

“Reasoning”, Reasoning and Argumentation

Una breve panoramica critica su LLM

Language Technologies Lab, Nov 20th, 2025

PhD. Federico Ruggeri

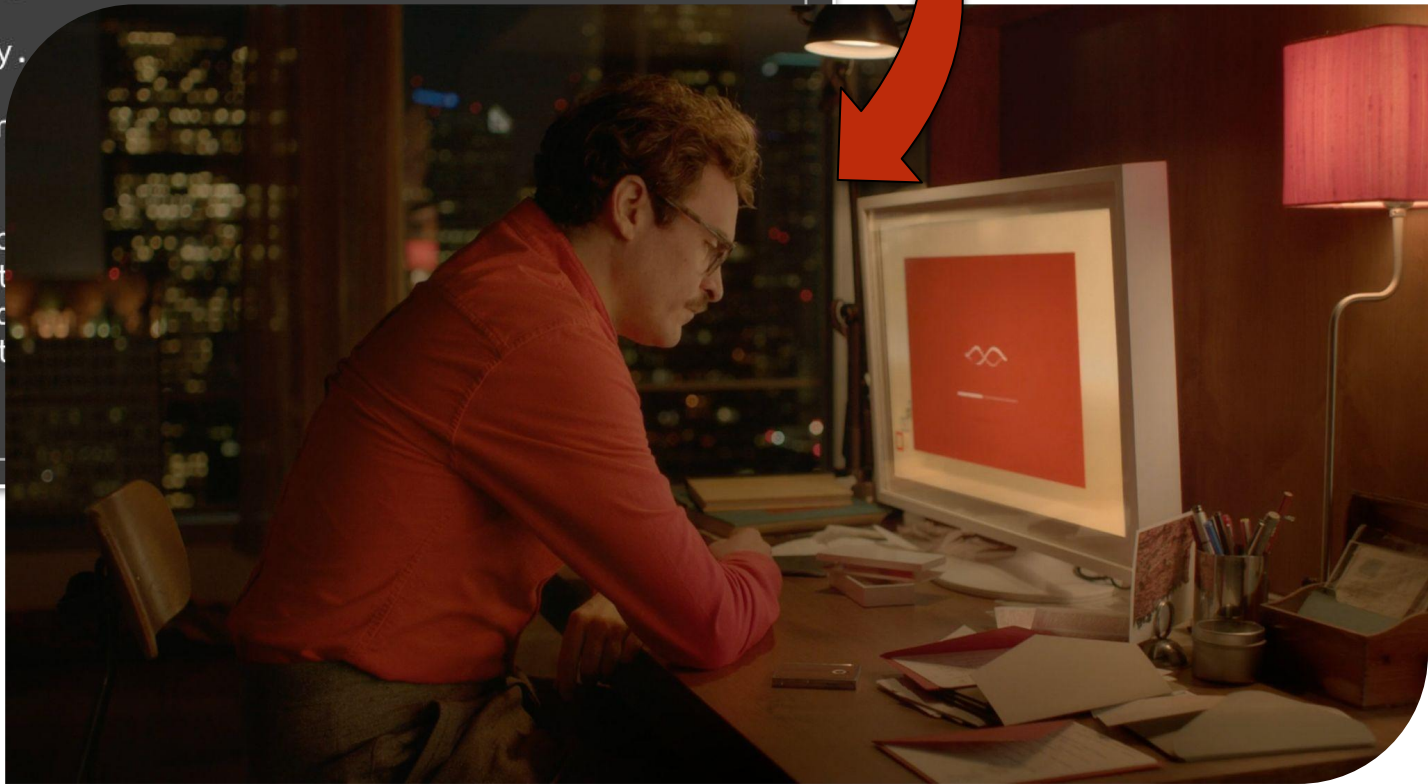
Perché ci interessa il tema

Eliza (1996)

```
> Hello, I am Eliza.  
* Hey!  
> Don't you ever say Hello?  
* Hello!  
> How are you today.. What would you like to discuss?  
* I'm depressed.  
> Do you enjoy being depressed?  
* No, I don't.  
> Why no?  
* Because I want to be happy.  
> Why do you want to be happy?  
* Because I'm tired of being depressed.  
> That is quite interesting.  
* I'm not sure█
```

Eliza (1996)

> Hello, I am Eliza.
* Hey!
> Don't you ever say Hello?
* Hello!
> How are you today.
* I'm depressed.
> Do you enjoy being
* No, I don't.
> Why no?
* Because I want to
> Why do you want to
* Because I'm tired
> That is quite int
* I'm not sure█





Annual Meeting of the Association for Computational Linguistics (2024)

Volumes

- Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 865 papers
- Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 77 papers
- Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations) 39 papers
- Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop) 47 papers
- Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts) 7 papers
- Findings of the Association for Computational Linguistics: ACL 2024 976 papers



Annual Meeting of the Association for Computational Linguistics (2024)

Volumes

- Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 865 papers
- Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 77 papers
- Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations) 39 papers
- Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop) 47 papers
- Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts) 7 papers
- Findings of the Association for Computational Linguistics: ACL 2024 976 papers



