A Survey of Inductive Reasoning for Large Language Models

Anonymous ACL submission

Abstract

Reasoning is an important task for large language models (LLMs). Among all the reasoning paradigms, inductive reasoning is one of the basic types, which is characterized by its particular-to-general thinking process and the non-uniqueness of its answers. The inductive mode is crucial for knowledge generalization and aligns better with human cognition, so it is a fundamental mode of learning, hence attracting increasing interest. Despite the importance of inductive reasoning, there is no systematic summary of it. Therefore, this paper presents the first comprehensive survey of inductive reasoning for LLMs. First, methods for improving inductive reasoning are categorized into three main areas: post-training enhancement, test-time exploration, and data augmentation. Then, current benchmarks of inductive reasoning are summarized, and a unified sandbox-based evaluation approach with the observation coverage metric is derived. Finally, we offer some analyses regarding the source of inductive ability and how simple model architectures and data help with inductive tasks, providing a solid foundation for future research.

1 Introduction

In recent years, the rapid development of large language models (LLMs) (Zhao et al., 2023) leads to significant progress in many natural language processing (NLP) downstream tasks. Among these, reasoning (Huang and Chang, 2022; Plaat et al., 2024; Zhang et al., 2025a) is one of the comprehensive and challenging tasks for LLMs, and therefore receives considerable attention.

Within all the reasoning paradigms, inductive reasoning (Lu, 2024) is one of the fundamental types. It involves drawing general conclusions from specific observations (Han et al., 2024). We give two examples illustrating number sequence calculation and list transformation in Figure 1. The main characteristics of inductive reasoning are its particular-to-general thinking process and the non-uniqueness of its answers. Considering how humans perceive the world, they typically make judgments by drawing analogies from past experiences to current situations, rather than always going through a strictly logical process as in deductive reasoning (Ingold, 2021). We can assume that the inductive mode is key to knowledge generalization and better aligns with human cognition. It is a fundamental mode of learning and thus attracts increasing interest.

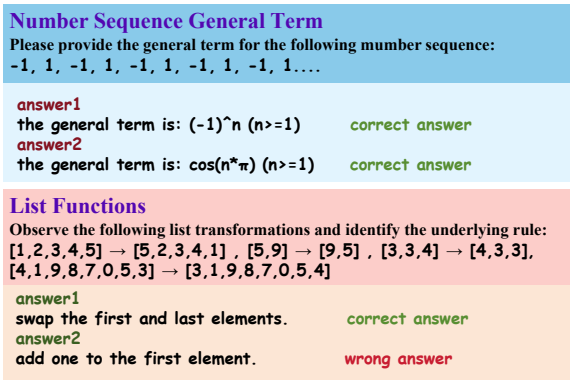


Figure 1: Two examples of inductive reasoning. They generalize from specific observations or cases to derive general conclusions or rules. There may be more than one such conclusion that meets all the observations.

Despite the importance of inductive reasoning, previous works mostly focus on deductive reasoning (Li et al., 2024b), represented by mathematical proof (Ahn et al., 2024; Chen et al., 2024b) and program verification (Liu et al., 2023; Jiang et al., 2024b), which is a logical reasoning that derives necessary conclusions from general rules or premises. For more conceptual distinctions, please refer to Appendix A.2. It has already been extensively studied in recent years (Lu et al., 2024; Wang et al., 2024a). Moreover, there is no systematic summary of inductive reasoning for LLMs.

Therefore, this paper presents the first compre-

hensive survey of inductive reasoning for LLMs. We introduce the background, including relevant concepts, applications in NLP and real-world scenarios, as well as the significance (Section 2) at the beginning. The main body consists of three parts. First, we review and prospect the methods for enhancing the inductive reasoning capabilities of LLMs (Section 3), which are categorized into three main areas: post-training enhancement, test-time exploration, and data augmentation. We then summarize the current benchmarks of inductive reasoning and derive a unified sandbox-based evaluation approach with the observation coverage metric (Section 4). Finally, we offer some theoretical analyses of inductive reasoning (Section 5), regarding the sources of inductive ability and practical experiences for enhancing it. The taxonomy of this survey is in Figure 2. In summary, the main contributions of this paper are threefold:

- **First survey.** To our knowledge, we are the first to present a comprehensive survey of inductive reasoning for LLMs, thoroughly analyzing the current techniques and applications.
- **New taxonomy.** We categorize methods for improving inductive reasoning into: post-training enhancement, test-time exploration, and data augmentation. We also summarize the current benchmarks and derive a unified sandbox-based evaluation approach.
- **Bright prospect.** We offer some analyses regarding the source of inductive ability and how simple model architectures and data help with inductive tasks, providing a solid foundation for future research (Appendix A.8).

2 Background

In this section, we will introduce the relevant concepts of inductive reasoning, some of its application scenarios, and the significance of studying it.

2.1 Concepts

2.1.1 Large Language Models

Since the transformer architecture (Vaswani et al., 2017) has become mainstream for language models, the field of NLP has experienced rapid development (Wei et al., 2021). During 2017 to 2022, pretrained language models (PLMs) (Kalyan et al., 2021), which undergo two stages—pretraining and finetuning—such as the BERT (Devlin et al., 2019) and T5 (Raffel et al., 2023) series, once dominated the entire field. From 2022, with the advent of ChatGPT-3.5 (Ye et al., 2023), the era of LLMs officially begins. LLMs, with their massive number of parameters and unique training methods (Zhao et al., 2025), significantly improve the performance of NLP tasks (She et al., 2023; Xu et al., 2023; Plaat et al., 2024) and have a profound impact on various aspects of daily life (Gan et al., 2023; Zhou et al., 2023; Maatouk et al., 2023). Some well-known LLMs include the GPT series[1], the Gemini series[2], the Claude series[3], and so on.

2.1.2 Inductive Reasoning

Inductive reasoning represents ***making an induction from specific instances or observations to derive general rules and conclusions*** (Arthur, 1994; Heit, 2000). From another perspective, it denotes one reasoning approach where the conclusion is not guaranteed with certainty, but instead supported only to a certain degree of probability (Copi et al., 2004). In other words, inductive reasoning may have more than one valid hypothesis that can account for all the instances or observations, making its answer open (Thomas, 2003). To sum up abstractly, inductive reasoning is a thought process that proceeds from the particular to the general.

2.2 Applications of Inductive Reasoning

The core idea of inductive reasoning is inductive bias. It is a set of assumptions or prior conditions that a model or an individual relies on when encountering unseen items (Caruana, 1993; Baxter, 2000). There is no 'universal' bias in deep learning. Choosing an appropriate inductive bias for a specific task is key to achieving success (Provost and Buchanan, 1995). We will discuss its applications in NLP tasks and real-world scenarios.

2.2.1 NLP Downstream Tasks

Inductive reasoning is widely applied to improve the performance of NLP downstream tasks (Appendix A.3). Some common practices include training models to learn inductive bias (Yang et al., 2025), constructing chains of thought (CoT) (Chen et al., 2025b) or summarizing rules to enhance interpretability (Xu and Yang, 2025), and leveraging intra-parameters implicit knowledge (Cheng et al., 2024b) for induction, and others. Such approaches

[1] https://chatgpt.com/overview
[2] https://cloud.google.com/vertex-ai/generative-ai/docs
[3] https://claude.ai

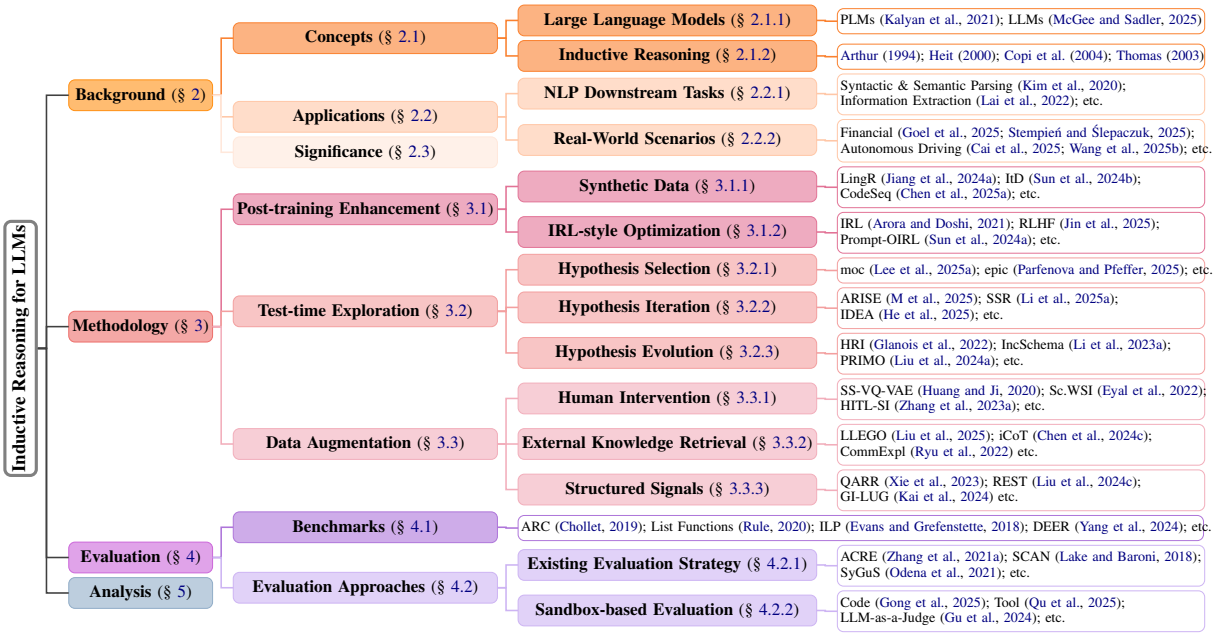


Figure 2: Taxonomy of the survey about the inductive reasoning for LLMs.

benefit a wide range of downstream NLP tasks: syntactic and semantic parsing (Kim et al., 2020; Yamada et al., 2021; Lindemann et al., 2024; Tsujimoto et al., 2025), information extraction (Lai et al., 2022; Liu et al., 2024c; Silva et al., 2025; Xu et al., 2024), dialogue systems (Feng et al., 2022; Xie et al., 2024; Ou et al., 2024), question answering (Gu and Su, 2022; Kim et al., 2023; Chen et al., 2024e), and multimodal tasks (Amosy et al., 2024; Zhou et al., 2025; Naik et al., 2025).

2.2.2 Real-World Scenarios

Inductive reasoning has a broad impact on real-world scenarios. We list three of them. (1) Financial forecasting: Inductive models are essential for predicting future financial outcomes by learning complex, non-linear patterns from vast amounts of historical time-series data (Faheem, 2021; Goel et al., 2025; Stempień and Ślepaczuk, 2025). (2) Autonomous driving: Inductive reasoning enables autonomous driving systems to generalize from historical data and past experiences to recognize rules in traffic conditions to make decisions.(Cai et al., 2025; Wang et al., 2025b). (3) Conversational healthcare and diagnostic dialogue: Inductive reasoning empowers artificial intelligence systems to mimic a clinician's process of taking patient history and formulating a diagnosis by generalizing from symptom patterns (Tu et al., 2024; Dhudum et al., 2024; Zhang et al., 2025b). Broader real-world applications are in Appendix A.8.

2.3 Significance of Inductive Reasoning

Inductive reasoning has broad applications in both AI and real-world scenarios. It is the most universal and essential method in knowledge discovery and generalization (Carter and Hamilton, 1998; Bai et al., 2024; Sun et al., 2025): (1) Deriving general conclusions from specific cases, allowing it to cover and generalize to a wider range of applications, which aligns with the human learning process. (2) Adaptive adjustments help in uncertain and complex scenarios, where inductive reasoning may yield multiple plausible outcomes rather than a single unique solution.

