continuously propose new insights. Instead, once an initial hypothesis is formed, we proactively recall our memories or explore further to gather new observations that either strengthen or weaken the hypothesis, allowing us to verify and refine our ideas.

Given a hypothesis, Li et al. (2024) and Jung et al. (2022) propose two proactive methods for hypothesis discovery that both leverage the LLM's parametric memory to generate evidence that either strengthens or weakens the hypothesis. In Hypothesis Testing Prompting, the model directly uses its internal reasoning to evaluate the generated evidence, determining which pieces are more convincing, and then decides whether the hypothesis is correct based on the balance of evidence that strengthens or weakens it. In contrast, Maieutic Prompting iteratively constructs a tree of evidence by generating both strengthening and weakening explanations. It then employs the LLM to assign a belief score (reflecting the model's confidence in the evidence) and a consistency score (measuring how well the evidence aligns with the hypothesis). Finally, a MAX-SAT solver is applied to select the subset of evidence that maximizes the overall scores, thereby determining whether to accept or reject the hypothesis.

Different from relying solely on an LLM's parametric memory to generate new evidence, Seals & Shalin (2024) propose a minimal setting for proactive hypothesis discovery. Inspired by the Wason Task from cognitive science, this task challenges LLMs to prove a formal language hypothesis of the form "if p then q." Here, both p and q are objects described in natural language, for example, "if a person is a man, then he

drinks alcohol." The task provides four cards, each with two sides representing different attributes. Initially, one side of each card is shown, displaying $p, q, \neg p$, and $\neg q$, while the other side reveals the state of another attribute. To rigorously validate the hypothesis "if p then q," one must flip the p card to confirm that its hidden side is q (modus ponens) and flip the $\neg q$ card to check that its hidden side is $\neg p$ (modus tollens). Flipping only these two cards provides sufficient evidence for the hypothesis, while the other two cards do not offer the necessary information. Thus, in this benchmark, by proactively flipping two cards, we can determine whether the LLM can correctly identify natural language expressions of p and q and validate the hypothesis using a minimal action space.

Moreover, Conti et al. (2024) propose APEx, a multimodal automatic benchmarking framework that evaluates hypotheses about large multimodal models in a fully automated and iterative fashion. For example, to test a hypothesis such as "a model is able to identify graffiti-styled images," APEx first leverages text-to-image retrieval and generation tools to create a tailored set of test images. It then employs a range of transformation tools to perform image augmentation, introducing variations that challenge the models' robustness. In an iterative experimental loop, the framework executes these experiments on a library of models, analyzes the results, and refines the testing protocol accordingly.

7.3 Complete Loop: Real-World Hypothesis Discovery Simulation

Other works equip LLM agents with interactive environments that more closely mirror the complexity of real-world hypothesis discovery by combining planning, acting, and evidence collection. For example, Xu et al. (2023) construct a Minecraft–like world in which a "vandal" agent performs up to 26 types of actions (e.g., moving, eating, crafting) to achieve a hidden goal (such as collecting lava or crafting a particular item) and leaves behind tracks as evidence. A detective agent—driven by reinforcement learning to maximize information gain—then gathers those tracks and presents them to an LLM, which must answer a multiple-choice question about the vandal's original objective. Because evidence collection relies on an RL policy rather than LLM planning, however, this setup evaluates only the model's capacity to interpret evidence, not its ability to proactively generate and test hypotheses in a dynamic setting.

Building on this approach, Wang et al. (2022a) introduce 30 scientific tasks drawn from five topics in fifth-grade curricula, ranging from measuring the friction coefficient of an inclined plane to testing electrical conductivity. Here, agents must execute long action sequences and apply deductive reasoning grounded in established theories and definitions to complete each task. Likewise, Jansen et al. (2024) design 120 experiments across eight subjects (e.g., Chemistry, Archaeology), each with three difficulty levels, and allow 14 coarse-grained actions (such as "take," "put," and "move"). Agents are evaluated on (1) task completion, (2) execution of key experimental steps, and (3) accurate hypothesis discovery compared to a ground truth. While these virtual labs simulate multi-step procedures and test hypothesis application, their restricted action spaces support only qualitative inference and preclude the fine-grained interventions needed for quantitative rule-learning.