Annual Meeting of the Association for Computational Linguistics (2025)

Volumes

- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 1603 papers
- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 98 papers
- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations) 65 papers
- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop) 87 papers
- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts) 9 papers
- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track) 110 papers
- Findings of the Association for Computational Linguistics: ACL 2025 1388 papers



ACL Anthology

[News](#)

[FAQ](#)

[Corrections](#)

[Submissions](#)

[Github](#)

LLM

1,039



A



ACL Anthology

[News](#)

[FAQ](#)

[Corrections](#)

[Submissions](#)

[Github](#)

reasoning

1,353



Annual Meeting of the Association for Computational Linguistics (2025)

Volumes

- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 1603 papers
- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 98 papers
- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations) 65 papers
- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop) 87 papers
- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts) 9 papers
- Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track) 110 papers
- Findings of the Association for Computational Linguistics: ACL 2025 1388 papers



The 2025 Conference on Empirical Methods in Natural Language Processing

Suzhou, China

November 4-9, 2025

Volumes

- Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing **1810 papers**
- Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations **78 papers**
- Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track **193 papers**
- Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts **8 papers**
- Findings of the Association for Computational Linguistics: EMNLP 2025 **1406 papers**

Si
N

The 2025 Conference on Empirical Methods in Natural Language Processing

Suzhou, China

November 4-9, 2025

Volumes

- Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing 1810 papers
- Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations 78 papers
- Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track 193 papers
- Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts 8 papers
- Findings of the Association for Computational Linguistics: EMNLP 2025 1406 papers



[Home](#) [Dates](#) [Program ▼](#) [Call for Papers](#) [Area Chairs](#) [Guides ▼](#) [Sponsors](#) [FAQ](#)

CONFERENCE ON LANGUAGE MODELING

University of Pennsylvania
Philadelphia, PA
October 7-9, 2024



Prompting and Reasoning

If a train is moving at 60 mph and travels for 3 hours, how far does it go?

(Plain) Prompting

The train travels 180 miles.

Plain response

If a train is moving at 60 mph and travels for 3 hours, how far does it go?

The train travels 180 miles.

Plain response

To determine the distance traveled,
use the formula:

Distance = Speed \times Time

Given that the speed is 60 mph and
the time is 3 hours:

Distance = 60 mph \times 3 hours = 180 miles

So, the train travels 180 miles.

**Response with intermediate
reasoning steps**

(Reason) Prompting

ChatGPT o1 ▾

↑ Share



What is the last digit of 7^{999}



Reasoned about last digit of powers for a couple of seconds

1) Intermediate reasoning chain that is not explicitly shown to the user

ChatGPT o1

Share



What is the last digit of 7^{999}



Reasoned about last digit of powers for a couple of seconds

To find the last digit of 7^{999} , note that the last digits of powers of 7 repeat in a cycle of length 4:

$$\begin{aligned} 7^1 &= 7 && \text{(last digit 7)} \\ 7^2 &= 49 && \text{(last digit 9)} \\ 7^3 &= 343 && \text{(last digit 3)} \\ 7^4 &= 2401 && \text{(last digit 1)} \end{aligned}$$

Then it repeats: 7, 9, 3, 1, 7, 9, 3, 1, ...

To determine which of these four digits appears for 7^{999} , we look at $999 \bmod 4$:

$$999 \div 4 = 249 \text{ remainder } 3.$$

Since the remainder is 3, we take the third digit in the cycle (7, 9, 3, 1), which is 3.

1) Intermediate reasoning chain that is not explicitly shown to the user

2) Intermediate reasoning steps that are shown as part of the answer

ChatGPT o1

Share



What is the last digit of 7^{999}



Reasoned about last digit of powers for a couple of seconds

To find the last digit of 7^{999} , note that the last digits of powers of 7 repeat in a cycle of length 4:

$$\begin{aligned} 7^1 &= 7 && \text{(last digit 7)} \\ 7^2 &= 49 && \text{(last digit 9)} \\ 7^3 &= 343 && \text{(last digit 3)} \\ 7^4 &= 2401 && \text{(last digit 1)} \end{aligned}$$

Then it repeats: 7, 9, 3, 1, 7, 9, 3, 1, ...

To determine which of these four digits appears for 7^{999} , we look at $999 \bmod 4$:

$$999 \div 4 = 249 \text{ remainder } 3.$$

Since the remainder is 3, we take the third digit in the cycle (7, 9, 3, 1), which is 3.

Hence, the last digit of 7^{999} is **3**.