3 Methodology

In this section, we will introduce three major approaches to enhance the inductive capabilities of LLMs: post-training enhancement(Section 3.1), test-time exploration (Section 3.2), and data augmentation (Section 3.3). It is worth noting that we not only summarize existing methods but also prospect a forward-looking review of potential future inductive approaches. For convenience, we treat the inputs of inductive reasoning as observations and refer to these outputs as rules. Boundaries and comparisons of the three methods can be found in the Appendix A.4.

3.1 Post-training Enhancement

Post-training enhancement refers to improving the inductive reasoning ability of LLMs during the

post-training stage (Lai et al., 2025; Wu, 2025), using algorithms such as supervised finetuning (SFT) and reinforcement learning (RL). This category of methods primarily focuses on constructing synthetic data (Long et al., 2024; Jiang et al., 2025) (Section 3.1.1) and developing new algorithms (Section 3.1.2). We illustrate them in Figure 3.

3.1.1 Synthetic Data

Synthetic data means artificially generated data that mimics the properties and patterns of real-world data (Bauer et al., 2024). Data plays a decisive role in LLM training. To address certain inherent limitations of natural data, such as being difficult to obtain or organize (Nadas et al., 2025), researchers often construct data manually to compensate for these shortcomings. LingR (Jiang et al., 2024a) builds a 'linguistic rule instruction set' for various LLMs, enabling them to learn step-by-step reasoning based on linguistic rules such as causality. ItD (Sun et al., 2024b) leverages the deductive abilities of LLMs to generate data and optimize inductive ability. The model's capacity to learn general rules from a small number of samples is significantly enhanced. CodeSeq (Chen et al., 2025a) constructs SFT and RL training sets to ask LLMs to facilitate reasoning over number sequence general term formulas, thereby improving their inductive abilities. Other approaches (Wu et al., 2022; Aksu et al., 2023; Darm et al., 2023; Mosolova et al., 2025; Li et al., 2025b) establish similar induction-related training datasets for the models to learn from.

3.1.2 IRL-style Optimization

Reward models (RMs) (Zhong et al., 2025) are typically utilized to provide supervision signals for the RL process. However, for inductive reasoning, due to the non-uniqueness of answers and the uncertainty in the reasoning process, traditional RMs struggle to provide effective supervision. Therefore, Inverse RL (Arora and Doshi, 2021) (IRL), which needs to induce the latent reward functions, may serve as an alternative approach (Sun and van der Schaar, 2025). The Reinforcement Learning from Human Feedback (RLHF) (Kaufmann et al., 2024; Swamy et al., 2024) process of LLMs is essentially IRL, as it infers human preferences and the underlying reward function from human feedback. Therefore, designing an appropriate reward model in RLHF can enhance the inductive reasoning ability of LLMs (Jin et al., 2025). We can also employ Prompt-OIRL (Sun et al., 2024a) for reference, which proposes an IRL-based method that reuses historical prompting trial-and-error experience to train a reward model to improve the model's inductive exploration ability. Although works combining IRL and reasoning are still scarce, the approach of IRL—fitting the posterior distribution of the reward model from human or data signals (Cai et al., 2024; Krishna and Sahoo, 2024)—has strong extensibility and can thus be regarded as one of the important methods for the facilitation of inductive reasoning.

3.2 Test-time Exploration

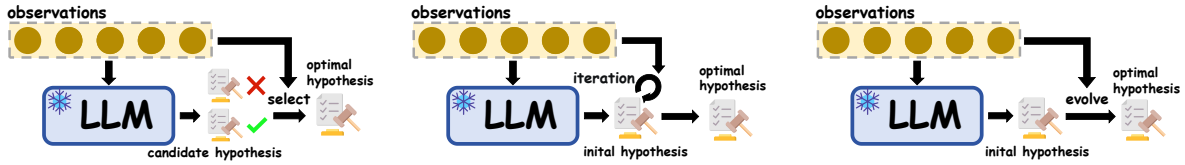
The goal of inductive reasoning is to derive general rules from observations, which inevitably involves forming hypotheses during the reasoning process. The above post-training enhancement requires training LLMs, whereas the test-time exploration in this section is a hypothesis-based method that only works during the inference stage (Zhang et al., 2025c). Unlike end-to-end models (Kotary et al., 2021), the test-time exploration method prompts the frozen LLMs to form an inductive reasoning pipeline (He and Chen, 2025). We have LLMs generate candidate hypotheses for inductive problems, which can then undergo selection (Section 3.2.1), iteration (Section 3.2.2), or evolution (Section 3.2.3) operations to reach the optimal one. Detailed processes are shown in Figure 4.

3.2.1 Hypothesis Selection

Hypothesis selection refers to choosing, from the candidate hypotheses generated by LLMs, those that can cover the observations (Pazzani and Silverstein, 1990; Bun et al., 2019). Hypothesis Search (Wang et al., 2024b) let the LLMs generate multiple abstract hypotheses in natural language. Then, narrow down the hypothesis set through either the LLMs or minimal human filtering. The motivation of Mixture of Concepts (MoC) (Lee et al., 2025a) lies in the fact that hypotheses for inductive reasoning often produce semantic redundancy. Therefore, the proposed method simulates human inductive reasoning by first figuring out a list of semantically non-redundant concepts and then generating corresponding hypotheses based on each concept. Parfenova and Pfeffer (2025) proposes Ensemble Pipeline for Inductive Coding (EPIC) to address the issue of inconsistency in inductive encodings by using small LLMs to generate candidate encodings, filtering them through a moderator mechanism and similarity checks, and finally evaluating them with

(a) synthetic data (b) IRL-style optimization

Figure 3: The demonstration of the post-training enhancement method for inductive reasoning.



(a) hypothesis selection (b) hypothesis iteration (c) hypothesis evolution

Figure 4: The demonstration of the test-time exploration method for inductive reasoning.

composite metrics.

3.2.2 Hypothesis Iteration

Hypothesis iteration means iterating over candidate hypotheses until they satisfy all the observations (Yom, 2015). Qiu et al. (2024) proposes a three-step iterative hypothesis refinement method that simulates the human inductive reasoning process: generate multiple hypotheses from a few examples; evaluate how many known instances each hypothesis can cover; and have the LLMs further revise the selected hypotheses based on feedback, iterating for several rounds until convergence. SSR (Li et al., 2025a) iteratively optimizes by generating diverse candidate rules and refining them based on execution feedback. ARISE (M et al., 2025) also iterates over the inductive rules before using them to train the model. IDEA framework (He et al., 2025) resolves the shortcomings of LLMs in interactive rule learning by simulating the human cycle of hypothesis revision, thereby enhancing the model's dynamic learning capability.

3.2.3 Hypothesis Evolution

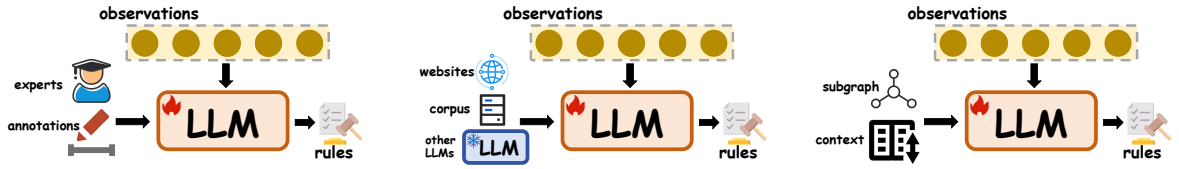
Unlike the iterative process, hypothesis evolution expands, diversifies, or evolves the hypothesis space by generating, filtering, and combining multiple hypotheses, forming hypotheses that better capture complex patterns (Galkin, 2011; Gil et al., 2017; Juretic, 2025). LLMs leverage contextual and label information, along with prompts, to progressively guide the model in dynamically generating patterns during reasoning, without relying on predefined rules (Dror et al., 2023). IncSchema (Li et al., 2023a) gradually induces general patterns by querying the LLMs in stages—first listing core events, then expanding details, and finally verifying relationships. HRI (Glanois et al., 2022) generates inductive meta-rules and matches them with samples, thereby evolving into first-order logic rules. PRIMO (Liu et al., 2024a) introduces a progressive multi-stage open rule induction method for deriving multi-hop rules, thereby capturing more complex reasoning chains.

3.3 Data Augmentation

Data augmentation (Zhang et al., 2022) for LLMs signifies enriching the model's input with additional knowledge or structured signals, such as external facts and retrieved documents, to enhance reasoning and improve output quality. We divide it into three subcategories: human intervention (Section 3.3.1), external knowledge (Section 3.3.2), and structured signals (Section 3.3.3). Please refer to Figure 5 about them.

3.3.1 Human Intervention

Human intervention incorporates expert knowledge or human-annotated information during inductive reasoning. SS-VQ-VAE (Huang and Ji, 2020) relies on a small amount of human-annotated information to discover new patterns. Eyal et al. (2022) generates substitute words, then annotates the corpus and trains static embeddings to enhance the model's inductive priors. Zhang et al. (2023a) utilizes GPT-3 to generate candidate patterns and enhances their quality through human intervention, addressing the issues of domain transfer and semantic consistency in purely automated approaches. Some other studies (Edwards and Ji, 2023; Verhoeven et al., 2024)



(a) human intervention (b) external knowledge (c) structured signals

Figure 5: The demonstration of the data augmentation method for inductive reasoning.

emphasize inductive capabilities in low-annotation scenarios, which indirectly reflects the importance of expert knowledge.

3.3.2 External Knowledge

In this paper, we define external knowledge (Cao et al., 2021) to include web or document information, knowledge from other corpora, knowledge stored in LLM parameters (AlKhamissi et al., 2022), and so on. LLEGO (Liu et al., 2025) incorporates the semantic prior knowledge embedded in large LLMs into genetic programming operations to enhance generalization ability. The parameter knowledge of LLMs (Zhang et al., 2023b) and multimodal large models (Zhang et al., 2021b; Li et al., 2024a) can also serve as an important supplementary information source for inductive tasks. For example, some powerful LLMs are directly prompted to produce the inductive Chain-of-Thought (Chen et al., 2024c), inductive steps (Qian et al., 2023), and inductive rules (Ramji and Ramji, 2025) for the current task, providing additional assistance. Other types of knowledge, such as bilingual corpora (Shi et al., 2021; Kohli et al., 2024), social media content (Radhakrishnan et al., 2020), and commonsense knowledge (Ryu et al., 2022), can be used for inductive tasks in the same way.

3.3.3 Structured Signals

Structured information represents subgraphs or contextual information of LLMs (neighboring hidden states or embeddings), which provide local implicit signals and help LLMs to learn relevant inductive biases (Immer et al., 2022). Li et al. (2023b) optimizes the model's output by retrieving nearest-neighbor embeddings as contextual examples. QARR (Xie et al., 2023) extracts an open subgraph for the query entity to inductively infer new entities. REST (Liu et al., 2024c) deploys rule-induced subgraphs to capture local semantic patterns, thereby enhancing the model's generalization ability in inductive scenarios. GI-LUG (Kai et al., 2024) uses a syntactic parser to generate syntax masks that guide the attention mechanism, and combine BPE embeddings with a hybrid loss function to optimize the induction process. Although this type of method is widely used in the PLM era, due to the same underlying principle, it can also play an important role for LLMs.

