Advancing Reasoning in LLMs: From Abstraction to General Intelligence

Boyuan Chen
Yuanpei College
Peking University

Abstract

The recent advancements in large language models (LLMs) have significantly enhanced their reasoning capabilities, particularly through the integration of System II thinking. However, LLMs remain far from achieving human-like reasoning, which is characterized by abstract reasoning—a cognitive ability that transcends dependence on specific details and involves the recognition of patterns, induction, analogy, and inference. This essay explores the core principles of abstraction reasoning and methodologies for simulating it in LLMs. We analyze the limitations of existing benchmarks in evaluating true reasoning capabilities, particularly their failure to distinguish between task similarity and task complexity. Furthermore, we propose novel task designs inspired by concepts such as Raven's Progressive Matrices and counterfactual reasoning to isolate and evaluate reasoning capabilities more effectively. Finally, we discuss the transition from prototype-centric to program-centric abstraction, emphasizing the integration of program synthesis and RL as potential pathways to achieving true general intelligence.

1 Introduction

To be rational is to possess the capacity for reasoning. The recent introduction of OpenAI o1 [12] has highlighted significant advancements in enabling System II thinking in large language models (LLMs), substantially enhancing their reasoning capabilities. However, current LLMs remain far from achieving human-level reasoning. Human reasoning is an advanced cognitive ability characterized by the recognition of patterns, induction, analogy, and inference to solve problems, transcending dependence on specific details. This capability involves abstracting information, identifying relationships between entities, and drawing conclusions without relying on concrete contexts or direct sensory input, often referred to as *abstraction reasoning* [5, 17].

In this essay, we first review the core principles of abstraction reasoning and methodologies for simulating it (Section 2). Subsequently, we discuss the limitations of existing benchmarks in accurately capturing and evaluating the true reasoning capabilities of LLMs, particularly due to their failure to distinguish between task similarity and task complexity (Section 3). Furthermore, we outline key features that tasks aimed at assessing general reasoning capability should exhibit. Finally, we explore the integration of deep learning and program synthesis, a potential pathway towards achieving true reasoning and realizing general artificial intelligence (Section 4).

2 Abstraction Reasoning: Core Principles and Insights

Abstraction reasoning is the human ability to generalize concrete information into abstract representations and utilize them for reasoning, creation, analogy, and innovation. This abstraction facilitates a vast generalization space for information, enabling humans to connect disparate concepts and derive new insights. The key components of abstraction reasoning can be summarized as follows:

- **Pattern Recognition**: The ability to observe and analyze data or phenomena to uncover regularities and patterns. Examples include identifying mathematical formulas, recognizing geometric shapes, or understanding syntactic rules in language.
- **Analogical Reasoning**: Drawing relationships based on similarities between two entities to infer additional properties or relationships. For instance, understanding that *birds can fly* and *bats can fly* leads to recognizing their shared characteristic as species adapted for flight.
- **Inductive Reasoning**: Summarizing general rules from specific instances. For example, observing several white swans and hypothesizing that "all swans are white," although this conclusion may not always be accurate.
- **Deductive Reasoning**: Inferring specific conclusions from general principles or premises. For instance, from the premises *all humans are mortal* and *Socrates is human*, we deduce that *Socrates is mortal*.

2.1 Concept Learning as a Manifestation of Abstraction Reasoning

A crucial manifestation of abstraction reasoning is *concept learning*, where humans can learn complex concepts from a simple example and apply these learned concepts to richer behaviors such as *action*, *imagination*, and *explanation* [7]. Concept learning illustrates the core insights of human abstraction reasoning, which include:

Reasoning Compositionality Human reasoning is inherently compositional, meaning it combines simple elements to form more complex structures. This enables humans to reuse learned components across tasks, facilitating systematic generalization.

Causality Causal reasoning allows humans to identify the relationships between causes and effects, enabling predictions, explanations, and interventions. This forms the basis of understanding and interacting with dynamic systems.

Learning to Learn Humans possess the ability to improve their learning processes over time, leveraging prior experiences to accelerate the acquisition of new concepts. This meta-learning capability allows efficient adaptation to novel situations.

Concept learning exemplifies how abstraction reasoning integrates these dimensions to generate robust and adaptable cognitive processes. The following sections will explore how to design systems based on these principles and how them are applied to evaluating and improving reasoning capabilities in modern AI systems.

2.2 Systematic-construction for Abstraction Reasoning

Early psychological theories proposed that human reasoning relies on formal rules of inference, akin to the framework of logical calculus, commonly referred to as *mental logic* [5]. However, this explanation is insufficient, as individuals without formal training in logic are still capable of making logical deductions, even in unfamiliar or abstract domains. Such observations challenge the idea that formal rules alone can fully capture the complexity and flexibility of human reasoning.

A more comprehensive perspective is provided by the theory of *mental models*. When perceiving the world, humans construct mental representations that describe "what" objects are and "where" they exist within a scene [9]. Similarly, understanding verbal or textual descriptions involves forming comparable mental models grounded in sensory inputs and prior knowledge. Reasoning within this framework entails refining these models to arrive at conclusions that are true, probable, or at least plausible, while aiming for novelty, simplicity, and informativeness [7, 6].