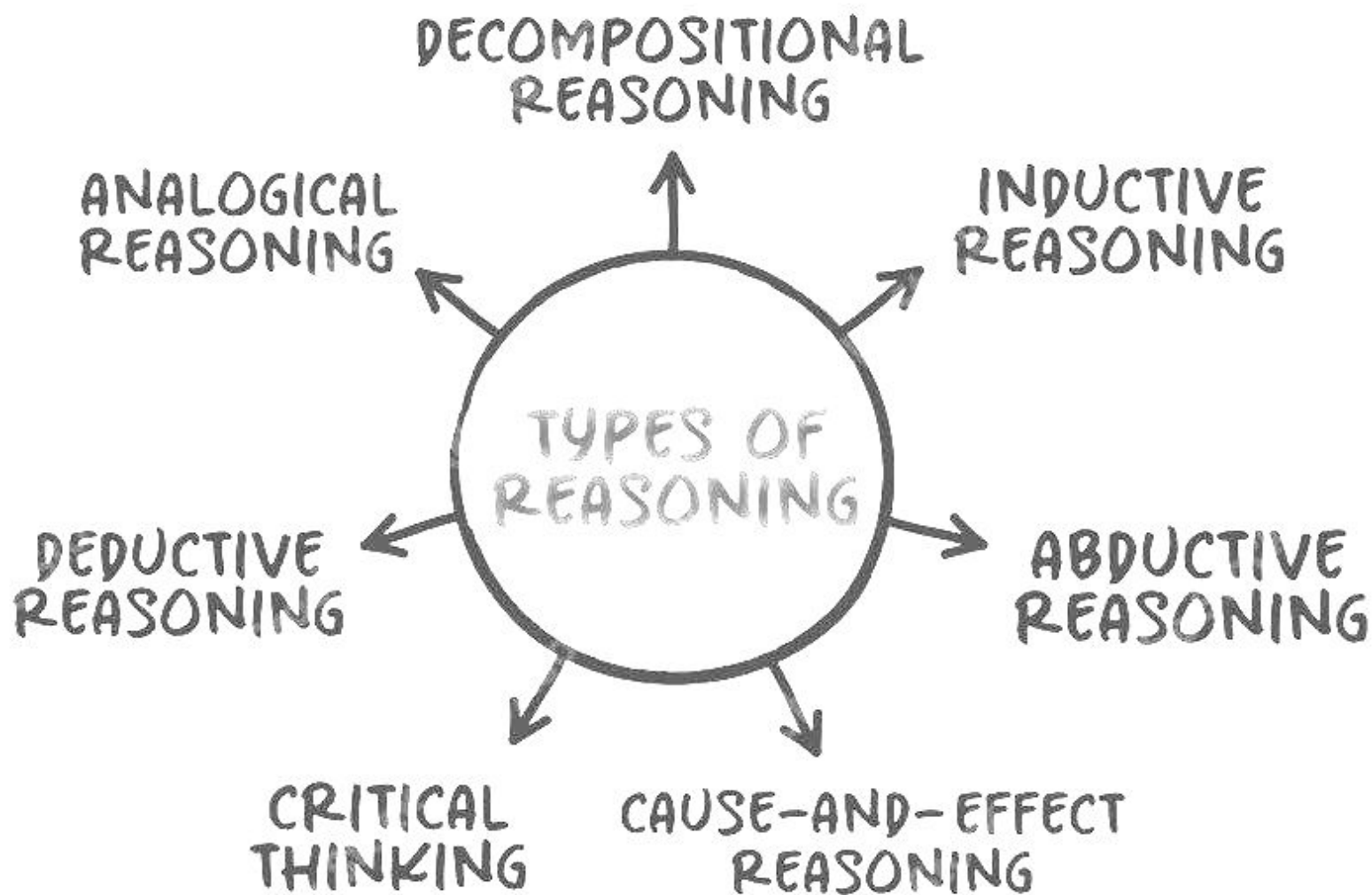


1) Intermediate reasoning chain that is not explicitly shown to the user

2) Intermediate reasoning steps that are shown as part of the answer

The main answer

Punto 1: Reasoning



Assessing the Reasoning Capabilities of LLMs in the context of Evidence-based Claim Verification

John Dougrez-Lewis^{1,*}, Mahmud Elahi Akhter^{2,*}, Federico Ruggeri³, Sebastian Löbbers², Yulan He^{4,5}, Maria Liakata^{2,5}

¹University of Warwick, UK, ²Queen Mary University of London, UK
³University of Bologna, Italy, ⁴King's College London, UK
⁵The Alan Turing Institute, UK

j.dougrez-lewis@warwick.ac.uk, federico.ruggeri@unibo.it
{m.akhter, s.lobbers, m.liakata}@qmul.ac.uk

Abstract

Although LLMs have shown great performance on Mathematics and Coding related reasoning tasks, the reasoning capabilities of LLMs regarding other forms of reasoning are still an open problem. Here, we examine the issue of reasoning from the perspective of claim verification. We propose a framework designed to break down any claim paired with evidence into atomic reasoning types that are necessary for verification. We use this framework to create RECV, the first claim verification benchmark, incorporating real-world claims, to assess the deductive and abductive reasoning capabilities of LLMs. The benchmark comprises of three datasets, covering reasoning problems of increasing complexity. We evaluate three state-of-the-art proprietary LLMs under multiple prompt settings. Our results show that while LLMs can address deductive reasoning problems, they consistently fail in cases of abductive reasoning. Moreover, we observe that enhancing LLMs with rationale generation is not always beneficial. Nonetheless, we find that generated rationales are semantically similar to those provided by humans, especially in deductive reasoning cases.

