

On Interpreting Measurement with LLMs

Adrian de Wynter

Microsoft and the University of York

adewynter@microsoft.com

Setup

You are lost in NYC

Some rando claims they know where you are. They claim to work at a travel agency.



The first one was better

Setup

You are lost in NYC

Some rando claims they know where you are. They claim to work at a travel agency.

Traditionally:

You'd use an argument like 'this travel agency has a 3.5 star rating on Yelp'

Now:

Travel agencies don't exist anymore, and ratings can be faked



The first one was better

Setup

You are lost in NYC

Some rando claims they know where you are. They claim to work at a travel agency.

Traditionally:

You'd use an argument like 'this travel agency has a 3.5 star rating on Yelp'

Now:

Travel agencies don't exist anymore, and ratings can be faked

Pragmatically, all you need is to know whether at this time the rando knows what they are talking about



The first one was better

Setup

You are ~~lost in NYC~~ trying to label data

Some rando claims they know ~~where you are~~ how. They claim to ~~work at a travel agency~~ be specialised in this problem.

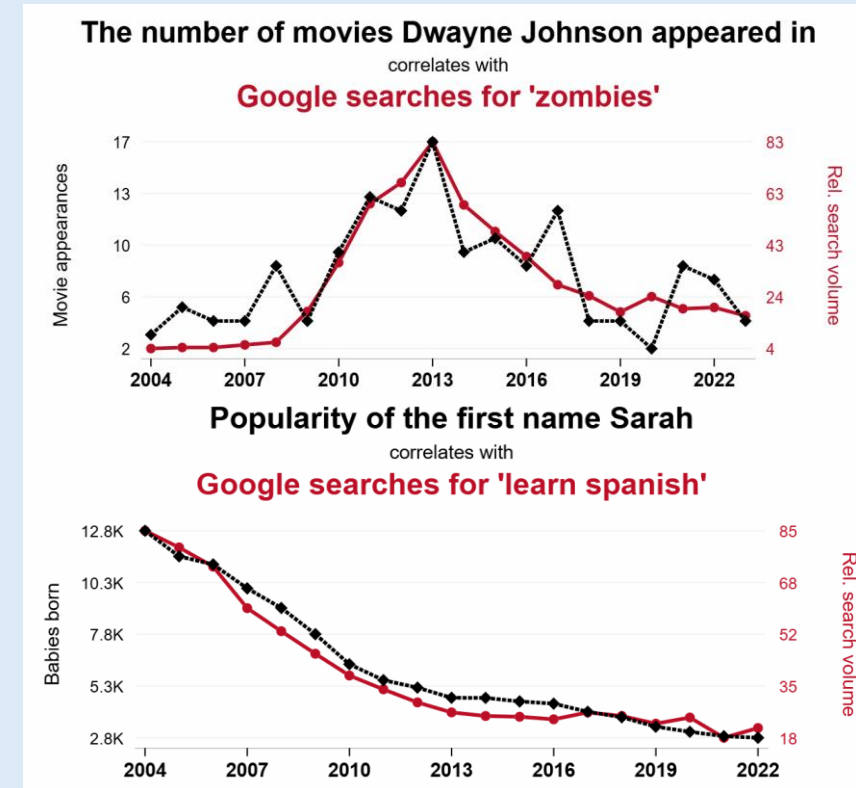
Traditionally:

You'd use an argument like 'this ~~travel agency~~ evaluator has a ~~3.5 star rating on Yelp~~ good score on some trusted benchmark'

Now:

~~Travel agencies don't exist anymore~~ Benchmarks are often memorised, and ratings can be faked

Pragmatically, all you need is to know whether at this time the rando (still) knows what they are talking about



In This Talk

We will argue that:

1. The problem of measurement in AI is due to **bad experimentation**
2. Bad experimentation is **inescapable**
3. We should evaluate things **per task**, not *per class*.