To address these limitations, He et al. (2024) propose puzzle environments in which agents can input arbitrary integers or letters and receive tailored feedback based on a hidden rule. In this framework, an LLM must iteratively probe the environment, uncover the underlying quantitative rule, and solve the puzzle. Performance is assessed not only by whether the agent solves the puzzle but also by human judgments of the clarity and rigor of its reasoning steps, thereby offering a finer-grained evaluation of both quantitative hypothesis generation and the quality of the model's deductive process.

7.4 Discussion and Future Directions in Hypothesis Discovery

Hypothesis discovery fundamentally differs from isolated reasoning tasks by requiring iterative learning and continuous refinement of hypotheses within dynamic, evolving contexts. Particularly in Real-World simulation scenarios, the decisions and actions taken by an LLM may lead to entirely different trajectories of observation collection, varied learning efficiencies, and alternative hypotheses.

Building effective benchmarks for hypothesis discovery requires constructing rich, realistic environments capable of simulating real-world complexities. These environments should contain diverse, comprehensive

action spaces and varied observational feedback mechanisms. Compared to traditional static, label-based datasets, creating such benchmarks is significantly more labor-intensive, demanding at least two key components: 1, **A set of rules unknown to the LLM** that can be learned within the environment. 2, **A sufficiently expressive action space** that allows the LLM to interact with the environment, receive feedback, and gather new information.

Given that current LLMs are trained on vast quantities of data, there is a risk of hypothesis leakage, where underlying rules might already be implicitly embedded in their parametric memory. For instance, benchmarks such as those introduced by Wang et al. (2022a) often rely on relatively straightforward tasks that do not genuinely necessitate novel hypothesis formation. Conversely, tasks proposed by He et al. (2024), despite aiming to encourage creative hypothesis formation, often yield simplistic, toy-like hypotheses with limited applicability to realistic scenarios. Therefore, future research should aim to develop environments with greater complexity and realism, fostering diverse and genuinely novel hypotheses. Benchmarks should be explicitly designed to push LLMs beyond their pretrained knowledge boundaries, and must provide practical tools for validating newly generated hypotheses. Such realistic simulation environments would address critical challenges such as hypothesis leakage and task oversimplification, ultimately fostering more robust and practical hypothesis discovery capabilities within LLMs.

8 Summary

In this survey, we have presented a comprehensive and structured framework for hypothesis discovery using LLMs, guided by Peirce's reasoning paradigm of abduction, deduction, and induction. Specifically, we systematically explored current methods and benchmarks across the three core components: hypothesis generation, hypothesis application, and hypothesis validation.

Our analysis identifies a significant gap between formal and natural language representations. While formal representations enable rigorous and objective evaluations, they often remain restricted to simplified, artificial scenarios lacking real-world complexity. Conversely, natural language representations effectively capture the nuanced complexities inherent in real-world reasoning tasks, yet suffer from a lack of reliable, rigorous evaluation metrics due to their inherently open-ended nature.

Existing methods, including prompt-based and fine-tuning approaches, demonstrate considerable potential but frequently isolate individual reasoning components. To move forward, we advocate for the development of integrated benchmarks and realistic, dynamic environments that more closely mimic real-world scientific inquiry and hypothesis discovery processes. Such benchmarks should provide rich intermediate Chain-of-Thought data, detailed commonsense reasoning steps, and comprehensive action spaces, thereby bridging the current divide between formal and informal reasoning representations.

Ultimately, establishing environments that demand proactive hypothesis generation, robust application to novel contexts, and rigorous validation against evolving evidence will be crucial. By addressing these challenges, future research will significantly advance the ability of LLMs to not merely execute instructions but to autonomously generate, refine, and validate hypotheses, thus realizing their potential as true engines of discovery and innovation.

References

Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G. Iyer, Mugdha S. Polimera, Michael J. Smith, Tirthankar Ghosal, Marc Huertas-Company, Sandor Kruk, Kevin Schawinski, and Ioana Ciucă. A survey on hypothesis generation for scientific discovery in the era of large language models, 2025. URL `https://arxiv.org/abs/2504.05496`.