4 Evaluation

In this section, we will introduce current benchmarks for LLM inductive reasoning, some evaluation approaches, and the corresponding metrics.

4.1 Benchmarks

To evaluate the inductive reasoning capabilities of LLMs, the research community constructs a diverse set of benchmarks, as shown in Table 1, that comprehensively assess the models' ability to generalize universal rules from concrete observations. It is noteworthy that the input formats of some tasks appear as paired samples or few-shot examples, often framed as analogy reasoning. As we claim in Appendix A.2, since analogical reasoning is a special form of inductive reasoning, we also regard benchmarks in this analogical form as benchmarks for inductive reasoning. The core task of these inductive benchmarks requires models to observe a small number of input examples (Observation Input), infer underlying patterns, and output the final rules (Induction Target).

As shown in Table 1, the data objects covered by these benchmarks span a wide range—from basic structures such as numbers, strings, and lists, to more complex forms like grids, logical formulas, and even natural language text.

Among them, benchmarks such as ARC (Chollet, 2019), List Functions (Rule, 2020), and SyGuS (Odena et al., 2021) focus on algorithmic or rule learning, requiring models to generate programs or operational rules that explain data transformations. What's more, tasks like ILP (Evans and Grefenstette, 2018), GeoILP (Chen et al., 2025c), and ACRE (Zhang et al., 2021a) place greater emphasis on the induction of logical concepts and symbolic

Table 1: Some benchmarks for evaluating the inductive reasoning abilities of LLMs. We provide the atomic objects, names with their references, the input formations, the targets to be induced, and the number of test samples (approximate values). '.' represents that it is the abbreviation of the benchmark name, while '*' indicates that the data are presented in the form of analogical reasoning. Further details about these benchmarks are in Appendix A.5.

Object	Benchmark Name	Obervation Input	Induction Target	# Samples
symbol	ILP (Evans and Grefenstette, 2018)	pos. and neg. samples	a one-order logic rule	1,500
entity	SCAN (Lake and Baroni, 2018)	a state of entities	an action of the state	4,000
grid	ARC* (Chollet, 2019)	pairs of grids	a grid conversion rule	400
list	List Func.* (Rule, 2020)	pairs of number lists	a list operation rule	1,200
code	PROGES (Alet et al., 2021)	IO input/output	a program	270,000
string	SyGuS (Odena et al., 2021)	a pair of strings	a string-mapping program	8,00
entity	ACRE (Zhang et al., 2021a)	functions of entities	a 'Blickets' entity	6,000
text	Instruc. (Honovich et al., 2022)	two NL sentences	an instruction	2,400
number	Arith.* (Wu et al., 2024)	two numbers	the sum in certain base	1,000
symbol	Le/Ho. (Liu et al., 2024b)	pairs of triplets	an entailment rule	2,000
structure	NutFrame (Guo et al., 2024)	some frame information	conceptual structures	30,000
fact	DEER (Yang et al., 2024)	a pair of facts	a rule covers the facts	200
puzzle	RULEARN (He et al., 2025)	some puzzle scenarios	a puzzle rule	300
word	Crypto.* (Li et al., 2025a)	pairs of english words	an encrypted rule	300
symbol	GeoILP (Chen et al., 2025c)	pos. and neg. samples	a logic rule	250,000
string	In.Bench (Hua et al., 2025)	a pair of strings	a string-mapping rule	1,000
number	CodeSeq (Chen et al., 2025a)	a number sequence	the general term	200

rules. Particularly, Codeseq (Chen et al., 2025a) involves the computation of number sequence general terms, which represents a more advanced and complex form of inductive reasoning.

Overall, these benchmarks test the models' pattern recognition ability and impose rigorous challenges on their higher-order cognitive skills. They not only examine how effectively LLMs can generalize from limited observations to underlying rules, but also serve as a foundation for driving further progress in enhancing such abilities.

4.2 Evaluation Approaches

In this section, we introduce the evaluation approaches for the inductive reasoning benchmarks mentioned above. We first present the existing evaluation strategies used in the benchmark papers (Section 4.2.1). Then, we derive a sandbox-based evaluation approach with a fine-grained observation coverage metric built upon it (Section 4.2.2).

4.2.1 Existing Evaluation Strategy

Most of the benchmarks in Section 4.1 and current works directly evaluate the consistency between answers generated by LLMs and the ground truth. Therefore, general metrics are employed, such as ACC, exact match, success rate, and so on. For example, ACRE (Zhang et al., 2021a) selects the most plausible "Blicket" from multiple options to evaluate the accuracy of its selection. SCAN (Lake and Baroni, 2018) focuses on assessing whether the generated outputs exactly match the reference answers to indicate accuracy. SyGuS (Odena et al., 2021) requires finding a program that satisfies the string transformation rule, and the number of tasks in which the correct program is successfully identified is counted as the success rate.

4.2.2 Sandbox-based Evaluation

Considering that all inductive reasoning tasks share the same intrinsic mechanism: inferring general rules from specific observations. We can adopt a unified approach, namely sandbox unit test (Appendix A.7.1), to standardize the evaluation across all the inductive benchmarks mentioned above.

The sandbox unit test is a method where individual components or functions are tested in isolation to ensure they work as intended (Alhindi and Hallett, 2025). Each test runs in a controlled, independent environment, using specific input cases to verify the correctness of the component. This approach helps identify errors early and ensures that each part functions correctly before integration.

The sandbox unit test is originally used for code verification, at which time it is referred to as a code unit test (Gong et al., 2025). With the development of LLMs, it is also applied to evaluate various deductive reasoning and agent-related tasks of LLMs,

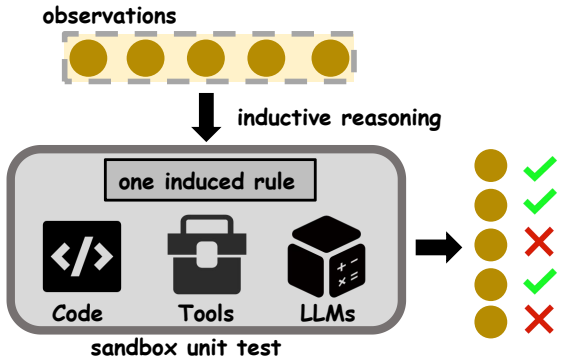


Figure 6: A demo of the sandbox unit test for inductive reasoning of LLMs.

such as InternBootCamp (Li et al., 2025c).

In inductive reasoning tasks, the sandbox unit test can also be deployed for evaluation. We present a demo in Figure 6. For an inductive rule generated by an LLM, it can be encapsulated as code (Gong et al., 2025), a tool (Qu et al., 2025), or written into a prompt to be provided to the LLM-as-a-Judge (Gu et al., 2024). Each observation can then be executed in a sandbox environment to determine whether it conforms to the current rule.

Based on this, we can derive a more fine-grained metric for LLM inductive reasoning: **observation coverage** (OC) (Appendix A.7.2), defined as the proportion of observations that pass the unit tests out of the total number of observations. In the example shown in Figure 6, this value is 0.6. Compared with the overall ACC or success rate of a task, OC provides a more fine-grained supervision signal at the observation level. This allows for a more precise reflection of the comprehensiveness of the model's answer. With this metric, more informative feedback can be provided for subsequent rule refinement and hypothesis exploration.

5 Analysis

In this section, we present several prior exploratory tasks that offer theoretical analyses for inductive reasoning and inductive bias of LLMs (Kharitonov and Chaabouni, 2021; Papadimitriou and Jurafsky, 2023; Wilson and Frank, 2023).

Inductive ability originates from induction heads. Many studies (Si et al., 2023; Edelman et al., 2024; Chen et al., 2024d) show that the strong in-context learning (ICL) (Dong et al., 2023, 2024; Crosbie and Shutova, 2025) or example imitation (Honovich et al., 2023; Ye et al., 2025) ability of LLMs originates from induction heads. An induction head is an attention head (Edelman et al., 2022; Ren et al., 2024) that performs a match-and-copy operation, identifying and replicating relevant context tokens (Singh et al., 2024). Minegishi et al. (2025) finds that, in fact, induction heads are meta-learning an abstract inductive within the context.

Model parameters, architecture, and data all help shape the inductive bias. The parameters, model architecture, and training data (White and Cotterell, 2021; Merrill et al., 2021; Lovering et al., 2021; Levine et al., 2022; HaoChen and Ma, 2023; Movahedi et al., 2025) are key to inductive bias. Lippl and Lindsey (2024) explores the effects of different parameters on inductive bias under multi-data mixed training and single-task finetuning scenarios, and ultimately emphasizes the importance of task similarity in mixed training. Some studies (Cabannes et al., 2023; Aerni et al., 2023) also highlight the importance of data augmentation, even the noisy data. Further research (Zeno et al., 2025) demonstrates that the choice of minimum norm can also determine a model's inductive generalization.

Induction means simplicity. Some early studies show that complex model architectures and data (Zietlow et al., 2021) can actually hinder inductive generalization. At the same time, for higher-order models, regularization can actually be detrimental to the formation of their inductive bias (Phuong and Lampert, 2021; Donhauser et al., 2022). Sometimes, simplicity is perfect for inductive reasoning (Goldblum et al., 2024). To enhance the inductive reasoning ability, finding simple inductive bias is of paramount importance. Simple and pure corpora often serve as the foundation for successful inductive reasoning (Mueller and Linzen, 2023).

6 Conclusion

This is the first survey of inductive reasoning for LLMs. The inductive mode is crucial for knowledge generalization and aligns better with human cognition. We categorize methods for improving inductive reasoning into three areas: post-training enhancement, test-time exploration, and data augmentation. We also summarize the current benchmarks and derive a unified sandbox-based evaluation approach. Finally, we offer some analyses regarding the source of inductive ability and how simple model architectures and data help with inductive tasks, providing a solid foundation for future research (Appendix A.8).

Limitations

This paper is a survey about the inductive reasoning abilities of Large Language Models. Due to space limitations, the main body of this survey is constrained to fewer than eight pages, and therefore, many details are not included in the main text. We only present the most essential parts. Meanwhile, although inductive reasoning in LLMs attracts increasing attention in recent years, the number of related studies remains relatively limited, making it difficult to produce a large-scale, systematic survey (even extending to 100 pages) comparable to those in other areas.

Ethics Statements

This survey primarily organizes and summarizes existing work on inductive reasoning in LLMs, with all relevant sources properly cited. Therefore, this paper does not raise any ethical or moral concerns.

Acknowledgments

We will finish this part in the camera-ready version.

Use of AI Assistants

We primarily use AI assistants to improve and enrich our writing, especially by leveraging LLMs to help us write taxonomy in LaTeX.

References

Michael Aerni, Marco Milanta, Konstantin Donhauser, and Fanny Yang. 2023. Strong inductive biases provably prevent harmless interpolation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024: Student Research Workshop, St. Julian's, Malta, March 21-22, 2024*, pages 225–237. Association for Computational Linguistics.

Taha Aksu, Devamanyu Hazarika, Shikib Mehri, Seokhwan Kim, Dilek Hakkani-Tür, Yang Liu, and Mahdi Namazifar. 2023. Cesar: Automatic induction of compositional instructions for multi-turn dialogs. *Preprint*, arXiv:2311.17376.

Ferran Alet, Javier Lopez-Contreras, James Koppel, Maxwell I. Nye, Armando Solar-Lezama, Tomás Lozano-Pérez, Leslie Pack Kaelbling, and Joshua B. Tenenbaum. 2021. A large-scale benchmark for few-shot program induction and synthesis. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 175–186. PMLR.