1 Introduction

Large Language Models (LLMs) have shown remarkable proficiency in complex tasks where reasoning capabilities, such as logical deduction and semantic comparison, are paramount. Notable examples include solving MBA exams (Terwiesch et al., 2023; Nori et al., 2023), performing quantitative reasoning (Lewkowycz et al., 2022), and communication games (Bakhtin et al., 2022; Gandhi et al., 2022).

that LLMs possess Theory of Mind (ToM) capabilities (Bubeck et al., 2023; Kosinski, 2023) were shown to be inaccurate (Ullman, 2023; Sileo and Lernould, 2023). In particular, despite appearing to manifest some form of ToM capabilities, LLMs mostly rely on shallow heuristics and spurious correlations (Shapira et al., 2023). Additionally, preliminary observations of emergent reasoning capabilities (Wei et al., 2022) were subsequently attributed to metric choice (Schaeffer et al., 2023), in-context learning (Lu et al., 2023b), and shortcuts (Kavumba et al., 2019).

These findings motivate the need for further research on the reasoning capabilities of LLMs, especially in high-stake real-world applications, where research on this topic is in its infancy. A notable example is fact-checking, where LLMs are considered to hold great potential for increased productivity even if at the same time they also facilitate bad actors in the proliferation of misinformation (Guo et al., 2023). Verifying information is challenging since models require both accurate veracity classification and strong rationale generation to be effective (Schlichtkrull et al., 2023). It is thus essential to understand the reasoning capabilities and limitations of LLMs in the context of fact-checking. In particular, we extend the current discussion around the reasoning abilities of LLMs, focusing on their ability to verify real-world claims.

In this work, we first propose a framework for breaking down complex claims into atomic reasoning steps. The motivation behind this is the lack of uniform terminology around reasoning capabilities. Most prominent examples of this are the

Assessing the Reasoning Capabilities of LLMs in the context of Evidence-based Claim Verification

John Dougrez-Lewis^{1,*}, Mahmud Elahi Akhter^{2,*}, Federico Ruggeri³, Sebastian L  bbers², Yulan He^{4,5}, Maria Liakata^{2,5}

¹University of Warwick, UK, ²Queen Mary University of London, UK
³University of Bologna, Italy, ⁴King's College London, UK
⁵The Alan Turing Institute, UK

j.dougrez-lewis@warwick.ac.uk, federico.ruggeri6@unib
 {m.akhter, s.lobbers, m.liakata}@qmul.ac.uk

Abstract

Although LLMs have shown great performance on Mathematics and Coding related reasoning tasks, the reasoning capabilities of LLMs regarding other forms of reasoning are still an open problem. Here, we examine the issue of reasoning from the perspective of claim verification. We propose a framework designed to break down any claim paired with evidence into atomic reasoning types that are necessary for verification. We use this framework to create RECV, the first claim verification benchmark incorporating real-world claims, to assess the deductive and abductive reasoning capabilities of LLMs. The benchmark comprises of three datasets, covering reasoning problems of increasing complexity. We evaluate three state-of-the-art proprietary LLMs under multiple prompt settings. Our results show that while LLMs can address deductive reasoning problems, they consistently fail in cases of abductive reasoning. Moreover, we observe that enhancing LLMs with rationale generation is not always beneficial. Nonetheless, we find that generated rationales are semantically similar to those provided by humans, especially in deductive reasoning cases.

1 Introduction

Large Language Models (LLMs) have shown remarkable proficiency in complex tasks where reasoning capabilities, such as logical deduction and semantic comparison, are paramount. Notable examples include solving MBA exams (Terwiesch et al., 2023; Nori et al., 2023), performing quantitative reasoning (Lewkowycz et al., 2022), and communication games (Bakhtin et al., 2022), and

that LLMs possess Theory of Mind (ToM) capabilities (Bubeck et al., 2023; Kosinski, 2023; L  moult, 2023). In particular, despite appearing to manifest some form of ToM capabilities, mostly rely on shallow heuristics and spurious relations (Shapira et al., 2023). Additional preliminary observations of emergent reasoning capabilities (Wei et al., 2022) were subsequently attributed to metric choice (Schaeffer et al., in-context learning (Lu et al., 2023b), and cuts (Kavumba et al., 2019).

These findings motivate the need for further research on the reasoning capabilities of LLMs, especially in high-stake real-world applications, where fact-checking, where LLMs are expected to hold great potential for increased productivity even if at the same time they also facilitate the proliferation of misinformation (Schlichtkrull et al., 2023). Verifying information is challenging since models require both accurate verification and strong rationale generation to be able to understand the reasoning capabilities and limitations of LLMs in the context of fact-checking. In particular, we extend the current discussion about the reasoning abilities of LLMs, focusing on the ability to verify real-world claims.

In this work, we first propose a framework for breaking down complex claims into atomic reasoning steps. The motivation behind this is the lack of uniform terminology around reasoning capabilities in the literature.