Francis Bacon. *Novum organum*. Clarendon press, 1878.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pp. 10308–10330, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.555. URL https://aclanthology.org/2024.acl-long.555/.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

Adib Bazgir, Rama chandra Praneeth Madugula, and Yuwen Zhang. Agentichypothesis: A survey on hypothesis generation using LLM systems. In *Towards Agentic AI for Science: Hypothesis Generation, Comprehension, Quantification, and Validation*, 2025. URL https://openreview.net/forum?id=UeeyfR4CUg.

BIG bench authors. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. Abductive commonsense reasoning, 2020. URL https://arxiv.org/abs/1908.05739.

Chen Bowen, Rune Sætre, and Yusuke Miyao. A comprehensive evaluation of inductive reasoning capabilities and problem solving in large language models. In Yvette Graham and Matthew Purver (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 323–339, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.22/.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Arthur W. Burks. Peirce's theory of abduction. *Philosophy of Science*, 13(4):301–306, 1946. ISSN 00318248, 1539767X. URL http://www.jstor.org/stable/185210.

Chengkun Cai, Xu Zhao, Haoliang Liu, Zhongyu Jiang, Tianfang Zhang, Zongkai Wu, Jenq-Neng Hwang, Serge Belongie, and Lei Li. The role of deductive and inductive reasoning in large language models, 2025. URL https://arxiv.org/abs/2410.02892.

Miaosen Chai, Emily Herron, Erick Cervantes, and Tirthankar Ghosal. Exploring scientific hypothesis generation with mamba. In Lotem Peled-Cohen, Nitay Calderon, Shir Lissak, and Roi Reichart (eds.), *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 197–207, Miami, FL, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.nlp4science-1.17. URL https://aclanthology.org/2024.nlp4science-1.17/.

Chunkit Chan, Xin Liu, Tsz Ho Chan, Jiayang Cheng, Yangqiu Song, Ginny Wong, and Simon See. Self-consistent narrative prompts on abductive natural language inference. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1040–1057, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.67. URL https://aclanthology.org/2023.ijcnlp-main.67/.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery, 2024. URL https://arxiv.org/abs/2410.05080.

Kewei Cheng, Jingfeng Yang, Haoming Jiang, Zhengyang Wang, Binxuan Huang, Ruirui Li, Shiyang Li, Zheng Li, Yifan Gao, Xian Li, Bing Yin, and Yizhou Sun. Inductive or deductive? rethinking the fundamental reasoning abilities of llms, 2024. URL https://arxiv.org/abs/2408.00114.

François Chollet. On the measure of intelligence, 2019. URL https://arxiv.org/abs/1911.01547.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Alessandro Conti, Enrico Fini, Paolo Rota, Yiming Wang, Massimiliano Mancini, and Elisa Ricci. Automatic benchmarking of large multimodal models via iterative experiment programming, 2024. URL https://arxiv.org/abs/2406.12321.

Maksym Del and Mark Fishel. True detective: A deep abductive reasoning benchmark undoable for GPT-3 and challenging for GPT-4. In Alexis Palmer and Jose Camacho-collados (eds.), *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pp. 314–322, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.starsem-1.28. URL https://aclanthology.org/2023.starsem-1.28/.

Igor Douven. Abduction. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition, 2021.

Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. Transforming science with large language models: A survey on ai-assisted scientific discovery, experimentation, content generation, and evaluation, 2025. URL https://arxiv.org/abs/2502.05151.

Harry G. Frankfurt. Peirce's notion of abduction. *The Journal of Philosophy*, 55(14):593–597, 1958. ISSN 0022362X. URL http://www.jstor.org/stable/2021966.

Pär Anders Granhag and Aldert Vrij. Deception detection. *Psychology and law: An empirical perspective*, pp. 43–92, 2005.

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. Agentic ai for scientific discovery: A survey of progress, challenges, and future directions, 2025. URL https://arxiv.org/abs/2503.08979.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alex Wardle-Solano, Hannah Szabo, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander R. Fabbri, Wojciech Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. Folio: Natural language reasoning with first-order logic, 2024. URL https://arxiv.org/abs/2209.00840.

Steve Hanks and Drew McDermott. Nonmonotonic logic and temporal projection. *Artificial intelligence*, 33 (3):379–412, 1987.

