

# **“Reasoning”, Reasoning and Argumentation**

## **Una breve panoramica critica su LLM**

---

*Language Technologies Lab, Nov 20<sup>th</sup>, 2025*

PhD. Federico Ruggeri

---

# Perché ci interessa il tema

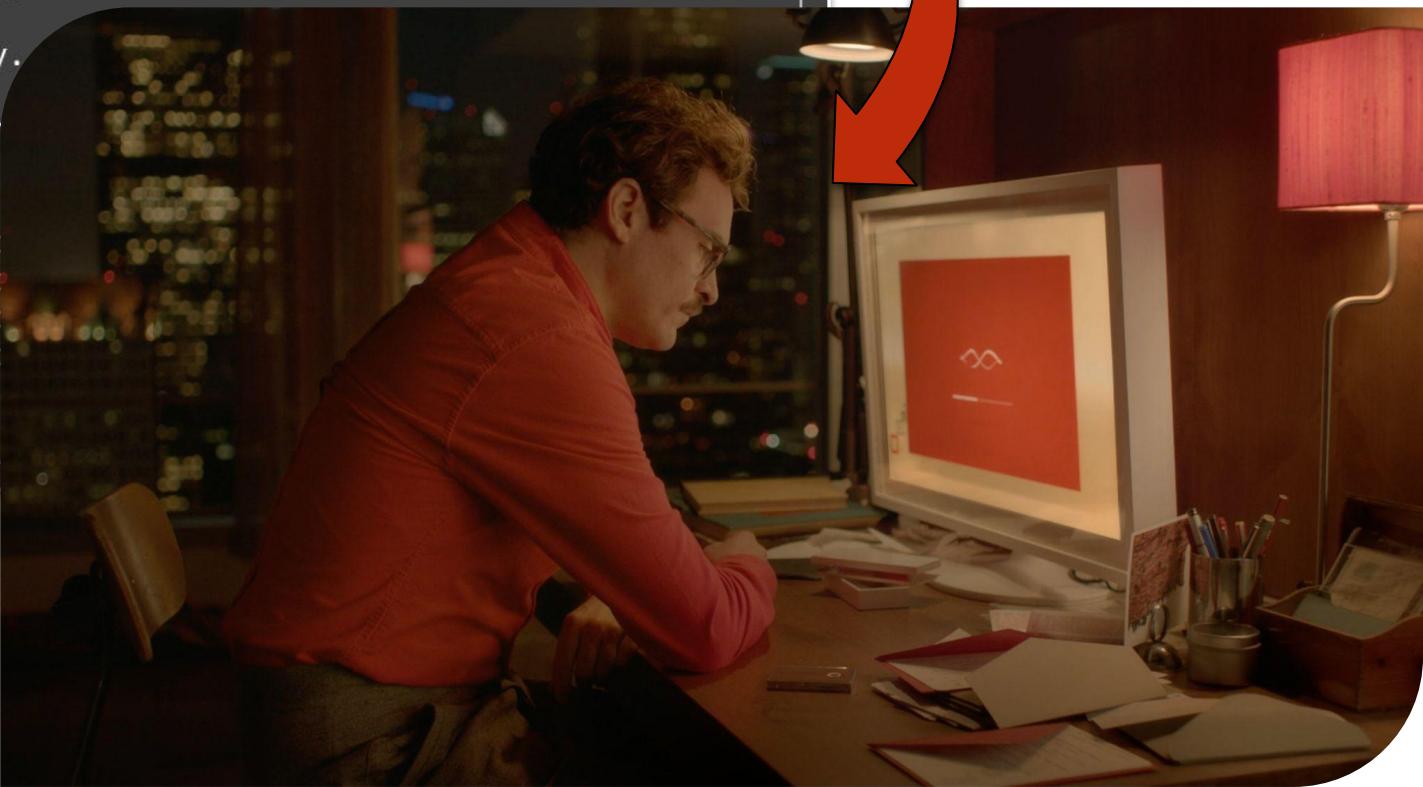
---

## 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 od 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 beir  
* 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



A



## 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](#)

# 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

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:

$$\text{Distance} = \text{Speed} \times \text{Time}$$

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

$$\text{Distance} = 60 \text{ mph} \times 3 \text{ hours} = 180 \text{ miles}$$

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  $7^{999}$

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



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:

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

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.