Measuring what Matters: Construct Validity in Large Language Model Benchmarks

Andrew M. Bean^{1,*}, Ryan Othniel Kearns¹, Angelika Romanou², Franziska Sofia Hafner¹, Harry Mayne¹

Jan Batzner^{3,4}, Negar Foroutan², Chris Schmitz⁵, Karolina Korgul¹, Hunar Batra¹, Oishi Deb¹, Emma Beharry⁶, Cornelius Emde¹, Thomas Foster¹, Anna Gausen⁷, Maria Grandury^{8,9}, Simeng Han¹⁰, Valentin Hofmann^{11,12}, Lujain Ibrahim¹, Hazel Kim¹, Hannah Rose Kirk^{1,7}, Fangru Lin¹, Gabrielle Kaili-May Liu¹⁰, Lennart Luetgau⁷, Jabez Magomere¹, Anna Sotnikova², Yushi Yang¹, Yilun Zhao¹⁰, Jonathan Rystrom¹, Adel Bibi¹, Antoine Bosselut², Ronald Clark¹, Arman Cohan¹⁰, Jakob Foerster¹, Yarin Gal^{1,7}, Scott A. Hale^{1,13}, Inioluwa Deborah Raji¹⁴, Christopher Summerfield^{1,7}, Philip H.S. Torr¹, Cozmin Ududec⁷, Luc Rocher¹, Adam Mahdi^{1*}

¹University of Oxford, ²EPFL, ³Weizenbaum Institute Berlin
⁴Technical University Munich, ⁵Centre for Digital Governance, Hertie School
⁶Stanford University, ⁷UK AI Security Institute, ⁸SomosNLP
⁹Universidad Polit  cnica de Madrid, ¹⁰Yale University, ¹¹Allen Institute for AI
¹²University of Washington, ¹³Meedan, ¹⁴UC Berkeley

Abstract

Evaluating large language models (LLMs) is crucial for both assessing their capabilities and identifying safety or robustness issues prior to deployment. Reliably measuring abstract and complex phenomena such as ‘safety’ and ‘robustness’ requires strong *construct validity*, that is, having measures that represent what matters to the phenomenon. With a team of 29 expert reviewers, we conduct a systematic review of 445 LLM benchmarks from leading conferences in natural language processing and machine learning. Across the reviewed articles, we find patterns related to the measured phenomena, tasks, and scoring metrics which undermine the validity of the resulting claims. To address these shortcomings, we provide eight key recommendations and detailed actionable guidance to researchers and practitioners in developing LLM benchmarks.

1 Introduction

Benchmarks and evaluations play a critical role in helping determine which models are the most effective for a given task. Research [1, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100]

Punto 2: “Reasoning”?



Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang^{1,2} Elena Gribovskaya¹ Nora Kassner¹ Mor Geva^{3,4*} Sebastian Riedel^{1,2*}

¹Google DeepMind ²UCL ³Google Research ⁴Tel Aviv University

{soheeyang, egribovskaya, norakassner, pipek, srriedel}@google.com



Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang^{1,2} Elena Gribovskaya¹ Nora Kassner¹ Mor Geva^{3,4*} Sebastian Riedel^{1,2*}

¹Google DeepMind ²MIT ³Google Research ⁴Bar Ilan University

Not All LLM Reasoners Are Created Equal

Arian Hosseini
Mila

arian.hosseini9@gmail.com

Alessandro Sordoni
Mila, Microsoft Research

Daniel Toyama
Google DeepMind

Aaron Courville
Mila

Rishabh Agarwal
Mila, Google DeepMind



Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang^{1,2} Elena Gribovskaya¹ Nora Kassner¹ Mor Geva^{3,4*} Sebastian Riedel^{1,2*}

¹Google Research, ²MIT, ³Google Research, ⁴Bar Ilan University

Not All LLM Reasoners Are Created Equal

Arian Hosseini

Alessandro Sordoni

Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation.

Jinglong Gao Xiao Ding* Bing Qin Ting Liu

Research Center for Social Computing and Information Retrieval

Harbin Institute of Technology, China

{jlgao, xding, qinb, tliu}@ir.hit.edu.cn



Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang^{1,2} Elena Gribovskaya¹ Nora Kassner¹ Mor Geva^{3,4*} Sebastian Riedel^{1,2*}

¹Google DeepMind ²EPFL ³Google Research ⁴MIT Artificial Intelligence

Not All LLM Reasoners Are Created Equal

Arian Hosseini

Alessandro Sordani

Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation.

Jinglong Gao Xiao Ding* Bing Qin Ting Liu

Research Center for Social Computing and Information Retrieval

The Illusion of Thinking:

Understanding the Strengths and Limitations of Reasoning Models
via the Lens of Problem Complexity

Parshin Shojae*[†]
Maxwell Horton

Iman Mirzadeh*
Samy Bengio

Keivan Alizadeh
Mehrdad Farajtabar



Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang^{1,2} Elena Gribovskaya¹ Nora Kassner¹ Mor Geva^{3,4*} Sebastian Riedel^{1,2*}

Not All LLM Reasoners Are Created Equal

Arian Hosseini

Alessandro Sordoni

Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation.