This view conceptualizes human reasoning as a process of narrowing a hypothesis space defined by mental models. Through abstraction, irrelevant details are excluded to focus on essential relationships, while heuristic search enables efficient exploration of potential solutions guided by intuition and prior experience. Moreover, given the limitations of human working memory, reasoning often incorporates inductive biases, favoring consideration of a single possibility to simplify inference (*e.g.*, switching between System I and System II thinking). Visual representations, when designed intuitively and aligned with the task, can further enhance abstraction by providing clear structures [17].

2.3 Modeling of Abstraction Reasoning

To address the limitations of artificial neural networks in systematic generalization—particularly their challenges in simulating human-like abstraction reasoning—various modeling approaches have been developed based on the core features of abstraction reasoning.

One prominent method is *Bayesian Program Learning* (BPL) [7], which represents concepts as simple probabilistic programs, expressed through abstract descriptive languages. These probabilistic generative models enable the learning of a wide range of visual concepts from just a single example and support extensive generalization. This approach shares similarities with *program synthesis*, where, given a set of examples, the goal is to generate programs via: (1) *Enumerative Search*: Exhaustively searching for hypotheses consistent with the examples, often guided by Bayesian inference. (2) *Neural-Enhanced Generation*: Employing neural networks to identify programs that maximize the explanatory power of the observed data.

This process exemplifies a high-level abstraction mechanism, akin to *compression*, as it reduces the Kolmogorov complexity of the representation [8]. In this sense, abstraction can be regarded as a behavior of intelligence, distilling patterns into concise and generalizable forms.

To tackle challenges in learning compositionality, meta-learning has been introduced. By training networks on a sequence of compositional tasks, meta-learning equips models with the capacity to perform systematic behaviors, effectively enabling *learning to learn* [6]. Furthermore, some approaches integrate visual information into joint training frameworks for perception and reasoning, thereby significantly enhancing capabilities in logical induction and graphical reasoning [15, 16].

Another critical challenge is balancing the tractability of reasoning with the expressivity of the hypothesis class [4]. This requires combining bottom-up and top-down capabilities: (1) *Bottom-Up Generalization*: Inferring universal rules from limited examples. (2) *Top-Down Inductive Priors*: Using neural networks to learn prior distributions for concept generation from human behavioral data. These priors act as inductive biases, aiding heuristic search and facilitating the generation of structured, generalizable abstractions.

3 Evaluating Abstraction Reasoning in LLMs

Current benchmarks for evaluating LLMs predominantly focus on assessing their knowledge and problem-solving capabilities, often through tasks derived from standardized exams. However, a growing body of research suggests that the reasoning abilities required to solve these tasks are, in many cases, more closely related to memorization than genuine reasoning [10, 11, 3, 13]. The central idea is that LLMs' success in these tasks depends primarily on the *task similarity* to examples encountered during pretraining, rather than on the *task complexity*. For instance, while LLMs can solve seemingly complex problems like the Monty Hall problem—because corresponding examples and solutions exist on the internet—they often fail at tasks requiring deeper reasoning, such as overcoming the *reversal curse* [1]. This underscores the need for benchmarks that can disentangle task similarity from task complexity to truly evaluate the reasoning processes of LLMs.

3.1 Designing Tasks to Isolate Reasoning Capability

To address the limitations of task similarity, we propose the development of tasks that eliminate reliance on pre-existing conceptual knowledge and require models to establish new concepts from scratch. A notable example of such tasks is Raven's Progressive Matrices (RPM), a widely used human IQ test designed to assess abstract reasoning and fluid intelligence. RPM achieves this by relying solely on visual patterns and relationships rather than prior knowledge, forcing participants to deduce rules and complete matrices through abstraction. Inspired by RPM, we can design tasks that incorporate combinatorial reasoning, variations in colors, objects, and relationships, enabling us to evaluate LLMs' reasoning capabilities without reliance on prior training examples.

3.2 Counterfactual Tasks for Robust Evaluation

Another promising approach is the use of counterfactual tasks, which deviate from the default assumptions of conventional tasks while preserving the underlying reasoning processes. These tasks challenge models to operate under novel conditions, testing their ability to generalize reasoning to unfamiliar contexts. Examples include arithmetic problems with base changes (*e.g.*, switching from decimal to nonary systems) and syntactic analysis in non-standard grammars [13, 14]. Such tasks demand abstraction and systematic generalization, directly reflecting the core requirements of abstract reasoning. By probing the model's ability to understand diverse rules, execute complex tasks, and adapt to new environments, counterfactual tasks provide a deeper evaluation of reasoning capabilities.

4 True Intelligence: Transitioning from Prototype-Centric to Program-Centric Abstraction

As discussed above, while LLMs appear to excel in certain reasoning tasks, much of their success stems from leveraging task similarity rather than demonstrating genuine reasoning or generalization. The core issue lies in the static nature of current LLMs, which exhibit low levels of abstraction. Their reasoning is grounded in large-scale data and primarily employs next-token-prediction as a form of compression, resulting in what can be described as *prototype-centric abstraction* [2]. This form of abstraction effectively captures shared structures across tasks, such as reusing abstract subroutines, but struggles with solving complex logical reasoning tasks. Such tasks are often discrete, independent, low-sample, and lack regularity, requiring *discrete search* and *symbolic tools* to solve "on the fly"—capabilities more aligned with *program-centric abstraction*.

