Dhruval Shah
Short Story Assignment- Summary paper

Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models - A Survey

Introduction

This survey paper dives deep into the methodologies that are traditionally used in testing accuracy of the Large Language Model and shows how the LLMs are capable enough of recognizing patterns that it has learnt from its training data but fail to perform well with new data presented. This survey paper shows that there is still a gap in our understanding of LLM's reasoning process as they usually rely on shallow patterns rather than focusing on deeper insights like a human brain does. This survey paper shows how there is a need for further research that can distinguish between AI and human reasoning. In this survey, the main focus was to see how an LLM is working out to solve the complex problems by studying its reasoning processes.

The Foundations of Reasoning: Key Concepts and Definitions

Reasoning is the first definition that is introduced to the readers and shows how it is not a new thing to explore reasoning in both AI and humans. It shows reasoning in humans is not an explicit thing but an implicit one which we use in common practice and most of the times we don't even realize that our end result is always based on some reasoning. Similarly, in LLMs for every end result there is some reasoning attached to it.

Reasoning behavior shows how the LLMs, when given a reasoning task, respond to it. It is about checking their actions, expressions, and the mechanisms they use to solve and process the task at hand. While humans tend to speak out the thought process or demonstrate physical cues while solving that particular problem, LLMs show their thoughts through computed responses to the stimuli presented by reasoning tasks.

Reasoning performance states how the model based on the Reasoning behavior performs and adheres to the accuracy metrics we use to judge the LLM.

The survey paper shows the tasks of how the reasoning tasks are categorized into two broader categories

A. Core Reasoning Tasks
B. Integrated Reasoning Tasks

A. Core Reasoning Tasks focuses on a single aspect of reasoning that is mathematical or logical or casual reasoning.
B. Integrated Reasoning Tasks takes two or more fundamental reasoning skills one at a time and utilizes them.

This survey takes into account just the core reasoning tasks into consideration for LLMs especially logical, mathematical, and causal reasoning tasks. This survey doesn't include the review of the models' reasoning behaviors within the

context of integrated reasoning tasks.

Dissecting Reasoning: LLMs Across Core Reasoning Tasks

This section of the survey paper shows how these models currently behave across three core
reasoning tasks:

A. Logical reasoning
B. Mathematical reasoning
C. Causal reasoning tasks.

A. **Logical Reasoning:** The core of logical reasoning is the ability to deduce results from premises following a structured set of rules. LLMs have demonstrated varying levels of adeptness in this arena, engaging in deductive, inductive, and abductive reasoning. However, this journey is not without its issues. Surveys in the paper show a tendency of LLMs to generate likely results but they ultimately turn out to be incorrect inferences, showing the LLMs tend to rely on a superficial pattern that they find out and not a logical deduction. This way the LLM usually takes a shortcut to reasoning, that sometimes produces the correct results but usually that is not the case. These results show that the superficial pattern recognition of the LLM does not come out on top when they are faced with new, novel and complex scenarios which are falling outside their training.

B. **Mathematical Reasoning:** This part of reasoning navigates the domain of mathematical reasoning. LLMs are usually given both structured arithmetic numerical and complex word problems, requiring not only computational skills but also the ability to apply and parse mathematical concepts. The LLMs usually can give amazing solutions to a problem that it has already seen, but their performance reveals a significant dependence on memorization rather than understanding a given problem. Whenever the

LLM is introduced to a slightly altered problem statement, the solution to the problem and the output suffer, significantly hampering their ability to generate and come to the correct solutions. This stresses upoon challenges in generalizing mathematical principles beyond just cramming up the solution.

C. **Causal Reasoning**: The survey paper shows that this is the most complex of reasoning tasks. Causal reasoning is like figuring out how a thing is done, like solving a puzzle or a board game. It can figure out how a part of the puzzle can be solved and the game can be moved forward as it has already seen that pattern before in it's training phase but the struggle begins when the tasks are novel. In the context of puzzle the struggle gets real when it encounters a position that it has never seen in the training part. This shows that the LLMs are not truly grasping the things and just memorizing the things.

So observing these core reasoning tasks, a constant theme comes to the surface and is easily visible that LLMs perform exceptionally well in the problems that they have already seen in the training phase, which showcases the ability of LLm to nicely imitate reasoning via pattern recognition and memorization. But, this kind of surface level knowledge and pattern recognition does not do good when a novel problem arises that requires a bit of thinking or applying analytical skills. This tendency to look forward to the learned patterns, is sometimes efficient, but usually leads to inaccuracies and oversights, particularly in novel environments.