Jinglong Gao Xiao Ding* Bing Qin Ting Liu

Research Center for Social Computing and Information Retrieval

The Illusion of Thinking:

Understanding the Strengths and Limitations of Reasoning Models
via the Lens of Problem Complexity

Parshin Shojaei*[†]
Maxwell Horton

Iman Mirzadeh*
Samy Bengio

Keivan Alizadeh
Mehrdad Farajtabar



Emergent Analogical Reasoning in Large Language Models

Taylor Webb^{1,*}, Keith J. Holyoak¹, and Hongjing Lu^{1,2}

¹Department of Psychology

²Department of Statistics

University of California, Los Angeles, CA, USA



Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang^{1,2} Elena Gribovskaya¹ Nora Kassner¹ Mor Geva^{3,4*} Sebastian Riedel^{1,2*}

¹Google DeepMind ²ICML ³Google Research ⁴MIT Artificial Intelligence

Not All LLM Reasoners Are Created Equal

Arian Hosseini

Alessandro Sordoni

Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation.

Jinglong Gao Xiao Ding* Bing Qin Ting Liu

Research Center for Social Computing and Information Retrieval

The Illusion of Thinking:

Understanding the Strengths and Limitations of Reasoning Models
via the Lens of Problem Complexity

Parshin Shojaei*
Maxwell Horton

Iman Mirzadeh*
Samy Bengio

Keivan Alizadeh
Mehrdad Farajtabar



Emergent Analogical Reasoning in Large Language Models

Taylor Webb^{1,*}, Keith J. Holyoak¹, and Hongjing Lu^{1,2}

¹Department of Psychology

LARGE LANGUAGE MODELS AS ANALOGICAL REASONERS

Michihiro Yasunaga,^{1,2} Xinyun Chen,¹ Yujia Li,¹ Panupong Pasupat,¹ Jure Leskovec,²
Percy Liang,² Ed H. Chi,¹ Denny Zhou¹

¹Google DeepMind ²Stanford University
myasu@cs.stanford.edu, {xinyunchen,dennyzhou}@google.com



Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang^{1,2} Elena Gribovskaya¹ Nora Kassner¹ Mor Geva^{3,4*} Sebastian Riedel^{1,2*}

¹Google DeepMind ²UC Berkeley ³Google Research ⁴Tel Aviv University

Not All LLM Reasoners Are Created Equal

Arian Hosseini

Alessandro Sordoni

Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation.

Jinglong Gao Xiao Ding* Bing Qin Ting Liu

Research Center for Social Computing and Information Retrieval

The Illusion of Thinking:

Understanding the Strengths and Limitations of Reasoning Models
via the Lens of Problem Complexity

Parshin Shojaei*
Maxwell Horton

Iman Mirzadeh*
Samy Bengio

Keivan Alizadeh
Mehrdad Farajtabar



Emergent Analogical Reasoning in Large Language Models

Taylor Webb^{1,*}, Keith J. Holyoak¹, and Hongjing Lu^{1,2}

¹Department of Psychology

LARGE LANGUAGE MODELS AS ANALOGICAL REASONERS

Michihiro Yasunaga,^{1,2} Xinyun Chen,¹ Yujia Li,¹ Panupong Pasupat,¹ Jure Leskovec,²
Percy Liang,² Ed H. Chi,¹ Denny Zhou¹

¹Google DeepMind ²Stanford University

Testing the General Deductive Reasoning Capacity of Large Language Models Using OOD Examples

Abulhair Saparov[†]

Richard Yuanzhe Pang[†]

Vishakh Padmakumar[†]

Nitish Joshi[†]

Seyed Mehran Kazemi^Δ

Najoung Kim^{Δ,β,*}

He He^{†,*}



VS.

Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang^{1,2} Elena Gribovskaya¹ Nora Kassner¹ Mor Geva^{3,4*} Sebastian Riedel^{1,2*}

Not All LLM Reasoners Are Created Equal

Arian Hosseini

Alessandro Sordoni

Is ChatGPT a Good Causal Reasoner? A Comprehensive Evaluation.

Jinglong Gao Xiao Ding* Bing Qin Ting Liu

Research Center for Social Computing and Information Retrieval

The Illusion of Thinking:

Understanding the Strengths and Limitations of Reasoning Models
via the Lens of Problem Complexity

Parshin Shojae*[†]
Maxwell Horton

Iman Mirzadeh*
Samy Bengio

Keivan Alizadeh
Mehrdad Farajtabar

Emergent Analogical Reasoning in Large Language Models

Taylor Webb^{1,*}, Keith J. Holyoak¹, and Hongjing Lu^{1,2}

LARGE LANGUAGE MODELS AS ANALOGICAL REASONERS

Michihiro Yasunaga,^{1,2} Xinyun Chen,¹ Yujia Li,¹ Panupong Pasupat,¹ Jure Leskovec,²
Percy Liang,² Ed H. Chi,¹ Denny Zhou¹