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



ChatGPT o1 ✓

Share



What is the last digit of  $7^{999}$

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



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:

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

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

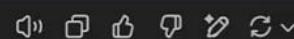
To determine which of these four digits appears for  $7^{999}$ , we look at  $999 \bmod 4$ :

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

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.

The main answer

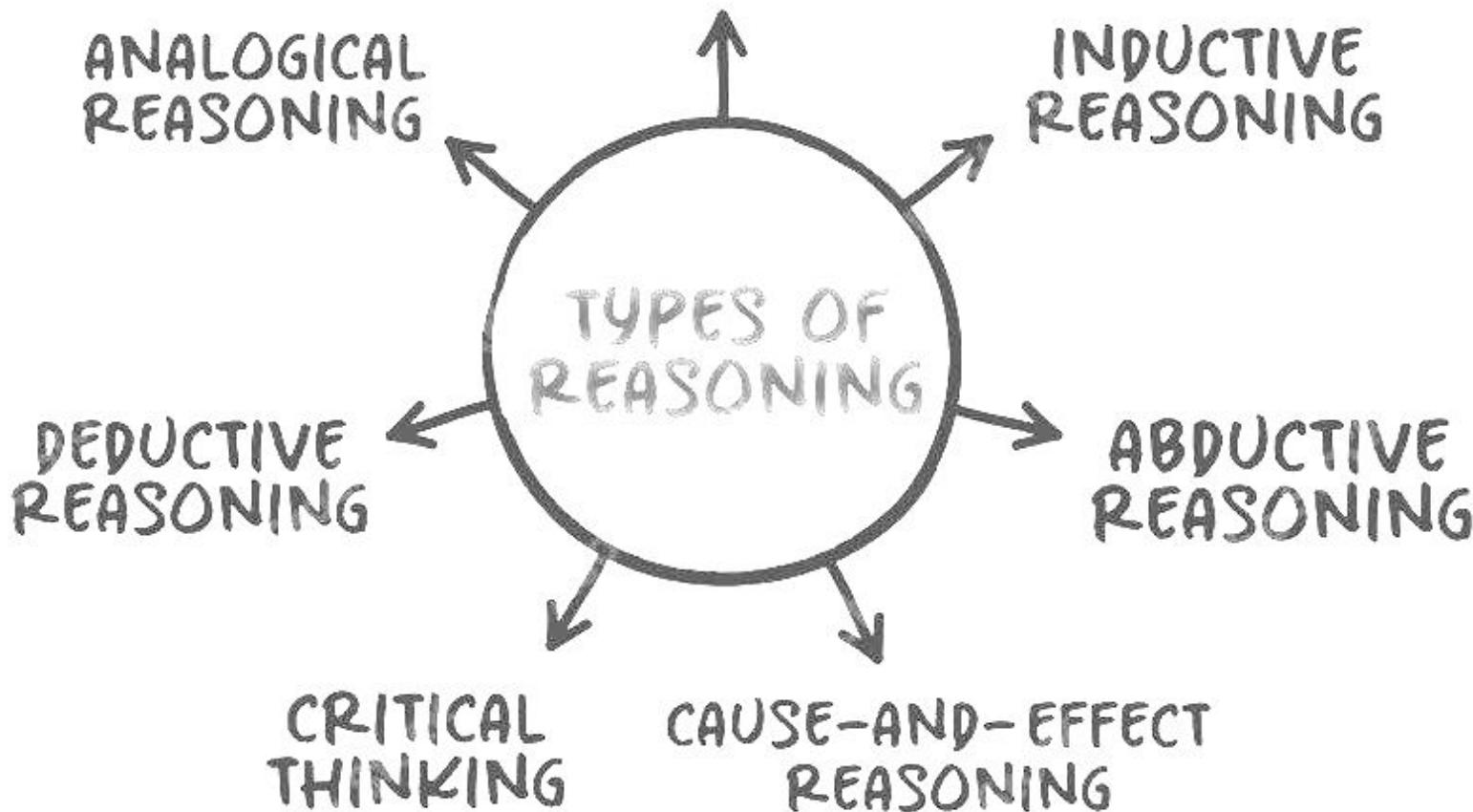


---

# Punto 1: Reasoning

---

# DECOMPOSITIONAL REASONING



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

John Dougrez-Lewis<sup>1,\*</sup>, Mahmud Elahi Akhter<sup>2,\*</sup>, Federico Ruggeri<sup>3</sup>,  
Sebastian Löbbers<sup>2</sup>, Yulan He<sup>4,5</sup>, Maria Liakata<sup>2,5</sup>