This examination of LLMs across various reasoning tasks not only shows the current capabilities and limitations of these models but also implies the importance of advancing beyond mere task performance metrics. To truly unlock and showcase the potential of LLMs in evaluating human-like reasoning, the focus should shift towards a deeper understanding of the processes underlying their responses, and taking a step forward from just pattern recognition to genuine understanding of the task.

Probing the Depths of Understanding: Pillars of Evaluation

As the paper dives deeper into the capabilities and complexities of the Large Language Models (LLMs) in reasoning tasks, the methodologies used to evaluate their performance are crucial. These approaches shed some light on the current capabilities of LLMs and also define the path for future advancements. The four pillars for evaluating the methodologies as stated in the paper, with each of them offering a unique insight into the reasoning behavior of LLMs:

A. Conclusion-based
B. Rationale-based
C. Interactive
D. Mechanistic evaluations

A. Conclusion-Based Evaluation: This methodology typically concentrates on just the end result of the reasoning process and not the thinking behind it. It measures and evaluates the accuracy and relevance of conclusions generated by LLMs in response to given tasks. However, it does not just evaluate whether an answer is right or wrong but through error analysis, it discovers patterns in the mistakes it made, such as some conceptual misunderstandings or any sensitivities to the context of task at hand. Dynamic benchmarks, such as functional benchmarks, further test out the model's adaptability and generalization capabilities by presenting variations on familiar tasks.

B. Rationale-Based Evaluation: The second way of evaluating the reasoning behavior as stated in the paper is rationale-based evaluations. It evaluates the 'thought process' of LLMs. This involves analyzing the chain of thoughts of the LLM that is evaluated and logical steps taken by the model in reaching their conclusions. The techniques that are mentioned in the paper like structured

parsing and interpretable quantitative metrics assess the coherence, validity, and factual consistency of these rationales. These evaluations are necessary because all of the time we just do not want an answer from the model but we require actual steps to reach that answer to find out and conclude if the LLM actually is mimicking the human chain of thought behavior.

C. Interactive Evaluation: Going a step ahead from basic static tasks, interactive evaluations engage LLMs in dynamic conversations that are based on the model's and human's responses like an AI chatbot that learns on the go. This method closely resembles human learning environments, where the feedback and new information constantly configure the reasoning processes. It's particularly useful in assessing the models' agility to adapt to the newer reasoning whenever there is some new information and the ability to defend their conclusions, and engage in more sophisticated conversation based reasoning tasks.

D. Mechanistic Evaluation: According to the paper this is the most thoughtful method as mechanistic evaluations tries to understand the internal workings of LLMs during the reasoning process. By probing the models' attention patterns, layer functions, and activation flows, surveyors gathered insights of the computational bases of LLM reasoning. This approach is like looking into the brain of the AI, trying to understand the neurological equivalent of reasoning steps and strategies.

In the survey paper all of the 4 methodologies offer a comprehensive look at evaluating the reasoning behavior behind LLMs. Yet, they also seem to present a crucial challenge, which is that the complexity of reasoning processes in LLMs cannot be fully captured by any single method. A multifaceted approach is necessary, one that combines various evaluation strategies to clear the picture of LLM reasoning capabilities. As the paper surveys further, we get to know what the LLMs can achieve, and refine their evaluation strategies which should be the key to unlocking deeper, resembling more human-like reasoning and behavior in the artificial mind of the LLM.

Beyond Computation: Insights and Implications

The paper reviews the reasoning capabilities of LLMs which offers deep and complex insights, extending the computational prowess and sheds light on both the benefits and drawbacks of Artificial Intelligence. As we navigate through the complex web of LLMs' reasoning behavior, we can come to a conclusion that though these LLMs have a strong and proven track record it can simulate or mimic human-like behavior and their thought processes.

The imitation of reasoning: LLMs have shown an impressive capability to tackle a wide range of reasoning tasks that quite often produce results that during a high level review feel like it resemble human-like thought processes. However, when we dig deeper into this insight it comes to light that the LLMs that are currently being evaluated often depend on just identifying the patterns of the problem it has seen during its training phase and not really depending upon a chain of thought process unlike a human. Due to this characteristic the models are not able to generalize and adapt to newer, novel problems which sometimes raises few questions on their reasoning capabilities.