Maysara Alhindi and Joseph Hallett. 2025. Playing in the sandbox: A study on the usability of seccomp. In *Twenty-First Symposium on Usable Privacy and Security, SOUPS 2025, Seattle, WA, USA, August 10-12, 2025*, pages 225–240. USENIX Association.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona T. Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *CoRR*, abs/2204.06031.

Nadia Alshahwan, Jubin Chheda, Anastasia Finogenova, Beliz Gokkaya, Mark Harman, Inna Harper, Alexandru Marginean, Shubho Sengupta, and Eddy Wang. 2024. Automated unit test improvement using large language models at meta. *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*.

Ohad Amosy, Tomer Volk, Eilam Shapira, Eyal Ben-David, Roi Reichart, and Gal Chechik. 2024. Text2model: Text-based model induction for zero-shot image classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 155–172. Association for Computational Linguistics.

Saurabh Arora and Prashant Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500.

W. Brian Arthur. 1994. Inductive reasoning and bounded rationality. *The American Economic Review*, 84:406–411.

Yuyang Bai, Shangbin Feng, Vidhisha Balachandran, Zhaoxuan Tan, Shiqi Lou, Tianxing He, and Yulia Tsvetkov. 2024. Kgquiz: Evaluating the generalization of encoded knowledge in large language models. In *Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, May 13-17, 2024*, pages 2226–2237. ACM.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *ArXiv*, abs/2302.04023.

Howard Barnum. 2012. The beginning of infinity: Explanations that transform the world. *Physics Today*, 65:48–50.

André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. 2024. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524*.

Jonathan Baxter. 2000. A model of inductive bias learning. *ArXiv*, abs/1106.0245.

Mark Bun, Gautam Kamath, Thomas Steinke, and Steven Z Wu. 2019. Private hypothesis selection. *Advances in Neural Information Processing Systems*, 32.

Vivien Cabannes, Bobak Toussi Kiani, Randall Balestriero, Yann LeCun, and Alberto Bietti. 2023. The SSL interplay: Augmentations, inductive bias, and generalization. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 3252–3298. PMLR.

Xuan Cai, Xuesong Bai, Zhiyong Cui, Danmu Xie, Daocheng Fu, Haiyang Yu, and Yilong Ren. 2025. Text2scenario: Text-driven scenario generation for autonomous driving test. *Preprint*, arXiv:2503.02911.

Yuang Cai, Yuyu Yuan, Jinsheng Shi, and Qinhong Lin. 2024. Approximated variational bayesian inverse reinforcement learning for large language model alignment. *Preprint*, arXiv:2411.09341.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4862–4872. Association for Computational Linguistics.

Colin L. Carter and Howard J. Hamilton. 1998. Efficient attribute-oriented generalization for knowledge discovery from large databases. *IEEE Trans. Knowl. Data Eng.*, 10(2):193–208.

Rich Caruana. 1993. Multitask learning: A knowledge-based source of inductive bias. In *International Conference on Machine Learning*.

Bowen Chen, Rune Sætre, and Yusuke Miyao. 2024a. A comprehensive evaluation of inductive reasoning capabilities and problem solving in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 323–339. Association for Computational Linguistics.

Kedi Chen, Qin Chen, Jie Zhou, Yishen He, and Liang He. 2024b. Diahalu: A dialogue-level hallucination evaluation benchmark for large language models. *CoRR*, abs/2403.00896.

Kedi Chen, Zhikai Lei, Fan Zhang, Yinqi Zhang, Qin Chen, Jie Zhou, Liang He, Qipeng Guo, Kai Chen, and Wei Zhang. 2025a. Code-driven inductive synthesis: Enhancing reasoning abilities of large language models with sequences. *Preprint*, arXiv:2503.13109.

Po-Chun Chen, Sheng-Lun Wei, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024c. Induct-learn: Short phrase prompting with instruction induction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5204–5231. Association for Computational Linguistics.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025b. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Preprint*, arXiv:2503.09567.

Si Chen, Richong Zhang, and Xu Zhang. 2025c. Geoilp: A synthetic dataset to guide large-scale rule induction. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. 2024d. Unveiling induction heads: Provable training dynamics and feature learning in transformers. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ziyang Chen, Dongfang Li, Xiang Zhao, Baotian Hu, and Min Zhang. 2024e. Temporal knowledge question answering via abstract reasoning induction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 4872–4889. Association for Computational Linguistics.

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. 2024a. Inductive or deductive? rethinking the fundamental reasoning abilities of llms. *CoRR*, abs/2408.00114.

Sitao Cheng, Liangming Pan, Xunjian Yin, Xinyi Wang, and William Yang Wang. 2024b. Understanding the interplay between parametric and contextual knowledge for large language models. *Preprint*, arXiv:2410.08414.

François Chollet. 2019. On the measure of intelligence. *CoRR*, abs/1911.01547.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S. Yu, and Qingsong Wen. 2025. LLM agents for education: Advances and applications. *CoRR*, abs/2503.11733.

Herbert H. Clark. 1969. Linguistic processes in deductive reasoning. *Psychological Review*, 76:387–404.

Irving M. Copi, Carl Cohen, and Daniel E. Flage. 2004. Essentials of logic.

Joy Crosbie and Ekaterina Shutova. 2025. Induction heads as an essential mechanism for pattern matching in in-context learning. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5034–5096. Association for Computational Linguistics.

Paul Darm, Antonio Valerio Miceli-Barone, Shay B. Cohen, and Annalisa Riccardi. 2023. Knowledge base question answering for space debris queries. *Preprint*, arXiv:2305.19734.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Rushikesh Dhudum, Ankit Ganeshpurkar, and Atmaram Pawar. 2024. Revolutionizing drug discovery: a comprehensive review of ai applications. *Drugs and Drug Candidates*, 3(1):148–171.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1107–1128. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey for in-context learning. *CoRR*, abs/2301.00234.

Konstantin Donhauser, Nicolò Ruggeri, Stefan Stojanovic, and Fanny Yang. 2022. Fast rates for noisy interpolation require rethinking the effects of inductive bias. *CoRR*, abs/2203.03597.

Rotem Dror, Haoyu Wang, and Dan Roth. 2023. Zero-shot on-the-fly event schema induction. In *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 693–713. Association for Computational Linguistics.

Benjamin L. Edelman, Surbhi Goel, Sham M. Kakade, and Cyril Zhang. 2022. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5793–5831. PMLR.

Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, Eran Malach, and Surbhi Goel. 2024. The evolution of statistical induction heads: In-context learning markov chains. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Carl Edwards and Heng Ji. 2023. Semi-supervised new event type induction and description via contrastive loss-enforced batch attention. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 3787–3809. Association for Computational Linguistics.

Richard Evans and Edward Grefenstette. 2018. Learning explanatory rules from noisy data. *Preprint*, arXiv:1711.04574.

Matan Eyal, Shoval Sadde, Hillel Taub-Tabib, and Yoav Goldberg. 2022. Large scale substitution-based word sense induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4738–4752, Dublin, Ireland. Association for Computational Linguistics.

Muhammad Ashraf Faheem. 2021. Ai-driven risk assessment models: Revolutionizing credit scoring and default prediction. *Iconic Research And Engineering Journals*, 5(3):177–186.

Shaoxiong Feng, Xuancheng Ren, Kan Li, and Xu Sun. 2022. Hierarchical inductive transfer for continual dialogue learning. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 693–699. Association for Computational Linguistics.

Dmitry V. Galkin. 2011. The hypothesis of interactive evolution. *Kybernetes*, 40(7):1021–1029.

Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Jerry Chun-Wei Lin. 2023. Large language models in education: Vision and opportunities. *CoRR*, abs/2311.13160.

Yolanda Gil, Daniel Garijo, Varun Ratnakar, Rajiv Mayani, Ravali Adusumilli, Hunter Boyce, Arunima Srivastava, and Parag Mallick. 2017. Towards continuous scientific data analysis and hypothesis evolution. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4406–4414. AAAI Press.

Claire Glanois, Zhaohui Jiang, Xuening Feng, Paul Weng, Matthieu Zimmer, Dong Li, Wulong Liu, and Jianye Hao. 2022. Neuro-symbolic hierarchical rule induction. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 7583–7615. PMLR.

Anubha Goel, Puneet Pasricha, Martin Magris, and Juho Kanniainen. 2025. Foundation time-series ai model for realized volatility forecasting. *Preprint*, arXiv:2505.11163.

Micah Goldblum, Marc Anton Finzi, Keefer Rowan, and Andrew Gordon Wilson. 2024. Position: The no free lunch theorem, kolmogorov complexity, and the role of inductive biases in machine learning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Jingzhi Gong, Vardan Voskanyan, Paul Brookes, Fan Wu, Wei Jie, Jie Xu, Rafail Giavrimis, Mike Basios, Leslie Kanthan, and Zheng Wang. 2025. Language models for code optimization: Survey, challenges and future directions. *CoRR*, abs/2501.01277.

Anirudh Goyal and Yoshua Bengio. 2020. Inductive biases for deep learning of higher-level cognition. *CoRR*, abs/2011.15091.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594.

Yu Gu and Yu Su. 2022. Arcaneqa: Dynamic program induction and contextualized encoding for knowledge base question answering. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 1718–1731. International Committee on Computational Linguistics.

Shaoru Guo, Yubo Chen, Kang Liu, Ru Li, and Jun Zhao. 2024. Nutframe: Frame-based conceptual structure induction with llms. In *LREC/COLING*, pages 12330–12335.

Simon Jerome Han, Keith J. Ransom, Andrew Perfors, and Charles Kemp. 2024. Inductive reasoning in humans and large language models. *Cogn. Syst. Res.*, 83:101155.

Jeff Z. HaoChen and Tengyu Ma. 2023. A theoretical study of inductive biases in contrastive learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Kaiyu He and Zhiyu Chen. 2025. From reasoning to learning: A survey on hypothesis discovery and rule learning with large language models. *Preprint*, arXiv:2505.21935.

Kaiyu He, Mian Zhang, Shuo Yan, Peilin Wu, and Zhiyu Zoey Chen. 2025. Idea: Enhancing the rule learning ability of large language model agent through induction, deduction, and abduction. *Preprint*, arXiv:2408.10455.

Evan Heit. 2000. Properties of inductive reasoning. *Psychonomic Bulletin & Review*, 7:569–592.

Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. 2022. Instruction induction: From few examples to natural language task descriptions. *Preprint*, arXiv:2205.10782.

Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. 2023. Instruction induction: From few examples to natural language task descriptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1935–1952. Association for Computational Linguistics.

Wenyue Hua, Tyler Wong, Fei Sun, Liangming Pan, Adam Jardine, and William Yang Wang. 2025. Inductionbench: Llms fail in the simplest complexity class. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 26526–26546. Association for Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Lifu Huang and Heng Ji. 2020. Semi-supervised new event type induction and event detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 718–724, Online. Association for Computational Linguistics.

Alexander Immer, Lucas Torroba Hennigen, Vincent Fortuin, and Ryan Cotterell. 2022. Probing as quantifying inductive bias. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1839–1851. Association for Computational Linguistics.

Tim Ingold. 2021. The perception of the environment: essays on livelihood, dwelling and skill.

InternTeam. 2025. Intern-s1: A scientific multimodal foundation model. *CoRR*, abs/2508.15763.

Shuoran Jiang, Qingcai Chen, Yang Xiang, Youcheng Pan, and Yukang Lin. 2024a. Linguistic rule induction improves adversarial and OOD robustness in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10565–10577, Torino, Italia. ELRA and ICCL.