Gilbert H. Harman. The inference to the best explanation. *The Philosophical Review*, 74(1):88–95, 1965. ISSN 00318108, 15581470. URL http://www.jstor.org/stable/2183532.

Jinwei He and Feng Lu. Causejudger: Identifying the cause with llms for abductive logical reasoning, 2024. URL https://arxiv.org/abs/2409.05559.

Kaiyu He, Mian Zhang, Shuo Yan, Peilin Wu, and Zhiyu Zoey Chen. Idea: Enhancing the rule learning ability of large language model agent through induction, deduction, and abduction, 2024. URL https://arxiv.org/abs/2408.10455.

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, 2024. URL https://arxiv.org/abs/2410.14255.

Wenyue Hua, Tyler Wong, Sun Fei, Liangming Pan, Adam Jardine, and William Yang Wang. Inductionbench: Llms fail in the simplest complexity class, 2025. URL https://arxiv.org/abs/2502.15823.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67/.

Kang il Lee, Hyukhun Koh, Dongryeol Lee, Seunghyun Yoon, Minsung Kim, and Kyomin Jung. Generating diverse hypotheses for inductive reasoning, 2025. URL https://arxiv.org/abs/2412.13422.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. *Advances in Neural Information Processing Systems*, 37:10088–10116, 2024.

Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. Brainteaser: Lateral thinking puzzles for large language models, 2023. URL https://arxiv.org/abs/2310.05057.

Cheng Jiayang, Lin Qiu, Tsz Chan, Tianqing Fang, Weiqi Wang, Chunkit Chan, Dongyu Ru, Qipeng Guo, Hongming Zhang, Yangqiu Song, Yue Zhang, and Zheng Zhang. StoryAnalogy: Deriving story-level analogies from large language models to unlock analogical understanding. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11518–11537, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.706. URL https://aclanthology.org/2023.emnlp-main.706/.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1266–1279, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.82. URL https://aclanthology.org/2022.emnlp-main.82/.

Aditya Kalyanpur, Kailash Karthik Saravanakumar, Victor Barres, Jennifer Chu-Carroll, David Melville, and David Ferrucci. Llm-arc: Enhancing llms with an automated reasoning critic, 2024. URL https://arxiv.org/abs/2406.17663.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 537–563, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.38. URL https://aclanthology.org/2022.findings-emnlp.38/.

Larry Laudan. William whewell on the consilience of inductions. *The Monist*, pp. 368–391, 1971.

Jiachun Li, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. Mirage: Evaluating and explaining inductive reasoning process in language models, 2025a. URL https://arxiv.org/abs/2410.09542.

Yitian Li, Jidong Tian, Hao He, and Yaohui Jin. Hypothesis testing prompting improves deductive reasoning in large language models, 2024. URL https://arxiv.org/abs/2405.06707.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. From system 1 to system 2: A survey of reasoning large language models, 2025b. URL https://arxiv.org/abs/2502.17419.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. Deductive verification of chain-of-thought reasoning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 36407–36433. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/72393bd47a35f5b3bee4c609e7bba733-Paper-Conference.pdf.

Emmy Liu, Graham Neubig, and Jacob Andreas. An incomplete loop: Instruction inference, instruction following, and in-context learning in language models, 2024. URL https://arxiv.org/abs/2404.03028.

Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. Logical reasoning in large language models: A survey, 2025. URL https://arxiv.org/abs/2502.09100.

Jorge Nathan Matias, Kevin Munger, Marianne Aubin Le Quere, and Charles R. Ebersole. The upworthy research archive, a time series of 32,487 experiments in U.S. media. *Scientific Data*, 8, 2021. URL https://api.semanticscholar.org/CorpusID:236883026.

John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pp. 431–450. Elsevier, 1981.

John Stuart Mill. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and Methods of Scientific Investigation: Vol. I*. BoD–Books on Demand, 2024.