The Importance of Out-of-Distribution Generalization: This paper identifies the most obvious limitations of state of the art LLMs, which is their struggle with unusual scenarios like the tasks that the LLM was not trained on. This challenge emphasizes a crucial and very visible disparity between human and machine reasoning: the human ability to apply and relearn the learned concepts flexibly and creatively in unfamiliar and unseen contexts, a feat that LLMs still need to get their feet on. If and when this limitation is addressed the LLMs will be really moving towards more genuine forms of reasoning and problem-solving and not just cramming things up and remembering the scenarios and the next steps.

Evolving Evaluation Methods: The survey paper assessed LLMs' reasoning behaviors and spotted that there is a clear need for more sophisticated and nuanced evaluation methods and not depend on just the outcome of it. Traditional error metrics just focus on task accuracy and are somewhat

insufficient for measuring the complexity of the reasoning processes. So a multifaceted approach that combines conclusion-based, rationale-based, interactive, and mechanistic evaluations, is crucial for a deeper understanding of LLMs' capabilities and limitations. Such comprehensive assessments combining multiple things can facilitate more targeted developments, steering LLM research towards models that can truly reason and think like a human brain.

Beyond Computation: Insights and Implications

The exploration into the reasoning capabilities of Large Language Models (LLMs) offers profound insights that extend well beyond computational prowess, shedding light on both the promise and limitations of these models. As we navigate through the complex process of LLMs' reasoning behavior, a critical understanding emerges: despite their remarkable achievements, LLMs' abilities to mimic human reasoning remain constrained by inherent limitations.

The Imitation of Reasoning: LLMs have demonstrated an impressive capacity to tackle a wide range of reasoning tasks, often producing results that superficially resemble human-like thought processes. However, a closer examination reveals that these models frequently rely on identifying patterns and correlations within their extensive training data, rather than engaging in the deliberate, analytical reasoning that characterizes human thought. This tendency not only limits their ability to generalize to novel scenarios but also raises questions about the depth and authenticity of their reasoning capabilities.

The Importance of Out-of-Distribution Generalization: One of the most glaring limitations of current LLMs is their struggle with out-of-distribution scenarios—situations or tasks that diverge from their training experiences. This challenge underscores a critical disparity between human and machine reasoning: the human ability to apply learned concepts flexibly and creatively in unfamiliar contexts, a feat that LLMs have yet to master. Addressing this limitation is essential for advancing LLMs towards more genuine forms of reasoning and problem-solving.

Evolving Evaluation Methods: The insights garnered from assessing LLMs' reasoning behaviors spotlight the need for more sophisticated and nuanced evaluation methods. Traditional metrics that focus solely on task accuracy are

insufficient for capturing the complexity of reasoning processes. Instead, a multifaceted approach, incorporating conclusion-based, rationale-based, interactive, and mechanistic evaluations, is vital for a deeper understanding of LLMs' capabilities and limitations. Such comprehensive assessments can facilitate more targeted developments, steering LLM research towards models that not only compute but truly reason.

Implications for Development and Understanding: The paper finds out using the surveys that there are significant implications for the development of LLMs. There is a paradigm shift in how the researchers approach the training and evaluation process of the models, urging a move towards methods that foster genuine reasoning abilities rather than just cramming things up. Furthermore, these insights contribute to our broader understanding of intelligence and reasoning of the LLM, that forces us to reconsider our definitions and expectations of the LLMs.

As the paper went on to prove that LLMs need to mimic human reasoning, it becomes clear that the path moving forward full of challenges and opportunities as well. This paper acknowledges the limitations of current models and embracing more advanced evaluation methods that are needed, we can pave the way for future advancements. The quest to bridge the gap between human and machine reasoning is not just about enhancing the results and outputs of computational LLMs; it's about the thought process behind it.

The Journey Ahead: Reasoning in the Age of AI

As we look to conclude the summary of the paper we can clearly see that the reasoning abilities of LLMs is still very much a thing in progress. This journey has revealed LLMs are impressive but ultimately there is a clear limit on the mimicry of human reasoning which emphasizes a reliance on pattern recognition over the newer(for LLM) and analytical thought processes that humans possess.

Future Directions for Research: Moving forward, research must pivot and truly focus on enhancing LLMs' abilities in generalizing the problem and developing evaluation methodologies that capture the depth of reasoning instead of cramming up the scenarios and their next steps to narrow down the gap between human and artificial intelligence.

In essence, the road ahead for LLMs and AI, in general, is one of immense possibility, tempered by the need for our thoughtful consideration of the implications of the advancements. As the researchers continue to push the boundaries of what AI can achieve, our guiding light should always be the betterment of humanity and focus on the chain of thought process of the LLM.