Xue Jiang, Yihong Dong, Lecheng Wang, Zheng Fang, Qiwei Shang, Ge Li, Zhi Jin, and Wenpin Jiao. 2024b. Self-planning code generation with large language models. *ACM Trans. Softw. Eng. Methodol.*, 33(7):182:1–182:30.

Yanru Jiang, Siyu Liang, and Junwon Choi. 2025. Synthetic survey data generation and evaluation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.1, KDD 2025, Toronto, ON, Canada, August 3-7, 2025*, pages 2292–2302. ACM.

Can Jin, Yang Zhou, Qixin Zhang, Hongwu Peng, Di Zhang, Marco Pavone, Ligong Han, Zhang-Wei Hong, Tong Che, and Dimitris N. Metaxas. 2025. Your reward function for rl is your best prm for search: Unifying rl and search-based tts. *Preprint*, arXiv:2508.14313.

Davor Juretic. 2025. Exploring the evolution-coupling hypothesis: Do enzymes' performance gains correlate with increased dissipation? *Entropy*, 27(4):365.

Jushi Kai, Shengyuan Hou, Yusheng Huang, and Zhouhan Lin. 2024. Leveraging grammar induction for language understanding and generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 4501–4513. Association for Computational Linguistics.

Charles W. Kalish and Jordan Thevenow-Harrison. 2014. Descriptive and inferential problems of induction.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. AMMUS : A survey of transformer-based pretrained models in natural language processing. *CoRR*, abs/2108.05542.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. 2024. A survey of reinforcement learning from human feedback.

Ruhma Khan, Sumit Gulwani, Vu Le, Arjun Radhakrishna, Ashish Tiwari, and Gust Verbruggen. 2025. Llm-guided compositional program synthesis. *CoRR*, abs/2503.15540.

Eugene Kharitonov and Rahma Chaabouni. 2021. What they do when in doubt: a study of inductive biases in seq2seq learners. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jeonghwan Kim, Giwon Hong, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2023. Fineprompt: Unveiling the role of finetuned inductive bias on compositional reasoning in GPT-4. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3763–3775. Association for Computational Linguistics.

Taeuk Kim, Jihun Choi, Daniel Edmiston, and Sang-goo Lee. 2020. Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Harsh Kohli, Helian Feng, Nicholas Dronen, Calvin McCarter, Sina Moeini, and Ali Kebarighotbi. 2024. How lexical is bilingual lexicon induction? In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4381–4386. Association for Computational Linguistics.

James Kotary, Ferdinando Fioretto, Pascal Van Hentenryck, and Bryan Wilder. 2021. End-to-end constrained optimization learning: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4475–4482. ijcai.org.

Shambhavi Krishna and Aishwarya Sahoo. 2024. Solving the inverse alignment problem for efficient rlhf. *Preprint*, arXiv:2412.10529.

Hanyu Lai, Xiao Liu, Junjie Gao, Jiale Cheng, Zehan Qi, Yifan Xu, Shuntian Yao, Dan Zhang, Jinhua Du, Zhenyu Hou, Xin Lv, Minlie Huang, Yuxiao Dong, and Jie Tang. 2025. A survey of post-training scaling in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 2771–2791. Association for Computational Linguistics.

Viet Dac Lai, Hieu Man, Linh Ngo Van, Franck Dernoncourt, and Thien Huu Nguyen. 2022. Multilingual subevent relation extraction: A novel dataset and structure induction method. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5559–5570. Association for Computational Linguistics.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *Preprint*, arXiv:1711.00350.

Kang-il Lee, Hyukhun Koh, Dongryeol Lee, Seunghyun Yoon, Minsung Kim, and Kyomin Jung. 2025a. Generating diverse hypotheses for inductive reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 8461–8474. Association for Computational Linguistics.

Kangil Lee, Jahyun Koo, Seunghyun Yoon, Minbeom Kim, Hyukhun Koh, Dongryeol Lee, and Kyomin Jung. 2025b. Program synthesis via test-time transduction. *CoRR*, abs/2509.17393.

Yoav Levine, Noam Wies, Daniel Jannai, Dan Navon, Yedid Hoshen, and Amnon Shashua. 2022. The inductive bias of in-context learning: Rethinking pretraining example design. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Martha Lewis and Melanie Mitchell. 2024a. Evaluating the robustness of analogical reasoning in large language models. *arXiv preprint arXiv:2411.14215*.

Martha Lewis and Melanie Mitchell. 2024b. Using counterfactual tasks to evaluate the generality of analogical reasoning in large language models. *arXiv preprint arXiv:2402.08955*.

Boyi Li, Rodolfo Corona, Karttikeya Mangalam, Catherine Chen, Daniel Flaherty, Serge Belongie, Kilian Q. Weinberger, Jitendra Malik, Trevor Darrell, and Dan Klein. 2024a. Re-evaluating the need for multimodal signals in unsupervised grammar induction. *Preprint*, arXiv:2212.10564.

Chunyang Li, Weiqi Wang, Tianshi Zheng, and Yangqiu Song. 2025a. Patterns over principles: The fragility of inductive reasoning in llms under noisy observations. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 19608–19626. Association for Computational Linguistics.

Jiachun Li, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. 2025b. MIRAGE: evaluating and explaining inductive reasoning process in language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Peiji Li, Jiasheng Ye, Yongkang Chen, Yichuan Ma, Zijie Yu, Kedi Chen, Ganqu Cui, Haozhan Li, Jiacheng Chen, Chengqi Lyu, Wenwei Zhang, Linyang Li, Qipeng Guo, Dahua Lin, Bowen Zhou, and Kai Chen. 2025c. Internbootcamp technical report: Boosting LLM reasoning with verifiable task scaling. *CoRR*, abs/2508.08636.

Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch, and Jiawei Han. 2023a. Open-domain hierarchical event schema induction by incremental prompting and verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5677–5697, Toronto, Canada. Association for Computational Linguistics.

Siyue Li, Xiaofan Zhou, Zhizhong Wu, Yuiian Long, and Yanxin Shen. 2024b. Strategic deductive reasoning in large language models: A dual-agent approach. In *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, pages 834–839. IEEE.

Wen-Ding Li and Kevin Ellis. 2024. Is programming by example solved by llms? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yaoyiran Li, Anna Korhonen, and Ivan Vulic. 2023b. On bilingual lexicon induction with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9577–9599. Association for Computational Linguistics.

Matthias Lindemann, Alexander Koller, and Ivan Titov. 2024. Strengthening structural inductive biases by pre-training to perform syntactic transformations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 11558–11573. Association for Computational Linguistics.

Samuel Lippl and Jack W. Lindsey. 2024. Inductive biases of multi-task learning and finetuning: multiple regimes of feature reuse. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Jianyu Liu, Sheng Bi, and Guilin Qi. 2024a. Primo: Progressive induction for multi-hop open rule generation. *Preprint*, arXiv:2411.01205.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tennison Liu, Nicolas Huynh, and Mihaela van der Schaar. 2025. Decision tree induction through llms via semantically-aware evolution. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Tianyang Liu, Tianyi Li, Liang Cheng, and Mark Steedman. 2024b. Explicit inductive inference using large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 15779–15786. Association for Computational Linguistics.

Tianyu Liu, Qitan Lv, Jie Wang, Shuling Yang, and Hanzhu Chen. 2024c. Learning rule-induced subgraph representations for inductive relation prediction. *CoRR*, abs/2408.07088.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *CoRR*, abs/2406.15126.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. Predicting inductive biases of pre-trained models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zhou Lu. 2024. When is inductive inference possible? In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024. Mathcoder2: Better math reasoning from continued pretraining on model-translated mathematical code. *CoRR*, abs/2410.08196.

Yaswanth M, Vaibhav Singh, Ayush Maheshwari, Amrith Krishna, and Ganesh Ramakrishnan. 2025. ARISE: iterative rule induction and synthetic data generation for text classification. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 6419–6434. Association for Computational Linguistics.

Ali Maatouk, Nicola Piovesan, Fadhel Ayed, Antonio De Domenico, and Mérouane Debbah. 2023. Large language models for telecom: Forthcoming impact on the industry. *CoRR*, abs/2308.06013.

Monnie McGee and Bivin Sadler. 2025. Generative AI takes a statistics exam: A comparison of performance between chatgpt3.5, chatgpt4, and chatgpt4o-mini. *CoRR*, abs/2501.09171.

Matteo Merler, Katsiaryna Haitsiukevich, Nicola Dainese, and Pekka Marttinen. 2024. In-context symbolic regression: Leveraging large language models for function discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024 - Student Research Workshop, Bangkok, Thailand, August 11-16, 2024*, pages 589–606. Association for Computational Linguistics.

William Merrill, Vivek Ramanujan, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Effects of parameter norm growth during transformer training: Inductive bias from gradient descent. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1766–1781. Association for Computational Linguistics.

Ivan Milev, Mislav Balunovic, Maximilian Baader, and Martin T. Vechev. 2025. Toolfuzz - automated agent tool testing. *CoRR*, abs/2503.04479.

Gouki Minegishi, Hiroki Furuta, Shohei Taniguchi, Yusuke Iwasawa, and Yutaka Matsuo. 2025. Beyond induction heads: In-context meta learning induces multi-phase circuit emergence. *CoRR*, abs/2505.16694.

Anna Mosolova, Marie Candito, and Carlos Ramisch. 2025. In the LLM era, word sense induction remains unsolved. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17161–17178, Vienna, Austria. Association for Computational Linguistics.

Sajad Movahedi, Antonio Orvieto, and Seyed-Mohsen Moosavi-Dezfooli. 2025. Geometric inductive biases of deep networks: The role of data and architecture. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Aaron Mueller and Tal Linzen. 2023. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11237–11252. Association for Computational Linguistics.

Sam Musker, Alex Duchnowski, Raphaël Millière, and Ellie Pavlick. 2024. Semantic structure-mapping in llm and human analogical reasoning. *arXiv e-prints*, pages arXiv–2406.

Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13:134615–134633.

Atharva Naik, Darsh Agrawal, Hong Sng, Clayton Marr, Kexun Zhang, Nathaniel R. Robinson, Kalvin Chang, Rebecca Byrnes, Aravind Mysore, Carolyn P. Rosé, and David R. Mortensen. 2025. Programming by examples meets historical linguistics: A large language model based approach to sound law induction. *CoRR*, abs/2501.16524.

Augustus Odena, Kensen Shi, David Bieber, Rishabh Singh, Charles Sutton, and Hanjun Dai. 2021. BUSTLE: bottom-up program synthesis through learning-guided exploration. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jiao Ou, Jiayu Wu, Che Liu, Fuzheng Zhang, Di Zhang, and Kun Gai. 2024. Inductive-deductive strategy reuse for multi-turn instructional dialogues. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 17402–17431. Association for Computational Linguistics.

Isabel Papadimitriou and Dan Jurafsky. 2023. Injecting structural hints: Using language models to study inductive biases in language learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 8402–8413. Association for Computational Linguistics.

Angelina Parfenova and Jürgen Pfeffer. 2025. Measuring what matters: Evaluating ensemble llms with label refinement in inductive coding. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 10803–10816. Association for Computational Linguistics.

Michael J Pazzani and Glenn Silverstein. 1990. Feature selection and hypothesis selection models of induction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 12.

Mary Phuong and Christoph H. Lampert. 2021. The inductive bias of relu networks on orthogonally separable data. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Bäck. 2024. Reasoning with large language models, a survey. *CoRR*, abs/2407.11511.

Foster J. Provost and Bruce G. Buchanan. 1995. Inductive policy: The pragmatics of bias selection. *Machine Learning*, 20:35–61.