Gerhard Minnameier. Peirce-suit of truth –why inference to the best explanation and abduction ought not to be confused. *Erkenntnis*, 60(1):75–105, 2004. doi: 10.1023/B:ERKE.0000005162.52052.7f. URL https://doi.org/10.1023/B:ERKE.0000005162.52052.7f.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

Rajiv Movva, Kenny Peng, Nikhil Garg, Jon Kleinberg, and Emma Pierson. Sparse autoencoders for hypothesis generation, 2025. URL https://arxiv.org/abs/2502.04382.

Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Basel Alomair, Dan Hendrycks, and David Wagner. Can llms follow simple rules?, 2024. URL https://arxiv.org/abs/2311.04235.

Ha-Thanh Nguyen, Randy Goebel, Francesca Toni, Kostas Stathis, and Ken Satoh. How well do sota legal reasoning models support abductive reasoning?, 2023. URL https://arxiv.org/abs/2304.06912.

Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges, 2024. URL https://arxiv.org/abs/2409.02387.

Chitu Okoli. Inductive, abductive and deductive theorising. *International Journal of Management Concepts and Philosophy*, 16(3):302–316, 2023.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5153–5176, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.313. URL https://aclanthology.org/2023.emnlp-main.313/.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3806–3824, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.248. URL https://aclanthology.org/2023.findings-emnlp.248/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Charles Sanders Peirce. *Collected papers of charles sanders peirce*, volume 5. Harvard University Press, 1974.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. TopicGPT: A prompt-based topic modeling framework. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2956–2984, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.164. URL https://aclanthology.org/2024.naacl-long.164/.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey, 2024. URL https://arxiv.org/abs/2407.11511.

Karl Popper. *The logic of scientific discovery*. Routledge, 2005.

Kevin Pu, KJ Kevin Feng, Tovi Grossman, Tom Hope, Bhavana Dalvi Mishra, Matt Latzke, Jonathan Bragg, Joseph Chee Chang, and Pao Siangliulue. Ideasynth: Iterative research idea development through evolving and composing idea facets with literature-grounded feedback. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–31, 2025.

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, 2024. URL https://arxiv.org/abs/2407.08940.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=bNt7oajl2a.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Chandan K Reddy and Parshin Shojaee. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28601–28609, 2025.

Raymond Reiter. A logic for default reasoning. *Artificial intelligence*, 13(1-2):81–132, 1980.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Thinking like a skeptic: Defeasible inference in natural language. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4661–4675, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.418. URL https://aclanthology.org/2020.findings-emnlp.418/.

Joshua Stewart Rule. *The child as hacker: building more human-like models of learning*. PhD thesis, Massachusetts Institute of Technology, 2020.

S Seals and Valerie Shalin. Evaluating the deductive competence of large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8614–8630, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.476. URL https://aclanthology.org/2024.naacl-long.476/.

Xiaoming Shi, Siqiao Xue, Kangrui Wang, Fan Zhou, James Zhang, Jun Zhou, Chenhao Tan, and Hongyuan Mei. Language models can improve event prediction by few-shot abductive reasoning. *Advances in Neural Information Processing Systems*, 36:29532–29557, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36: 8634–8652, 2023.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers, 2024. URL https://arxiv.org/abs/2409.04109.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4506–4515, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1458. URL https://aclanthology.org/D19-1458/.

Friedrich Stadler. The road to experience and prediction from within: Hans reichenbach's scientific correspondence from berlin to istanbul. *Synthese*, 181:137–155, 2011.

Wangtao Sun, Chenxiang Zhang, XueYou Zhang, Xuanqing Yu, Ziyang Huang, Pei Chen, Haotian Xu, Shizhu He, Jun Zhao, and Kang Liu. Beyond instruction following: Evaluating inferential rule following of large language models, 2024. URL https://arxiv.org/abs/2407.08440.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

Chenhao Tan, Lillian Lee, and Bo Pang. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of ACL*, 2014.

Fuchun Wang, Xian Zhou, Wenpeng Hu, Zhunchen Luo, Wei Luo, and Xiaoying Bai. Llm assists hypothesis generation and testing for deliberative questions. In Derek F. Wong, Zhongyu Wei, and Muyun Yang (eds.), *Natural Language Processing and Chinese Computing*, pp. 424–436, Singapore, 2025. Springer Nature Singapore. ISBN 978-981-97-9434-8.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D. Goodman. Hypothesis search: Inductive reasoning with language models, 2024. URL https://arxiv.org/abs/2309.05660.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader?, 2022a. URL https://arxiv.org/abs/2203.07540.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. Scienceworld: Is your agent smarter than a 5th grader?, 2022b. URL https://arxiv.org/abs/2203.07540.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks, 2015. URL https://arxiv.org/abs/1502.05698.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. Reframing human-AI collaboration for generating free-text explanations. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 632–658, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.47. URL https://aclanthology.org/2022.naacl-main.47/.