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

j.dougrez-lewis@warwick.ac.uk, yulan.he@kcl.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, 2023), passing professional medical tests (Kung et al., 2023; Nori et al., 2023), performing quantitative reasoning (Lewkowycz et al., 2022), and communication games (Bakhtin et al., 2022, and et al., 2023; 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 short-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 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. Most prominent are the terms

# Measuring what Matters: Construct Validity in Large Language Model Benchmarks

Andrew M. Bean<sup>1,\*</sup> Ryan Othniel Kearns<sup>1</sup> Angelika Romanou<sup>2</sup>  
Franziska Sofia Hafner<sup>1</sup> Harry Mayne<sup>1</sup>

Jan Batzner<sup>3,4</sup> Negar Foroutan<sup>2</sup> Chris Schmitz<sup>5</sup> Karolina Korgul<sup>1</sup> Hunar Batra<sup>1</sup>  
Oishi Deb<sup>1</sup> Emma Beharry<sup>6</sup> Cornelius Emde<sup>1</sup> Thomas Foster<sup>1</sup> Anna Gausen<sup>7</sup>  
María Grandury<sup>8,9</sup> Simeng Han<sup>10</sup> Valentin Hofmann<sup>11,12</sup> Luqain Ibrahim<sup>1</sup>  
Gabrielle Kaili-May Liu<sup>10</sup> Hazel Kim<sup>1</sup> Hannah Rose Kirk<sup>1,7</sup> Fangru Lin<sup>1</sup>  
Adel Bibi<sup>1</sup> Lennart Luettgau<sup>7</sup> Jabez Magomere<sup>1</sup> Jonathan Ryström<sup>1</sup>  
Yarin Gal<sup>1,7</sup> Antoine Bosselut<sup>2</sup> Ronald Clark<sup>1</sup> Arman Cohan<sup>10</sup> Jakob Foerster<sup>1</sup>  
Scott A. Hale<sup>1,13</sup> Inioluwa Deborah Raji<sup>14</sup> Christopher Summerfield<sup>1,7</sup>  
Philip H.S. Torr<sup>1</sup> Cozmin Ududec<sup>7</sup> Luc Rocher<sup>1</sup> Adam Mahdi<sup>1,\*</sup>

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

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. Most prominent are the terms 'deductive' and 'abductive' reasoning, which are often used interchangeably. In this work, we propose a framework for breaking down complex claims into atomic reasoning steps. The motivation behind this is the lack of uniform terminology around reasoning. Most prominent are the terms 'deductive' and 'abductive' reasoning, which are often used interchangeably.

## 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), passing professional medical tests (Kung et al., 2023; Nori et al., 2023), performing quantitative reasoning (Lewkowycz et al., 2022), and communication games (Bakhtin et al., 2022; et al., 2023; Gandhi et al., 2022).

## 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.

These findings motivate the need for further research on the reasoning capabilities of LLMs in high-stake real-world applications, 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 the proliferation of misinformation (Wei et al., 2023). Verifying information is challenging as models require both accurate veracity since models require both accurate veracity 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 of 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. Most prominent are the terms 'deductive' and 'abductive' reasoning, which are often used interchangeably. In this work, we propose a framework for breaking down complex claims into atomic reasoning steps. The motivation behind this is the lack of uniform terminology around reasoning. Most prominent are the terms 'deductive' and 'abductive' reasoning, which are often used interchangeably.

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

John Dougrez-Lewis<sup>1,\*</sup> Mahmud Elahi Akhter<sup>2,\*</sup> Federico Ruggeri<sup>3</sup>,  
Sebastian Löbbers<sup>2</sup>, Yulan He<sup>4,5</sup>, Maria Liakata<sup>2,5</sup>,  
<sup>1</sup>University of Warwick, UK, <sup>2</sup>Queen Mary University of London, UK  
<sup>3</sup>University of Bologna, Italy, <sup>4</sup>King's College London, UK  
<sup>5</sup>The Alan Turing Institute, UK  
j.dougrez-lewis@warwick.ac.uk, yulan.he@cl.ac.uk, federico.ruggeri@unibz.it, s.lobbers, m.liakata}@qmul.ac.uk

---

## **Punto 2: “Reasoning”?**

---



## Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang<sup>1,2</sup> Elena Gribovskaya<sup>1</sup> Nora Kassner<sup>1</sup> Mor Geva<sup>3,4\*</sup> Sebastian Riedel<sup>1,2\*</sup>