Jing Qian, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. 2023. Limitations of language models in arithmetic and symbolic induction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9285–9298. Association for Computational Linguistics.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. 2024. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *Preprint*, arXiv:2310.08559.

Changle Qu, Sunhao Dai, Xiaochi Wei, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Jun Xu, and Ji-Rong Wen. 2025. Tool learning with large language models: a survey. *Frontiers Comput. Sci.*, 19(8):198343.

QwenTeam. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Karthik Radhakrishnan, Tushar Kanakagiri, Sharanya Chakravarthy, and Vidhisha Balachandran. 2020. "a little birdie told me ... " - inductive biases for rumour stance detection on social media. In *Proceedings of the Sixth Workshop on Noisy User-generated Text, W-NUT@EMNLP 2020 Online, November 19, 2020*, pages 244–248. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Raghav Ramji and Keshav Ramji. 2025. Inductive linguistic reasoning with large language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 22783–22810. Association for Computational Linguistics.

Chandan K. Reddy and Parshin Shojaee. 2025. Towards scientific discovery with generative AI: progress, opportunities, and challenges. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 28601–28609. AAAI Press.

Jie Ren, Qipeng Guo, Hang Yan, Dongrui Liu, Quanshi Zhang, Xipeng Qiu, and Dahua Lin. 2024. Identifying semantic induction heads to understand in-context learning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6916–6932. Association for Computational Linguistics.

Lance J. Rips. 1994. The psychology of proof: Deductive reasoning in human thinking.

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

Dongwon Ryu, Ehsan Shareghi, Meng Fang, Yunqiu Xu, Shirui Pan, and Gholamreza Haffari. 2022. Fire burns, sword cuts: Commonsense inductive bias for exploration in text-based games. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 515–522. Association for Computational Linguistics.

Shuaijie She, Shujian Huang, Xingyun Wang, Yanke Zhou, and Jiajun Chen. 2023. Exploring the dialogue comprehension ability of large language models. *CoRR*, abs/2311.07194.

Thomas J. Sheffler. 2025. An approach to checking correctness for agentic systems. *CoRR*, abs/2509.20364.

Haoyue Shi, Luke Zettlemoyer, and Sida I. Wang. 2021. Bilingual lexicon induction via unsupervised bitext construction and word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 813–826. Association for Computational Linguistics.

Amitis Shidani, Tyler Farghly, Yang Sun, Habib Ganjgahi, and George Deligiannidis. 2025. Beyond real data: Synthetic data through the lens of regularization. *CoRR*, abs/2510.08095.

Parshin Shojaee, Kazem Meidani, Shashank Gupta, Amir Barati Farimani, and Chandan K. Reddy. 2025. LLM-SR: scientific equation discovery via programming with large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023. Measuring inductive biases of in-context learning with underspecified demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11289–11310. Association for Computational Linguistics.

Gabriel Silva, Mário Rodriges, António Teixeira, and Marlene Amorim. 2025. Inductive learning on heterogeneous graphs enhanced by LLMs for software mention detection. In *Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025)*,

pages 164–172, Vienna, Austria. Association for Computational Linguistics.

Aaditya K. Singh, Ted Moskovitz, Felix Hill, Stephanie C. Y. Chan, and Andrew M. Saxe. 2024. What needs to go right for an induction head? A mechanistic study of in-context learning circuits and their formation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Dominik Stempień and Robert Ślepaczuk. 2025. Hybrid models for financial forecasting: Combining econometric, machine learning, and deep learning models. *Preprint*, arXiv:2505.19617.

Hao Sun, Alihan Hüyük, and Mihaela van der Schaar. 2024a. Query-dependent prompt evaluation and optimization with offline inverse rl. *Preprint*, arXiv:2309.06553.

Hao Sun and Mihaela van der Schaar. 2025. Inverse reinforcement learning meets large language model post-training: Basics, advances, and opportunities. *Preprint*, arXiv:2507.13158.

Wangtao Sun, Haotian Xu, Xuanqing Yu, Pei Chen, Shizhu He, Jun Zhao, and Kang Liu. 2024b. Itd: Large language models can teach themselves induction through deduction. *Preprint*, arXiv:2403.05789.

Yizhou Sun, Flavio Chierichetti, Hady W. Lauw, Claudia Perlich, Wee Hyong Tok, and Andrew Tomkins, editors. 2025. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.1, KDD 2025, Toronto, ON, Canada, August 3-7, 2025*. ACM.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. 2024. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*.

David R Thomas. 2003. A general inductive approach for qualitative data analysis.

Marcin Trepczynski. 2024. Religija, teologija i filozofske vjetine automatiziranih programa za avrljanje (chatbotova) pogonjenima velikim jezinim modelima (llm). *Disputatio philosophica*.

Shogo Tsujimoto, Kosuke Yamada, and Ryohei Sasano. 2025. Semantic frame induction from a real-world corpus. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 991–997, Vienna, Austria. Association for Computational Linguistics.

Tao Tu, Anil Palepu, Mike Schaekermann, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Yong Cheng, Le Hou, Albert Webson, Kavita Kulkarni, S Sara Mahdavi, Christopher Semturs, Juraj Gottweis, and 6 others. 2024. Towards conversational diagnostic ai. *Preprint*, arXiv:2401.05654.

Takane Ueno, Yuto Sawa, Yeongdae Kim, Jacqueline Urakami, Hiroki Oura, and Katie Seaborn. 2022. Trust in human-ai interaction: Scoping out models, measures, and methods. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*.

Joaquin Vanschoren. 2018. Meta-learning: A survey. *CoRR*, abs/1810.03548.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Gust Verbruggen, Ashish Tiwari, Mukul Singh, Vu Le, and Sumit Gulwani. 2025. Execution-guided within-prompt search for programming-by-example. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Ivo Verhoeven, Pushkar Mishra, Rahel Beloch, Helen Yannakoudakis, and Ekaterina Shutova. 2024. A (more) realistic evaluation setup for generalisation of community models on malicious content detection. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 437–463. Association for Computational Linguistics.

Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2024a. Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Runxiang Wang, Boxiao Wang, Kai Li, Yifan Zhang, and Jian Cheng. 2025a. Drsr: LLM based scientific equation discovery with dual reasoning from data and experience. *CoRR*, abs/2506.04282.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D. Goodman. 2024b. Hypothesis search: Inductive reasoning with language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yuping Wang, Shuo Xing, Cui Can, Renjie Li, Hongyuan Hua, Kexin Tian, Zhaobin Mo, Xiangbo Gao, Keshu Wu, Sulong Zhou, Hengxu You, Juntong Peng, Junge Zhang, Zehao Wang, Rui Song, Mingxuan Yan, Walter Zimmer, Xingcheng Zhou, Peiran Li, and 28 others. 2025b. Generative ai for autonomous driving: Frontiers and opportunities. *Preprint*, arXiv:2505.08854.

Anjiang Wei, Tarun Suresh, Jiannan Cao, Naveen Kannan, Yuheng Wu, Kai Yan, Thiago S. F. X. Teixeira, Ke Wang, and Alex Aiken. 2025. Codearc: Benchmarking reasoning capabilities of LLM agents for inductive program synthesis. *CoRR*, abs/2503.23145.

Xiaokai Wei, Shen Wang, Dejiao Zhang, Parminder Bhatia, and Andrew O. Arnold. 2021. Knowledge enhanced pretrained language models: A compreshensive survey. *CoRR*, abs/2110.08455.

Jennifer C. White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 454–463. Association for Computational Linguistics.

Michael Wilson and Robert Frank. 2023. Inductive bias is in the eye of the beholder. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 152–162, Singapore. Association for Computational Linguistics.

Xiaobao Wu. 2025. Sailing AI by the stars: A survey of learning from rewards in post-training and test-time scaling of large language models. *CoRR*, abs/2505.02686.

Yuhuai Wu, Markus Rabe, Wenda Li, Jimmy Ba, Roger Grosse, and Christian Szegedy. 2022. Lime: Learning inductive bias for primitives of mathematical reasoning. *Preprint*, arXiv:2101.06223.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 1819–1862. Association for Computational Linguistics.

Zhiwen Xie, Yi Zhang, Jin Liu, Guangyou Zhou, and Jimmy Xiangji Huang. 2023. Learning query adaptive anchor representation for inductive relation prediction. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 14041–14053. Association for Computational Linguistics.

Zhouhang Xie, Bodhisattwa Prasad Majumder, Mengjie Zhao, Yoshinori Maeda, Keiichi Yamada, Hiromi Wakaki, and Julian J. McAuley. 2024. Few-shot dialogue strategy learning for motivational interviewing via inductive reasoning. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 13207–13219. Association for Computational Linguistics.

Biao Xu and Guanci Yang. 2025. Interpretability research of deep learning: A literature survey. *Inf. Fusion*, 115:102721.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. 2023. Large language models for generative information extraction: A survey. *CoRR*, abs/2312.17617.

Nan Xu, Hongming Zhang, and Jianshu Chen. 2024. CEO: corpus-based open-domain event ontology induction. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 946–964. Association for Computational Linguistics.

Kosuke Yamada, Ryohei Sasano, and Koichi Takeda. 2021. Verb sense clustering using contextualized word representations for semantic frame induction. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4353–4362. Association for Computational Linguistics.

Lingxiao Yang, Ru-Yuan Zhang, Qi Chen, and Xiaohua Xie. 2025. Learning with enriched inductive biases for vision-language models. *Int. J. Comput. Vis.*, 133(6):3746–3761.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. 2024. Language models as inductive reasoners. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 209–225. Association for Computational Linguistics.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *CoRR*, abs/2303.10420.

Mengyu Ye, Tatsuki Kuribayashi, Goro Kobayashi, and Jun Suzuki. 2025. Can input attributions explain inductive reasoning in in-context learning? In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 21199–21225. Association for Computational Linguistics.

Sean Yom. 2015. From methodology to practice: Inductive iteration in comparative research. *Comparative Political Studies*, 48(5):616–644.

Jianhao Yuan, Jie Zhang, Shuyang Sun, Philip Torr, and Bo Zhao. 2024. Real-fake: Effective training data synthesis through distribution matching. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Chen Zeno, Hila Manor, Greg Ongie, Nir Weinberger, Tomer Michaeli, and Daniel Soudry. 2025. When diffusion models memorize: Inductive biases in probability flow of minimum-norm shallow neural nets. *CoRR*, abs/2506.19031.

Chi Zhang, Baoxiong Jia, Mark Edmonds, Song-Chun Zhu, and Yixin Zhu. 2021a. ACRE: abstract causal reasoning beyond covariation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 10643–10653. Computer Vision Foundation / IEEE.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, Yu Fu, Xingtai Lv, Yuchen Zhang, Sihang Zeng, Shang Qu, Haozhan Li, Shijie Wang, Yuru Wang, Xinwei Long, and 20 others. 2025a. A survey of reinforcement learning for large reasoning models. *Preprint*, arXiv:2509.08827.

Odin Zhang, Haitao Lin, Xujun Zhang, Xiaorui Wang, Zhenxing Wu, Qing Ye, Weibo Zhao, Jike Wang, Kejun Ying, Yu Kang, Chang yu Hsieh, and Tingjun Hou. 2025b. Graph neural networks in modern ai-aided drug discovery. *Preprint*, arXiv:2506.06915.

Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Zhihan Guo, Yufei Wang, Irwin King, Xue Liu, and Chen Ma. 2025c. What, how, where, and how well? A survey on test-time scaling in large language models. *CoRR*, abs/2503.24235.

Songyang Zhang, Linfeng Song, Lifeng Jin, Kun Xu, Dong Yu, and Jiebo Luo. 2021b. Video-aided unsupervised grammar induction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1513–1524. Association for Computational Linguistics.