Guangzhi Xiong, Eric Xie, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. Improving scientific hypothesis generation with knowledge grounded large language models, 2024. URL https://arxiv.org/abs/2411.02382.

Manjie Xu, Guangyuan Jiang, Wei Liang, Chi Zhang, and Yixin Zhu. Active reasoning in an open-world environment. *Advances in Neural Information Processing Systems*, 36:11716–11736, 2023.

Yang Yan, Yu Lu, Renjun Xu, and Zhenzhong Lan. Do phd-level llms truly grasp elementary addition? probing rule learning vs. memorization in large language models. *arXiv preprint arXiv:2504.05262*, 2025.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. Language models as inductive reasoners. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 209–225, St. Julian's, Malta, March 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.13/.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13545–13565, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.804. URL https://aclanthology.org/2024.findings-acl.804/.

Zonglin Yang, Xinya Du, Rui Mao, Jinjie Ni, and Erik Cambria. Logical reasoning over natural language as knowledge representation: A survey, 2024c. URL https://arxiv.org/abs/2303.12023.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses, 2025. URL https://arxiv.org/abs/2410.07076.

Nathan Young, Qiming Bao, Joshua Bensemann, and Michael Witbrock. AbductionRules: Training transformers to explain unexpected inputs. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 218–227, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.19. URL https://aclanthology.org/2022.findings-acl.19/.

Fei Yu, Hongbo Zhang, Prayag Tiwari, and Benyou Wang. Natural language reasoning, a survey. *ACM Comput. Surv.*, 56(12), October 2024a. ISSN 0360-0300. doi: 10.1145/3664194. URL https://doi.org/10.1145/3664194.

Qingchen Yu, Shichao Song, Ke Fang, Yunfeng Shi, Zifan Zheng, Hanyu Wang, Simin Niu, and Zhiyu Li. Turtlebench: Evaluating top language models via real-world yes/no puzzles, 2024b. URL https://arxiv.org/abs/2410.05262.

Siyu Yuan, Jiangjie Chen, Xuyang Ge, Yanghua Xiao, and Deqing Yang. Beneath surface similarity: Large language models make reasonable scientific analogies after structure abduction. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2446–2460, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.160. URL https://aclanthology.org/2023.findings-emnlp.160/.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. WinoWhy: A deep diagnosis of essential commonsense knowledge for answering Winograd schema challenge. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5736–5745, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.508. URL https://aclanthology.org/2020.acl-main.508/.

Yue Zhang, Liqiang Jing, and Vibhav Gogate. Defeasible visual entailment: Benchmark, evaluator, and reward-driven optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25976–25984, 2025.

Wenting Zhao, Justin T. Chiu, Claire Cardie, and Alexander M. Rush. Abductive commonsense reasoning exploiting mutually exclusive explanations, 2023. URL https://arxiv.org/abs/2305.14618.

Wenting Zhao, Justin Chiu, Jena Hwang, Faeze Brahman, Jack Hessel, Sanjiban Choudhury, Yejin Choi, Xiang Li, and Alane Suhr. UNcommonsense reasoning: Abductive reasoning about uncommon situations. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8487–8505, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.469. URL https://aclanthology.org/2024.naacl-long.469/.

Ruiqi Zhong, Heng Wang, Dan Klein, and Jacob Steinhardt. Explaining datasets in words: Statistical models with natural language parameters. *Advances in Neural Information Processing Systems*, 37:79350–79380, 2024.

Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. Hypothesis generation with large language models. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pp. 117–139. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.nlp4science-1.10. URL http://dx.doi.org/10.18653/v1/2024.nlp4science-1.10.

Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. Large language models can learn rules, 2024. URL https://arxiv.org/abs/2310.07064.