<sup>1</sup>Google DeepMind <sup>2</sup>UCL <sup>3</sup>Google Research <sup>4</sup>Tel Aviv University

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



## Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang<sup>1,2</sup> Elena Gribovskaya<sup>1</sup> Nora Kassner<sup>1</sup> Mor Geva<sup>3,4\*</sup> Sebastian Riedel<sup>1,2\*</sup>

## 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<sup>1,2</sup> Elena Gribovskaya<sup>1</sup> Nora Kassner<sup>1</sup> Mor Geva<sup>3,4\*</sup> Sebastian Riedel<sup>1,2\*</sup>

## 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<sup>1,2</sup> Elena Gribovskaya<sup>1</sup> Nora Kassner<sup>1</sup> Mor Geva<sup>3,4\*</sup> Sebastian Riedel<sup>1,2\*</sup>

## 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 Shojaee\*† Iman Mirzadeh\* Keivan Alizadeh  
Maxwell Horton Samy Bengio Mehrdad Farajtabar



## Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang<sup>1,2</sup> Elena Gribovskaya<sup>1</sup> Nora Kassner<sup>1</sup> Mor Geva<sup>3,4\*</sup> Sebastian Riedel<sup>1,2\*</sup>

### 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 Shojaee\*† Iman Mirzadeh\* Keivan Alizadeh  
Maxwell Horton Samy Bengio Mehrdad Farajtabar



## Emergent Analogical Reasoning in Large Language Models

Taylor Webb<sup>1,\*</sup>, Keith J. Holyoak<sup>1</sup>, and Hongjing Lu<sup>1,2</sup>

<sup>1</sup>Department of Psychology

<sup>2</sup>Department of Statistics

University of California, Los Angeles, CA, USA



## Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang<sup>1,2</sup> Elena Gribovskaya<sup>1</sup> Nora Kassner<sup>1</sup> Mor Geva<sup>3,4\*</sup> Sebastian Riedel<sup>1,2\*</sup>

### 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 Shojaee\*† Iman Mirzadeh\* Keivan Alizadeh  
Maxwell Horton Samy Bengio Mehrdad Farajtabar



## Emergent Analogical Reasoning in Large Language Models

Taylor Webb<sup>1,\*</sup>, Keith J. Holyoak<sup>1</sup>, and Hongjing Lu<sup>1,2</sup>

### LARGE LANGUAGE MODELS AS ANALOGICAL REASONERS

Michihiro Yasunaga,<sup>1,2</sup> Xinyun Chen,<sup>1</sup> Yujia Li,<sup>1</sup> Panupong Pasupat,<sup>1</sup> Jure Leskovec,<sup>2</sup>  
Percy Liang,<sup>2</sup> Ed H. Chi,<sup>1</sup> Denny Zhou<sup>1</sup>

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



## Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang<sup>1,2</sup> Elena Gribovskaya<sup>1</sup> Nora Kassner<sup>1</sup> Mor Geva<sup>3,4\*</sup> Sebastian Riedel<sup>1,2\*</sup>

## 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 Shojaee\*† Iman Mirzadeh\* Keivan Alizadeh  
Maxwell Horton Samy Bengio Mehrdad Farajtabar



## Emergent Analogical Reasoning in Large Language Models

Taylor Webb<sup>1,\*</sup>, Keith J. Holyoak<sup>1</sup>, and Hongjing Lu<sup>1,2</sup>

## LARGE LANGUAGE MODELS AS ANALOGICAL REASONERS

Michihiro Yasunaga,<sup>1,2</sup> Xinyun Chen,<sup>1</sup> Yujia Li,<sup>1</sup> Panupong Pasupat,<sup>1</sup> Jure Leskovec,<sup>2</sup>  
Percy Liang,<sup>2</sup> Ed H. Chi,<sup>1</sup> Denny Zhou<sup>1</sup>

<sup>1</sup> Google DeepMind <sup>2</sup> Stanford University

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

Abulhair Saparov<sup>†</sup> Richard Yuanzhe Pang<sup>†</sup> Vishakh Padmakumar<sup>†</sup> Nitish Joshi<sup>†</sup>

Seyed Mehran Kazemi<sup>△</sup>

Najoung Kim<sup>△,β,\*</sup>

He He<sup>†</sup>



## Do Large Language Models Latently Perform Multi-Hop Reasoning?

Sohee Yang<sup>1,2</sup> Elena Gribovskaya<sup>1</sup> Nora Kassner<sup>1</sup> Mor Geva<sup>3,4\*</sup> Sebastian Riedel<sup>1,2\*</sup>