4.1 From Prototype-Centric to Program-Centric Abstraction

Pretrained models trained on large datasets excel at capturing implicit task structures, providing valuable intuition for navigating program search spaces and solving higher-level problems. However, transitioning from prototype-centric to program-centric abstraction requires integrating LLM training methodologies with program synthesis techniques. This involves mapping discrete task concepts to program sketches that serve as structured solution frameworks, generating candidate code snippets with fuzzy behavioral characterizations for iterative refinement, using embeddings to represent intermediary or incomplete programs for step-by-step problem-solving, and leveraging pretrained models to guide intermediate outputs and branching decisions, enhancing efficiency and accuracy in program synthesis.

This hybrid approach combines the data-driven intuition of LLMs with the precision and flexibility of symbolic reasoning, allowing for more robust abstraction capabilities and higher reasoning complexity.

4.2 Incorporating Reinforcement Learning into Abstraction Learning

An additional promising path for advancing abstraction learning is incorporating RL to incentivize program-centric abstraction. RL provides a framework for aligning model behavior with desired reasoning objectives by introducing reward mechanisms. These mechanisms can:

- Encourage models to explore diverse solution strategies beyond pre-trained patterns, fostering creativity and adaptability.
- Reward the generation of modular and reusable subprograms, promoting systematic generalization across tasks.
- Penalize reliance on heuristics that prioritize task similarity over task complexity, shifting focus to solving novel and challenging problems.

By integrating RL into the abstraction learning pipeline, LLMs could dynamically refine their reasoning processes, adapt to unseen challenges, and transition more effectively from prototype-centric to program-centric abstraction. This hybrid paradigm represents a step toward equipping AI systems with the flexibility and depth required for true reasoning and generalization.

5 Conclusion

In this essay, we examined the critical role of abstraction reasoning in human cognition and its implications for developing reasoning capabilities in LLMs. While current LLMs exhibit impressive performance in some reasoning tasks, their reliance on task similarity rather than task complexity highlights a significant gap in genuine reasoning and generalization. To address this, a transition from prototype-centric to program-centric abstraction is essential, leveraging the integration of program synthesis and reinforcement learning. By designing tasks that emphasize systematic generalization and abstraction, such as those inspired by Raven's Progressive Matrices and counterfactual reasoning, we can better evaluate and enhance the reasoning processes of LLMs. These insights pave the way for building models that align more closely with human-like reasoning, marking an essential step towards achieving true general intelligence.

References

[1] Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*, 2023. 3

[2] François Chollet. It's not about scale, it's about abstraction. `https://www.youtube.com/watch?v=s7_NlkBwdj8`, 2024. 4

[3] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[4] Kevin Ellis. Human-like few-shot learning via bayesian reasoning over natural language. *Advances in Neural Information Processing Systems*, 36:13149–13178, 2023. 3

[5] Philip N Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010. 1, 2

[6] Brenden M Lake and Marco Baroni. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623(7985):115–121, 2023. 2, 3

[7] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 2, 3

[8] Ming Li, Paul Vitányi, et al. *An introduction to Kolmogorov complexity and its applications*, volume 3. Springer, 2008. 3

[9] D Man and A Vision. A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company, San Francisco*, 1: 1, 1982. 2

[10] Marianna Nezhurina, Lucia Cipolina-Kun, Mehdi Cherti, and Jenia Jitsev. Alice in wonderland: Simple tasks showing complete reasoning breakdown in state-of-the-art large language models. *arXiv preprint arXiv:2406.02061*, 2024. 3

[11] Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. *arXiv preprint arXiv:2410.21272*, 2024. 3

[12] OpenAI. o1. `https://openai.com/index/learning-to-reason-with-llms/`, 2024. 1

[13] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. *arXiv preprint arXiv:2307.02477*, 2023. 3, 4

[14] Sirui Xie, Xiaojian Ma, Peiyu Yu, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Halma: Humanlike abstraction learning meets affordance in rapid problem solving. *arXiv preprint arXiv:2102.11344*, 2021. 4

[15] Chi Zhang, Baoxiong Jia, Feng Gao, Yixin Zhu, Hongjing Lu, and Song-Chun Zhu. Learning perceptual inference by contrasting. *Advances in neural information processing systems*, 32, 2019. 3

[16] Chi Zhang, Sirui Xie, Baoxiong Jia, Ying Nian Wu, Song-Chun Zhu, and Yixin Zhu. Learning algebraic representation for systematic generalization in abstract reasoning. In *European Conference on Computer Vision*, pages 692–709. Springer, 2022. 3

[17] Yixin Zhu, Tao Gao, Lifeng Fan, Siyuan Huang, Mark Edmonds, Hangxin Liu, Feng Gao, Chi Zhang, Siyuan Qi, Ying Nian Wu, et al. Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, 6(3):310–345, 2020. 1, 2