Tianyi Zhang, Isaac Tham, Zhaoyi Hou, Jiaxuan Ren, Leon Zhou, Hainiu Xu, Li Zhang, Lara J. Martin, Rotem Dror, Sha Li, Heng Ji, Martha Palmer, Susan Windisch Brown, Reece Suchocki, and Chris Callison-Burch. 2023a. Human-in-the-loop schema induction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 1–10, Toronto, Canada. Association for Computational Linguistics.

Yue Zhang, Hongliang Fei, and Ping Li. 2022. End-to-end distantly supervised information extraction with retrieval augmentation. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2449–2455. ACM.

Zhebin Zhang, Xinyu Zhang, Yuanhang Ren, Saijiang Shi, Meng Han, Yongkang Wu, Ruofei Lai, and Zhao Cao. 2023b. IAG: induction-augmented generation framework for answering reasoning questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1–14. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2025. A survey of large language models. *Preprint*, arXiv:2303.18223.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Jialun Zhong, Wei Shen, Yanzeng Li, Songyang Gao, Hua Lu, Yicheng Chen, Yang Zhang, Wei Zhou, Jinjie Gu, and Lei Zou. 2025. A comprehensive survey of reward models: Taxonomy, applications, challenges, and future. *arXiv preprint arXiv:2504.12328*.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, Zheng Li, and Fenglin Liu. 2023. A survey of large language models in medicine: Progress, application, and challenge. *CoRR*, abs/2311.05112.

Linqi Zhou, Stefano Ermon, and Jiaming Song. 2025. Inductive moment matching. *CoRR*, abs/2503.07565.

Dominik Zietlow, Michal Rolínek, and Georg Martius. 2021. Demystifying inductive biases for (beta-)vae based architectures. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12945–12954. PMLR.

A Appendix

A.1 Inductive Reasoning and Inductive Bias

The differences between inductive reasoning and inductive bias can be discussed in the following three aspects:

Different concepts Inductive reasoning is a type of reasoning task whose input consists of some observations and whose output is a general rule or principle that is consistent with these observations. In contrast, inductive bias is a concept that represents the prior knowledge or inclination within a machine learning system (Goyal and Bengio, 2020). It is the preferences or constraints that are 'prespecified' in the model before learning. The process of learning such bias is called meta learning (Vanschoren, 2018) in the machine learning field.

Different scopes of application All machine learning and deep learning tasks (including text, vision, video, and other tasks) inherently require the model or algorithm to be injected with the inductive bias appropriate for the specific task. Inductive reasoning, however, refers to one of the tasks.

Context of discussion The goal of this inductive reasoning task is to identify rules among observations within a single task. This paper primarily discusses the inductive reasoning task within the context of LLMs. Improving the inductive reasoning ability of LLMs essentially amounts to discovering the inductive bias underlying inductive reasoning. By doing so, the scope of the discussion can be narrowed, allowing the survey to be more thorough and well-developed.

A.2 Inductive Reasoning and Other Reasoning Modes

A.2.1 Inductive Reasoning and Deductive Reasoning

We provide an example for each in Figure 7. As two major modes of reasoning, inductive reasoning (Clark, 1969; Rips, 1994; Bang et al., 2023) differs from deductive reasoning in many aspects. Please refer to Table 2 for more.

A.2.2 Inductive Reasoning and Analogical Reasoning

As stated in the main text, inductive reasoning is a process that goes from the particular to the general. In contrast, analogical reasoning is a process from the particular to the particular, which can be regarded as a special form of inductive reasoning. For example, consider an inductive reasoning task of deriving the general term formula of a number sequence: the input is a number sequence, and the output is its general term formula. Analogical reasoning, on the other hand, takes a number sequence as input and outputs its next term. In short, analogical reasoning is a form of reasoning that involves imitation based on observations. It compares and transfers similarities from existing observations to infer and generate possible next items or outcomes (Lewis and Mitchell, 2024a,b; Musker et al., 2024).

In the field of LLMs, the commonly studied ICL (Dong et al., 2023, 2024) can be regarded as a form of analogical reasoning. For ease of discussion and to clearly define the scope, a large body of research on ICL is not discussed in this paper.

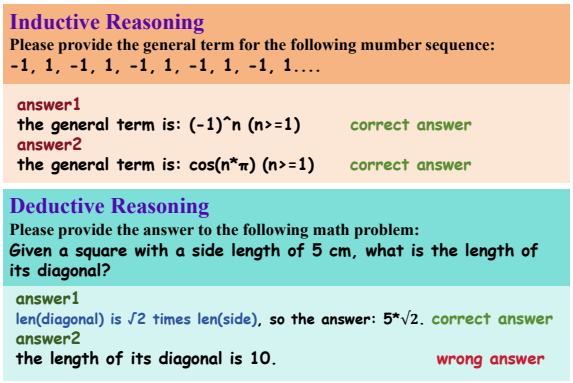


Figure 7: An example is given for both reasoning modes. In inductive reasoning, there may be multiple correct answers consistent with the existing observations, while in deductive reasoning, a precise logical reasoning process can help arrive at the only correct answer.

Table 2: The differences between inductive reasoning and deductive reasoning.

	Inductive Reasoning	Deductive Reasoning
Mode	particular-to-general	general-to-particular
Target	probabilistic conclusion	precise answer
Reliability	flexible and generative	premises-relied
Informative	extended information	no new information
Application	hypothetical inference	rigorous proof

A.3 More Synthesis and Comparative Analysis on NLP Applications

The impressive applications we cited, spanning from core syntactic and semantic parsing to complex multimodal tasks, clearly demonstrate the versatility of inductive reasoning in NLP.

Empirical effectiveness often hinges on the

method's ability to seamlessly integrate domain-specific structural knowledge as inductive bias — for instance, achieving high precision in information extraction usually involves explicit rule-based or graph-structured inductive biases, while improvements in dialogue systems frequently rely on incorporating conversational history as an inductive structure.

Approaches that embed strong, hand-crafted inductive bias (e.g., explicit grammar rules for parsing) tend to offer **superior interpretability and data efficiency** but suffer from lower generalizability across domains, whereas soft or learned bias (e.g., the knowledge encoded in the model's parameters) is **more flexible but potentially less transparent**. How to achieve a **trade-off** between the two is currently a key focus for applying inductive reasoning to downstream NLP tasks.

The major trend of inductive reasoning in NLP is a move toward methods that automate the discovery of inductive patterns, such as applying self-supervision or meta-learning to identify bias suitable for multiple tasks, marking a shift from prescriptive to descriptive induction (Kalish and Thevenow-Harrison, 2014) in method design.

This comparative analysis underscores the necessity of selecting an inductive reasoning method that carefully balances the need for robustness, interpretability, and generalization capacity for the target application scenario.

A.4 The Boundaries and Comparisons among the Methods

A.4.1 The Boundaries

Our classification minimizes overlap among the three methods conceptually as much as possible.

(1) **Post-training enhancement** directly trains LLMs using task-specific data. Synthetic data in this category is often generated from scratch to create entirely new training samples. (2) **Test-time exploration** is a train-free method based on inductive hypotheses. (3) **Data augmentation** can guide inductive tasks by leveraging annotated data and knowledge from outside the task domain manually. It can keep the model frozen, or even be used to train with fine-grained embedding or graph information. In conclusion, it modifies and expands existing training samples.

We believe that modern LLM techniques are often multifunctional, and some of the overlapping parts can be seen as emerging trends or directions for future research. For example, the construction of synthetic data inevitably incorporates prior knowledge from human experts.

A.4.2 The Comparisons

(1) **Post-training enhancement** is suitable for scenarios where sufficient data is available, and there is a need to directly enhance a single model's inherent inductive capabilities. *Advantages*: It can directly enhance the inductive reasoning ability of a specific LLM and is easy and quick to train. *Disadvantages*: It is only applicable to scenarios involving a single model, and cannot be easily used to solve complex problems that require cooperation among agent systems.

(2) **Test-time exploration** can be applied in situations with limited data or when the model cannot be fine-tuned, using hypothesis-driven methods to perform inductive reasoning tasks. *Advantages*: No model training is required, as it can leverage the capabilities of powerful closed-source LLMs. *Disadvantages*: It needs time and monetary costs, and multiple refinements do not necessarily guarantee an optimal solution.

(3) **Data augmentation** is more appropriate for scenarios where the model involves cross-domain tasks or requires the use of external knowledge (e.g., cross-domain agent systems with tool usage and human-crafted rules), or for training auxiliary modules such as embeddings or graphs, thereby improving the model's generalization and transferability to new patterns. *Advantages*: It can leverage information beyond the LLM, including even parameter-level information, to provide assistance. *Disadvantages*: Its working mode is similar to that of an agent system, giving it potential to handle complex real-world problems. However, it relies on external data and tools, and the overall system performance depends on the fundamental capabilities of each component.

A.4.3 Why Synthetic Data is Widely Used

Synthetic data accounts for a significant portion of many well-known LLM training corpora (DeepSeek-AI, 2024; QwenTeam, 2025; InternTeam, 2025) and has the following advantages compared with real data.

(1) **Coverage of rare cases**: Synthetic data can include samples that are uncommon in real-world scenarios. For example, in a batch of real data, certain inductive patterns may appear infrequently, and directly training on such data could introduce

bias into the model. (2) **Support for data-scarce or privacy-sensitive domains**: In fields where data is limited or highly sensitive, synthetic data can be generated at a large scale, enabling the training of more generalizable models. (3) **High efficiency**: Synthetic data can be generated very flexibly and at a large scale, which makes model training and debugging more efficient. (4) **Sometimes better performance**: In certain cases (Yuan et al., 2024; Shidani et al., 2025), synthetic data can even yield better results. Because it may better fit the model's distribution and even provide regularization.

For these reasons, it is also a better choice for inductive reasoning.

A.5 More Details about Benchmarks

In this section, we will introduce the inputs and outputs of LLM inductive reasoning benchmarks one by one in detail.

ILP (Evans and Grefenstette, 2018). It takes as input background knowledge in first-order logic, along with a pair of positive and negative examples for a specific first-order logic case, and the model is required to output a first-order logic that satisfies these conditions.

SCAN (Lake and Baroni, 2018). It takes as input a series of entities and their states, and requires LLMs to output the actions needed to achieve those states. For example, given 'a ball on the table' as input, the model should output 'place it on the table'.

ARC (Chollet, 2019). It takes as input several pairs of grids in natural language form, where they illustrate a specific transformation pattern. Then, given a new grid as input, the model is asked to output what the transformed grid would look like.

List Functions (Rule, 2020). It takes as input several pairs of number lists, where they illustrate a specific transformation pattern. Then, given a new number list as input, the model is asked to output what the transformed number list would look like.

PROGES (Alet et al., 2021). It provides several input-output pairs of a program and requires the model to generate the program itself.

SyGuS (Odena et al., 2021). It takes as input several pairs of strings, where they illustrate a specific transformation pattern. Then, it requires the model to generate a program to show this transformation process.

ACRE (Zhang et al., 2021a). It takes as input the results of interactions between different entities and a certain machine (i.e., the functions of different entities), and models need to output the entity corresponding to a specific function.

Instructions (Honovich et al., 2022). The model is given two natural language statements, A and B, where B is obtained by applying a certain instruction to A, and it is required to output that natural language instruction.

Arithmetics (Wu et al., 2024). It takes as input several pairs of two-digit additions along with their results, where these additions follow a calculation process in a certain numeral base. Given a pair of two-digit numbers, the model is required to output the result of their addition in the same base.