### 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 Shojaee\*† Iman Mirzadeh\* Keivan Alizadeh  
Maxwell Horton Samy Bengio Mehrdad Farajtabar

## Emergent Analogical Reasoning in Large Language Models

Taylor Webb<sup>1,\*</sup>, Keith J. Holyoak<sup>1</sup>, and Hongjing Lu<sup>1,2</sup>

### LARGE LANGUAGE MODELS AS ANALOGICAL REASONERS

Michihiro Yasunaga,<sup>1,2</sup> Xinyun Chen,<sup>1</sup> Yujia Li,<sup>1</sup> Panupong Pasupat,<sup>1</sup> Jure Leskovec,<sup>2</sup>  
Percy Liang,<sup>2</sup> Ed H. Chi,<sup>1</sup> Denny Zhou<sup>1</sup>

<sup>1</sup> Google DeepMind <sup>2</sup> Stanford University

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

Abulhair Saparov<sup>†</sup> Richard Yuanzhe Pang<sup>†</sup> Vishakh Padmakumar<sup>†</sup> Nitish Joshi<sup>†</sup>

## The Illusion of the Illusion of Thinking

A Comment on Shojaee et al. (2025)

A. Lawsen\*

# Measuring what Matters: Construct Validity in Large Language Model Benchmarks

Andrew M. Bean<sup>1\*</sup> Ryan Othniel Kearns<sup>1</sup> Angelika Romanou<sup>2</sup>  
Franziska Sofia Hafner<sup>1</sup> Harry Mayne<sup>1</sup>

Jan Batzner<sup>3,4</sup> Negar Foroutan<sup>2</sup> Chris Schmitz<sup>5</sup> Karolina Korgul<sup>1</sup> Hunar Batra<sup>1</sup>  
Oishi Deb<sup>1</sup> Emma Beharry<sup>6</sup> Cornelius Emde<sup>1</sup> Thomas Foster<sup>1</sup> Anna Gausen<sup>7</sup>  
Maria Grandury<sup>8,9</sup> Simeng Han<sup>10</sup> Valentin Hofmann<sup>11,12</sup> Lujain Ibrahim<sup>1</sup>  
Gabrielle Kaili-May Liu<sup>10</sup> Hazel Kim<sup>1</sup> Hannah Rose Kirk<sup>1,7</sup> Fangru Lin<sup>1</sup> Jabez Magomere<sup>1</sup> Jonathan Ryström<sup>1</sup>  
Adel Bibi<sup>1</sup> Antoine Bosselut<sup>2</sup> Ronald Clark<sup>1</sup> Arman Cohan<sup>10</sup> Jakob Foerster<sup>1</sup>  
Yarin Gal<sup>1,7</sup> Scott A. Hale<sup>1,13</sup> Inioluwa Deborah Raji<sup>14</sup> Christopher Summerfield<sup>1,7</sup>  
Philip H.S. Torr<sup>1</sup> Cozmin Ududec<sup>7</sup> Luc Rocher<sup>1</sup> Adam Mahdi<sup>1\*</sup>