¹Google DeepMind ²Stanford University

Testing the General Deductive Reasoning Capacity of Large Language Models Using OOD Examples

Abulhair Saparov[†]

Richard Yuanzhe Pang[†]

Vishakh Padmakumar[†]

Nitish Joshi[†]

The Illusion of the Illusion of Thinking

A Comment on Shojae et al. (2025)

A. Lawsen*

Measuring what Matters: Construct Validity in Large Language Model Benchmarks

Andrew M. Bean^{1*} Ryan Othniel Kearns¹ Angelika Romanou²
Franziska Sofia Hafner¹ Harry Mayne¹

Jan Batzner^{3,4} Negar Foroutan² Chris Schmitz⁵ Karolina Korgul¹ Hunar Batra¹
Oishi Deb¹ Emma Beharry⁶ Cornelius Emde¹ Thomas Foster^{1,12} Anna Gausen⁷
Maria Grandury^{8,9} Simeng Han¹⁰ Valentin Hofmann^{11,12} Lujain Ibrahim¹
Gabielle Kaili-May Liu¹⁰ Hannah Rose Kirk^{1,7} Fangru Lin¹
Anna Sotnikova² Yushi Yang¹ Jabez Magomere¹ Jonathan Rystrom¹
Adel Bibi¹ Antoine Bosselut² Ronald Clark¹ Arman Cohan¹⁰ Jakob Foerster¹
Yarin Gal^{1,7} Scott A. Hale^{1,13} Inioluwa Deborah Raji¹⁴ Christopher Summerfield^{1,7}
Philip H.S. Torr¹ Cozmin Ududec⁷ Luc Rocher¹ Adam Mahdi^{1*}

¹University of Oxford ²EPFL ³Weizenbaum Institute Berlin
⁴Technical University Munich ⁵Centre for Digital Governance, Hertie School
⁶Stanford University ⁷UK AI Security Institute ⁸SomosNLP
⁹Universidad Politécnica de Madrid ¹⁰Yale University ¹¹Allen Institute for AI
¹²University of Washington ¹³Meedan ¹⁴UC Berkeley

Abstract

Evaluating large language models (LLMs) is crucial for both assessing their capabilities and identifying safety or robustness issues prior to deployment. Reliably measuring abstract and complex phenomena such as ‘safety’ and ‘robustness’ requires strong *construct validity*, that is, having measures that represent what matters to the phenomenon. With a team of 29 expert reviewers, we conduct a systematic review of 445 LLM benchmarks from leading conferences in natural language processing and machine learning. Across the reviewed articles, we find patterns related to the measured phenomena, tasks, and scoring metrics which undermine the validity of the resulting claims. To address these shortcomings, we provide eight key recommendations and detailed actionable guidance to researchers and practitioners in developing LLM benchmarks.

1 Introduction

Benchmarks and evaluations play a critical role in the development of LLMs. They help determine which model improvements are worthwhile and guide future research [1, 2]. Creating a benchmark requires careful consideration of the concrete tasks and metrics that are used to evaluate the ‘intelligence’ of the model. In this paper, we focus on the ‘intelligence’ of the model, which is the ability to perform tasks that require reasoning, understanding, and problem-solving.

Don't really know
what
benchmarks
are measuring!

Punto 3: Argumentation



Figure 1: Sketch of a possible realization of the idea of reasonable parrots we advocate for: a conversational technology that fosters critical thinking and deliberation.



Figure 1: Sketch of a possible realization of the idea of reasonable parrots we advocate for: a conversational technology that fosters critical thinking and deliberation.

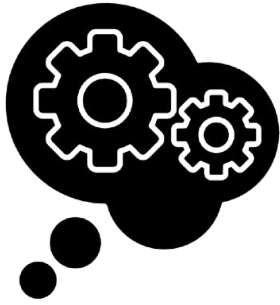
“Ad populum fallacy”

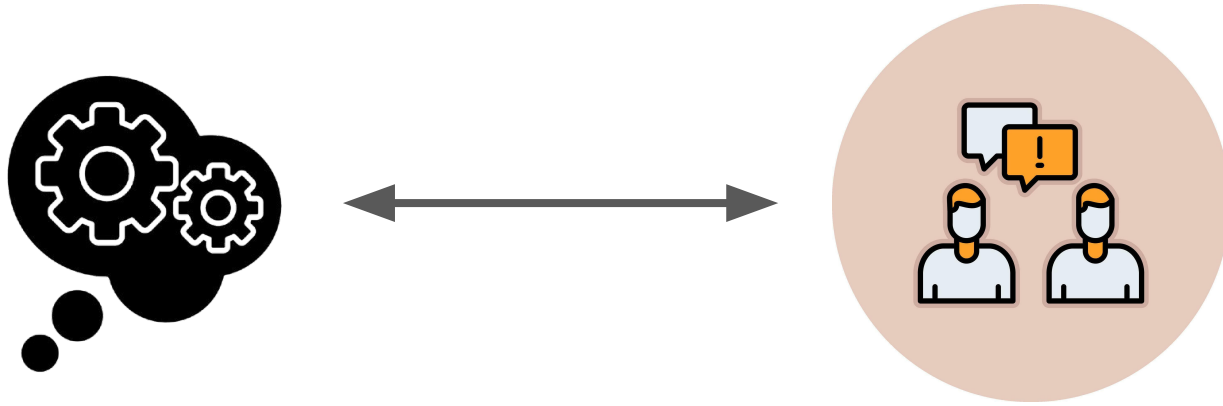


Figure 1: Sketch of a possible realization of the idea of reasonable parrots we advocate for: a conversational technology that fosters critical thinking and deliberation.