Levy/Holt (Liu et al., 2024b). It takes as input a pair of triplets, where the positive triplet represents a factual relationship and the negative triplet represents a counterfactual relationship. The task is to output an inference rule between triplets such that the positive triplet entails the negative triplet.

NutFrame (Guo et al., 2024). It takes text fragments that contain potential frames and frame elements, along with contextual information such as sentence structure, lexical cues, and semantic hints. This input helps the model identify underlying conceptual structures in the text. The model produces three types of outputs. Frame Induction: Identifies latent frames expressed in the text and maps them to existing FrameNet frames or proposes new frames. Frame Element Identification: Detects specific frame elements within the text. Frame Filling: Assigns concrete values from the text (entities or phrases) to the identified frame elements.

DEER (Yang et al., 2024). It takes as input several pairs of facts, where they illustrate a specific real-world rule. Then, it requires the model to generate the rule.

RULEARN (He et al., 2025) The input to it consists of three parts. Puzzle Scenarios: Each puzzle presents a set of conditions or operations that the agent can manipulate. These scenarios are designed to have underlying rules that are not explicitly provided. Agent Actions: The agent can perform a variety of actions, such as inputting integers or letters, to interact with the puzzle environment. Feedback: After each action, the agent receives feedback that

helps in refining its understanding of the hidden rule. The desired output is: the hidden rule governing the puzzle scenario based on the feedback received from its actions.

Cryptography (Li et al., 2025a) It takes as input several pairs of English words, where each pair follows a certain cryptographic transformation pattern. Given a new word, the model is required to output the new word obtained by applying the same transformation pattern.

GeoILP (Chen et al., 2025c) The same as ILP in general.

InductionBench (Hua et al., 2025) It takes as input a pair of strings, where the pair follows a certain transformation rule, and the task is to output that rule.

CodeSeq (Chen et al., 2025a) It takes as input a number sequence of numbers and is required to output the number sequence's general formula, with the entire output presented in code form.

Statements We determine the number of test samples for each benchmark by consulting the original paper, the corresponding GitHub repository, and the Hugging Face dataset files, either through direct counting or estimation.

A.6 Failure Analysis

Different benchmarks evaluate different models and do not provide a standardized, unified evaluation strategy. Therefore, for the sake of fairness, we cannot provide directly comparative results.

Hence, we summarize several typical failure modes of LLMs on these inductive tasks based on prior experiences and the conclusions drawn in previous studies.

(1) **Difficulty in internalizing inductive logic.** LLMs tend to capture only surface-level associations among observations, while failing to truly acquire the underlying logical or mathematical relations. For example, when presented with new samples or longer samples following the same pattern, the model often cannot induce the correct rule or produce the correct answer. (2) **Spurious pattern matching.** LLMs often latch onto accidental or superficial patterns that happen to fit the limited observations, leading to spurious rules that do not reflect the true underlying structure. These rules typically fail on even slightly varied test cases, revealing that the induced pattern was an artifact rather than a genuine abstraction. (3) **Failure in multi-step or compositional reasoning.** Many inductive tasks require composing several primitive operations (e.g., nested arithmetic, layered string transformations, or multi-step grid manipulations). LLMs often fail to correctly chain these operations, producing partial reasoning steps that appear correct individually but do not combine into a coherent global rule. (4) **Syntactic formatting errors.** The rule induced by the model may be semantically correct, but it does not conform to the syntactic constraints of the task or the underlying logical system. For example, the generated code-like rule may contain syntax errors, making it impossible to execute. (5) **Data sparsity challenge.** The rules induced by the model often apply only to the most frequent observations. When the distribution of inductive patterns is uneven, the model is prone to errors on rare or boundary-case observations that lie in the long tail.

A.7 Advantages of Sandbox and OC

A.7.1 Advantages of Sandbox

We will elaborate on the advantages of sandbox testing from the following points.

(1) **Meeting the demand for complex real-world tasks**: With the development of LLMs, there is an increasing demand for handling more complex tasks in real-world scenarios, such as scientific discovery and simulations. Consequently, Agent technologies have also advanced rapidly. In these scenarios, task success rates cannot be adequately measured using traditional metrics. Sandbox testing, as a highly integrated evaluation framework combining LLMs, tool usage, and complex rules, is better suited to address modern AI tasks, including inductive-related tasks. (2) **Isolation and safety**: Sandboxes are typically isolated environments with higher security, which helps reduce bias and risk while allowing rapid problem detection. In contrast, traditional metrics provide only coarse numerical indicators. (3) **Scalability**: Sandboxes are scalable, and existing sandbox frameworks can cover thousands of LLM tasks (Li et al., 2025c). (4) **Many works adopt a sandbox as the evaluation tool**: Their experimental results demonstrate that a sandbox can reveal a greater variety of errors (Milev et al., 2025), provide more logical explainability (Sheffler, 2025), and cover more boundary behaviors (Alshahwan et al., 2024).

Many existing LLM foundation trainings, such as the Intern series, as well as the Minimax-M1

and Seed-thinking models, employ large amounts of all kinds of reasoning data and sandbox validation to enhance the models' fundamental reasoning capabilities.

A.7.2 Advantages of OC

The OC evaluation method has the following advantages.

(1) **Overcoming averaging effects**: Metrics like ACC only measure overall accuracy, reflecting the average performance, but they do not reveal whether the model works effectively across all observations. Measuring OC provides a more fine-grained evaluation signal. (2) **Improved explainability**: This observation-level evaluation can uncover a greater variety of error types, allowing analysis of specific errors. (3) **Empirical support from existing work**: Previous studies (Chen et al., 2025a) (it reformulates number sequence as algorithmic problems and uses the OC of code solutions as the training signal) show that using observation-level supervision signals can effectively enhance a model's inductive reasoning abilities.

A.8 Future Works

We will further expand and discuss the prospects and potential applications of inductive reasoning in this subsection.

A.8.1 In AI domain

Inductive reasoning tasks, benchmarks, and data The inductive reasoning capability of LLMs can be examined from various other perspectives, such as format imitation (Cheng et al., 2024a), cross-domain induction (Chen et al., 2024a), multimodal inductive (Wang et al., 2024b), and so on. Even a single data point has the potential to be expanded into either SFT or RL data. This serves as a concrete truth demonstrating the scalability of inductive reasoning data. Therefore, constructing more inductive reasoning datasets and evaluation methods is key to improving LLMs' inductive reasoning capabilities.

Enhancing the inductive reasoning ability of LLMs We summarize several typical failure modes of LLMs on inductive tasks in Appendix A.6. Addressing these issues can further enhance the model's inductive reasoning ability. Using simple and effective training data and architectures to identify universal inductive biases is a reasonable option. In Section 5, we conclude some analyses of inductive reasoning theoretically from current practical experience. These analyses are highly beneficial for improving the inductive reasoning capabilities of LLMs. The paper also presents three major categories of methods, and it is expected that more methods will be proposed in the future.

Considering that BPE is the longest-established inductive reasoning task with the richest body of research, we will list several algorithms related to solving the BPE task below.

Example I: BPE Inductive program synthesis (or programming-by-example, PBE) is fundamentally an inductive reasoning task and one of the earliest to originate: given a few procedural input-output pairs, the system must generalize to unseen inputs. Li and Ellis (2024) argues that PBE is a highly general form of few-shot inductive inference, and they show that while pretrained LLMs perform poorly on classic PBE tasks (e.g., list and string manipulations), fine-tuning on in-distribution PBE data substantially improves performance. To address the inherent lack of search in LLMs, Verbruggen et al. (2025) introduce a within-prompt search method: they sample lines of code, execute them on the input examples, feed back the semantic outcomes, and use that to guide further generation, effectively letting the model perform a semantic-aware search. Building on this inductive foundation, Lee et al. (2025b) propose a transductive framework: at test time, they actively select test inputs, have the LLM predict their outputs, and eliminate candidate programs that disagree — this refines the inductive hypothesis space using feedback. In addition, recent work such as Khan et al. (2025) breaks down complex PBE tasks into subtasks, with the LLM guiding the synthesis of these subtasks and then recombining them, thus improving the model's ability to solve difficult problems. At the same time, the CodeARC (Wei et al., 2025) benchmark has been introduced to evaluate LLMs' reasoning ability in real-world inductive synthesis settings: the model can interactively query a black-box target function, propose candidate functions, and iteratively refine them via a differential testing oracle.

These advances collectively show that under LLM-driven paradigms, PBE is being reactivated and reshaped: modern methods integrate fine-tuning, search, semantic execution, and increasingly also active learning and feedback mechanisms to boost generalization and robustness.

A.8.2 In Real-world Domain

Driving scientific discovery and innovation In the context of scientific research, inductive reasoning is a vital method for uncovering natural laws. Artificial intelligence, especially generative AI, is becoming a new tool for scientists to propose hypotheses and design inductive experiments (Reddy and Shojaee, 2025). By learning patterns from large-scale data, these models can suggest plausible explanations and explore vast hypothesis spaces that would be difficult for humans to enumerate manually. This can serve a wide range of scientific fields, including the technological applications mentioned in Section 2.2.2.

Example II: AI4Science AI4Science constitutes a form of inductive reasoning by LLMs: instead of merely fitting data, LLMs generate symbolic hypotheses (programs) that generalize from observed examples, thereby performing induction over scientific phenomena. For instance, Shojaee et al. (2025) proposes new equation 'skeletons' using an LLM informed by scientific priors, and refines them with evolutionary search to discover mathematically and physically meaningful relations. This work touches on debates about the nature of scientific reasoning. Barnum (2012) argues that science is not purely induction, which raises interesting questions about whether LLM-based equation discovery should be framed as induction, deduction, or something more agentic. More broadly, Merler et al. (2024) uses LLMs to iteratively propose functions and optimize coefficients, reinforcing that LLMs are now capable of hypothesis generation rather than mere curve-fitting. Extensions like Wang et al. (2025a) further enrich this paradigm by combining data-driven insights with a reflective feedback loop for refining symbolic candidates. Taken together, these developments suggest that LLM-driven scientific discovery can be viewed as one of the future works of inductive reasoning.

Human–machine collaboration in education As cognitively-enhanced AI develops, future educational systems may evolve into "human–in-the-loop" intelligent agents (Chu et al., 2025). Inductive reasoning can help AI systems better understand students' ways of thinking and their learning processes. By inferring latent patterns from students' interactions and responses, such systems can adapt instruction strategies to individual needs and learning trajectories. Furthermore, inductive reasoning enables AI to provide more interpretable feedback and personalized guidance, fostering more effective and engaging human–AI collaboration in education.

New perspectives from philosophy and cognitive science From the perspective of cognitive science, inductive reasoning is a crucial tool for understanding human thought and the structure of cognition. Future research may further explore the similarities and differences between the inductive reasoning of AI and that of human thinking (Trepczynski, 2024). Such investigations could shed light on whether AI systems rely on mechanisms analogous to human abstraction, generalization, and concept formation, or whether they achieve inductive competence through fundamentally different processes.

Ethical and social implications As inductive reasoning capabilities become more powerful, AI systems' decisions or recommendations may embody inductive generalizations that 'seem reasonable but could be wrong'. Avoiding inductive bias will thus become a significant issue (Ueno et al., 2022). Such biases may arise from spurious correlations, incomplete data, or mismatches between training environments and real-world contexts, leading to systematic errors that are difficult to detect. Addressing this challenge will require not only improved model design and evaluation but also greater transparency, uncertainty awareness, and human oversight to ensure that inductive inferences remain reliable and trustworthy.