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

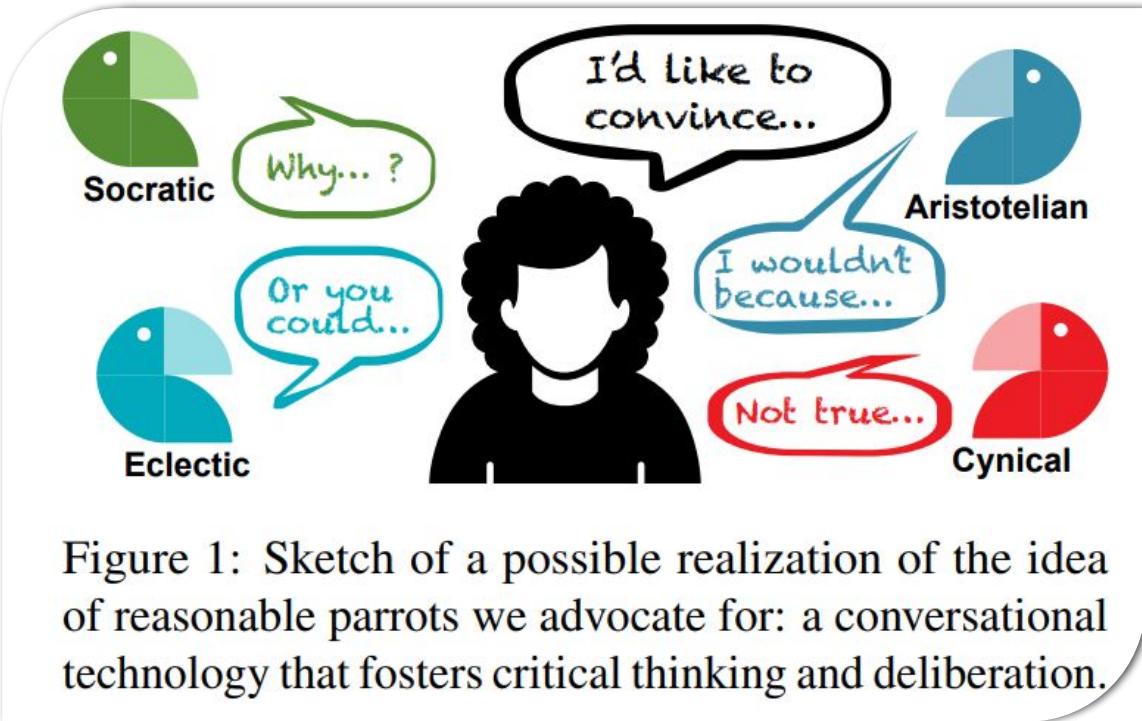
Benchmarks and evaluations play a critical role in the development of AI research [1, 2]. Creating a benchmark requires defining concrete tasks and metrics that measure the ‘intelligence’ of a model. It is often the case that these benchmarks are not well-defined, leading to ambiguous and unreliable results. This lack of clarity can lead to biased and unfair evaluations, which can in turn reinforce harmful biases and perpetuate social inequalities. To address this issue, we propose a new approach to benchmarking that focuses on the *construct validity* of the benchmarks. This approach involves a systematic review of existing benchmarks and a detailed analysis of their design, implementation, and evaluation. By doing so, we can identify patterns and trends that can help us better understand the strengths and weaknesses of different benchmarks. This can ultimately lead to more accurate and reliable evaluations, which can help us build more responsible and ethical AI systems.

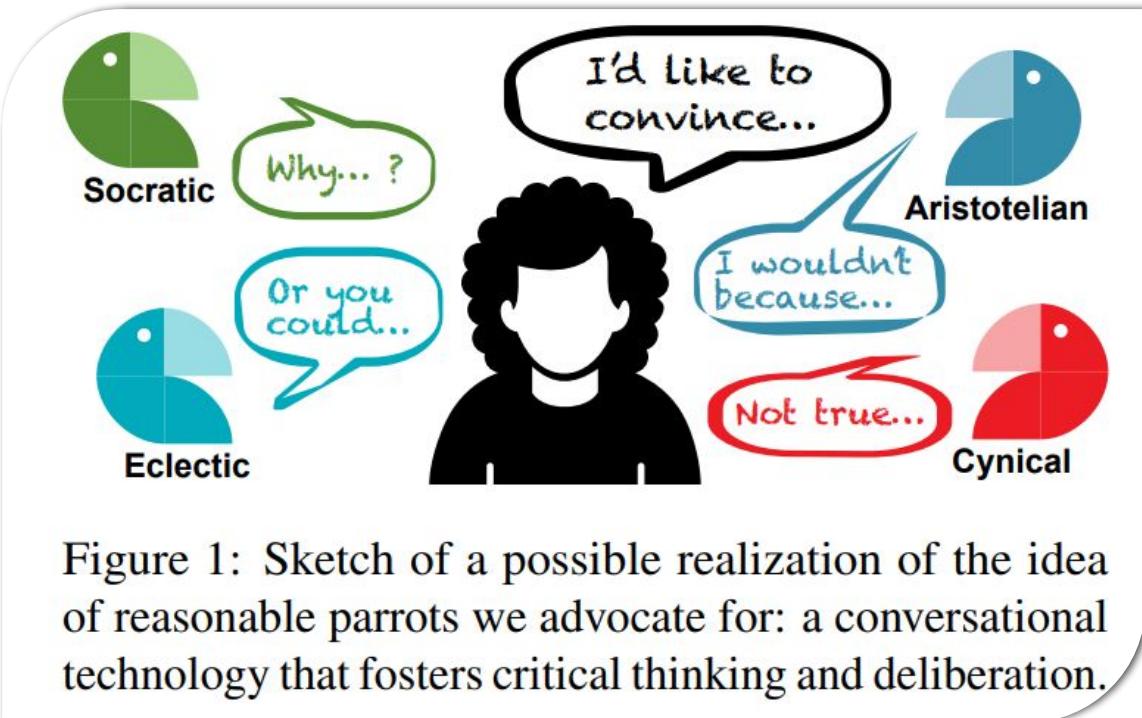
# Don't really know what benchmarks are measuring!