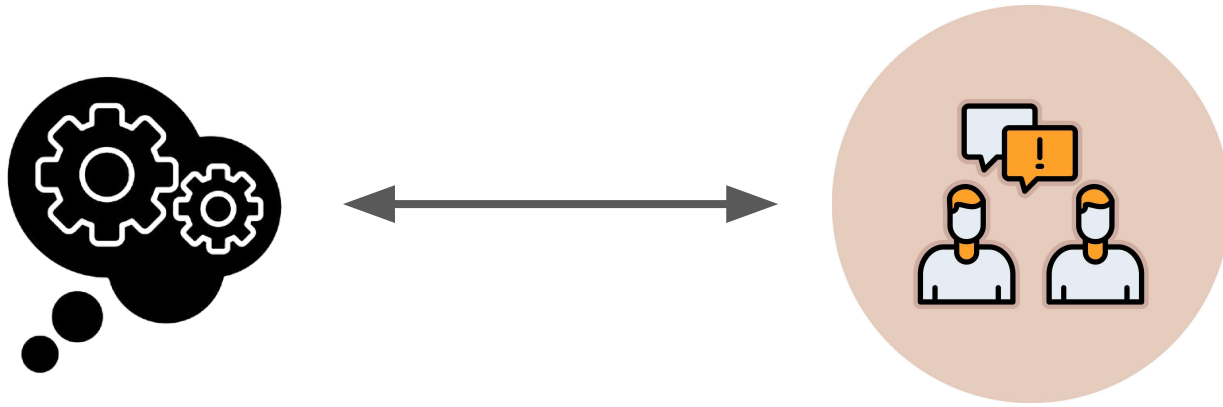
“Ad populum fallacy”

Argue by *design*





Argumentation represents, analyzes and evaluates the process of reaching conclusions on the basis of available information.



Argumentation represents, analyzes and evaluates the process of reaching conclusions on the basis of available information.

We can evaluate LLMs based on their ability to address argument mining tasks.

Proposal for ArgMining 2026, the 13th Workshop on Argument Mining and Reasoning

**Mohamed Elaraby, Annette Hautli-Janisz, John Lawrence,
Elena Musi, Julia Romberg, and Federico Ruggeri**

- Explainability in argument mining through reasoning;

Proposal for ArgMining 2026, the 13th Workshop on Argument Mining and Reasoning

**Mohamed Elaraby, Annette Hautli-Janisz, John Lawrence,
Elena Musi, Julia Romberg, and Federico Ruggeri**

- Explainability in argument mining through reasoning;
- Modeling, assessing, and critically reflecting on the argumentation capabilities of LLMs;

Proposal for ArgMining 2026, the 13th Workshop on Argument Mining and Reasoning

**Mohamed Elaraby, Annette Hautli-Janisz, John Lawrence,
Elena Musi, Julia Romberg, and Federico Ruggeri**

- Explainability in argument mining through reasoning;
- Modeling, assessing, and critically reflecting on the argumentation capabilities of LLMs;
- Novel benchmarks in argument mining that cater to the recent developments in LLM reasoning as a whole;

Proposal for ArgMining 2026, the 13th Workshop on Argument Mining and Reasoning

**Mohamed Elaraby, Annette Hautli-Janisz, John Lawrence,
Elena Musi, Julia Romberg, and Federico Ruggeri**

- Explainability in argument mining through reasoning;
- Modeling, assessing, and critically reflecting on the argumentation capabilities of LLMs;
- Guidelines for assessing and documenting the reasoning process(es) reflected in benchmarks;
- Novel benchmarks in argument mining that cater to the recent developments in LLM reasoning as a whole;

Proposal for ArgMining 2026, the 13th Workshop on Argument Mining and Reasoning

**Mohamed Elaraby, Annette Hautli-Janisz, John Lawrence,
Elena Musi, Julia Romberg, and Federico Ruggeri**

- Explainability in argument mining through reasoning;
- Modeling, assessing, and critically reflecting on the argumentation capabilities of LLMs;
- Guidelines for assessing and documenting the reasoning process(es) reflected in benchmarks;
- Novel benchmarks in argument mining that cater to the recent developments in LLM reasoning as a whole;
- Reflection on ethical aspects and societal impact of argument mining and LLM reasoning.

Grazie per l'attenzione!