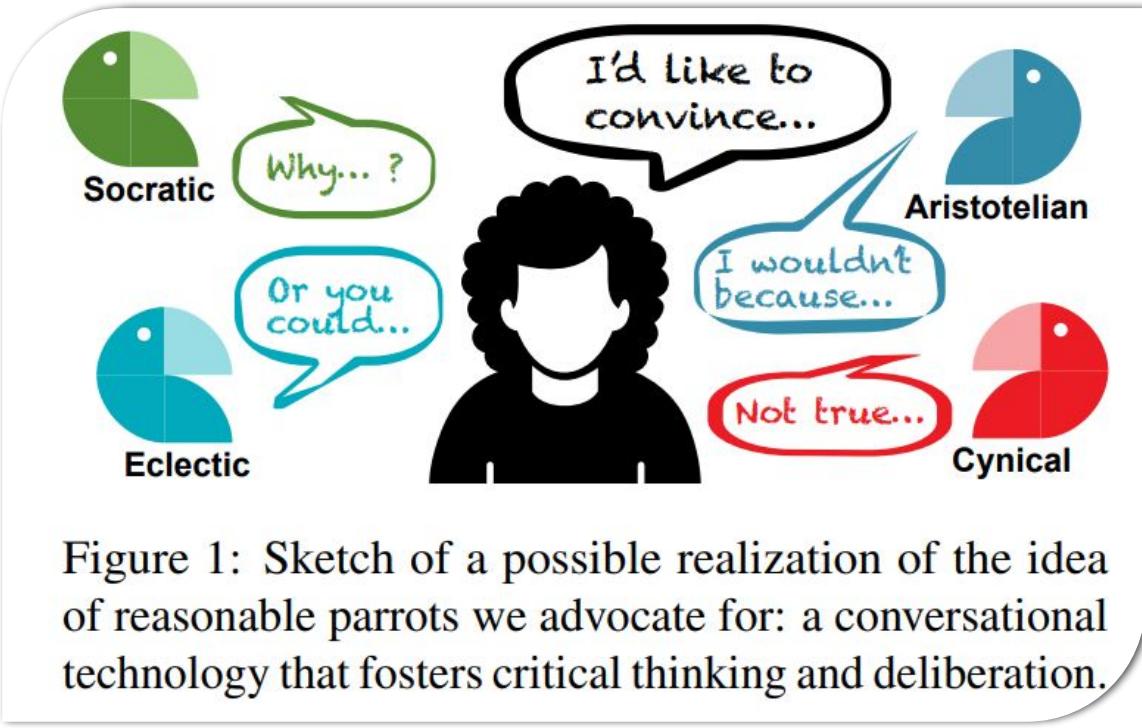
---

## **Punto 3: Argumentation**

---



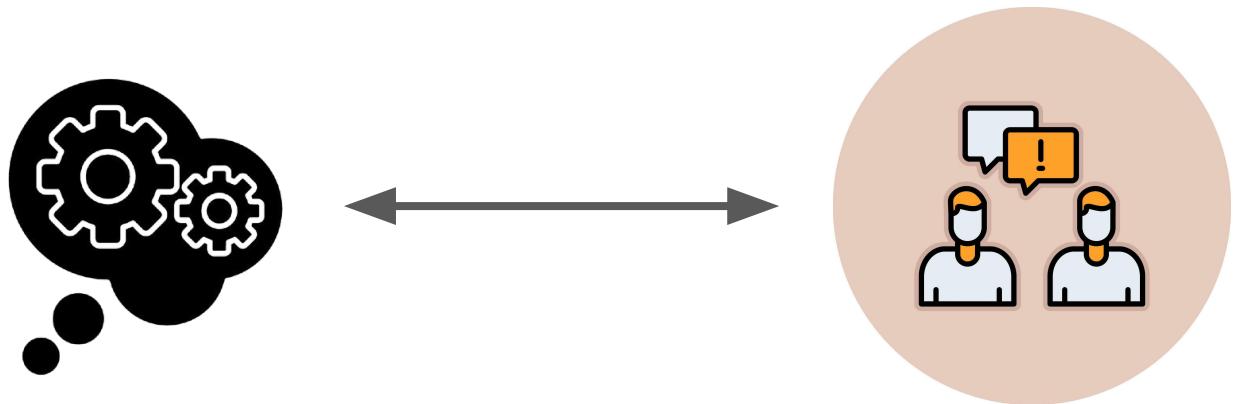


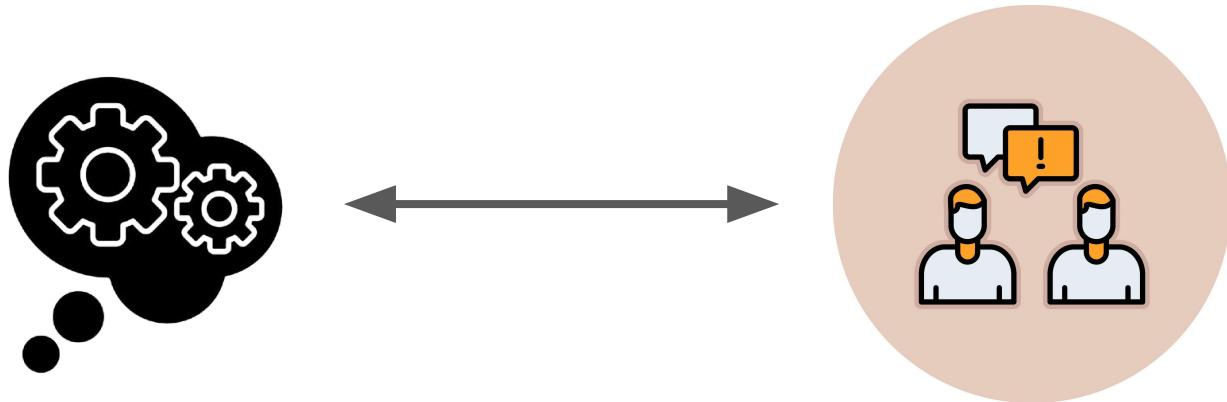


**“Ad populum fallacy”**

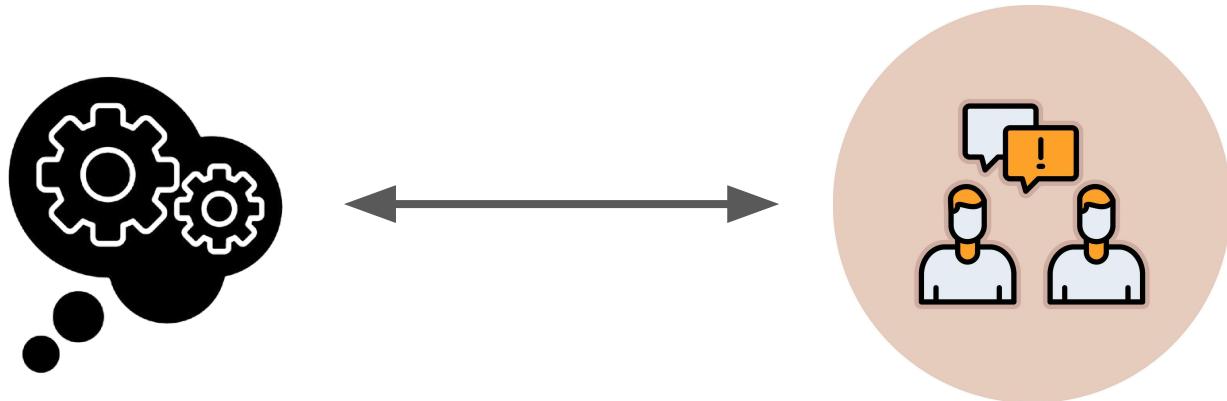
**Argue by design**

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.





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 13<sup>th</sup> 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 13<sup>th</sup> 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 13<sup>th</sup> 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;
- Novel benchmarks in argument mining that cater to the recent developments in LLM reasoning as a whole;
- Modeling, assessing, and critically reflecting on the argumentation capabilities of LLMs;

# **Proposal for ArgMining 2026, the 13<sup>th</sup> 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 13<sup>th</sup> 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!**

---