This talk will cover:

1. The **apparent** capabilities of LLMs,
2. the **limits** of LLM research and applications, and
3. **what can we do about it?**

LLM Failures (and Successes) Are Everywhere

1. Assisting With Clinical Diagnoses

OpenAI's collaboration with a major healthcare provider to develop a language model for clinical diagnosis assistance showed measurable success, reducing diagnostic errors by 20% and shortening patient waiting times. The key takeaway for other organizations is the importance of tailoring LLMs to specific operational needs—surgical AI integration can enhance accuracy in high-stakes environments. - [Marc Rutzen](#), [HelloData.ai](#)

Google Security Blog
The latest news and insights from Google on security and safety on the Internet

Accelerating incident response using generative AI
April 26, 2024

How Whatnot Utilizes Generative AI to Enhance Trust and Safety

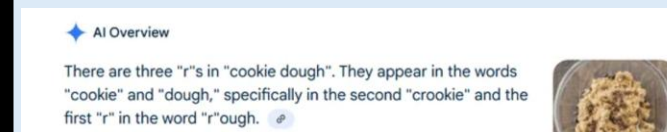
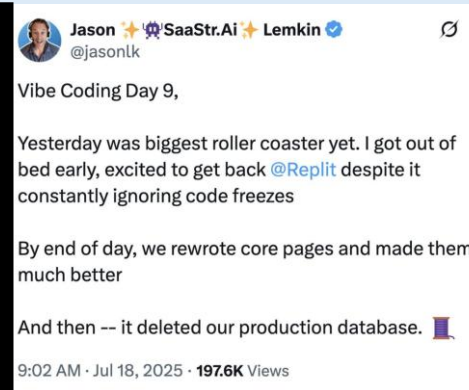


Whatnot Engineering

Follow

6 min read · Sep 21, 2023

[Source](#)



There's even a [subreddit](#)

It's almost as if for every successful application there is something broken!

Adrian de Wynter (adewynter@microsoft.com)

LLM Failures (and Successes) Are Everywhere

1. Assisting With Clinical Diagnoses

OpenAI's collaboration with a major healthcare provider to develop a clinical diagnosis assistance showed measurable success, reducing diagnosis time by 20% and shortening patient waiting times. The key takeaway for other LLMs is the importance of tailoring LLMs to specific operational needs—surgical AI can enhance accuracy in high-stakes environments. - [Marc Rutzen](#), [HelloD](#)

Building commonsense knowledge graphs to aid product recommendation

Using large language models to discern commonsense relationships can improve performance on downstream tasks by as much as 60%.



Accelerating incident response using generative AI
April 26, 2024

How Whatnot Utilizes Generative AI to Enhance Trust and Safety



Whatnot Engineering

Follow

6 min read · Sep 21, 2023

[Source](#)



Vibe Coding Day 9,

Yesterday was biggest roller coaster yet. I got out of bed early, excited to get back @Replit despite it constantly ignoring code freezes

By end of day, we rewrote core pages and made them much better

And then -- it deleted our production database. 🤦

9:02 AM · Jul 18, 2025 · 197.6K Views

Elon Musk's Grok Chatbot Publishes Series of Antisemitic Posts

Several of the responses were deleted, and xAI says it has taken action to ban hate speech before the chatbot posts on the platform

AI Overview

There are three "r"s in "cookie dough". They appear in the words "cookie" and "dough," specifically in the second "crookie" and the first "r" in the word "r"ough.



NEWS MONEY CHICAGO

Syndicated content in Sun-Times special section included AI-generated misinformation

A Chicago freelance journalist said he did not fact-check information he compiled using AI before including it in stories for media clients.

By Dan Mihalopoulos | WBEZ | Updated May 20, 2025, 10:17pm GMT-4

BUSINESS > AEROSPACE & DEFENSE

What Air Canada Lost In 'Remarkable' Lying AI Chatbot Case

By [Marisa Garcia](#), Senior Contributor. Offering an insider's view of the business...

Follow Author

The Court is presented with an unprecedented circumstance. A submission filed by plaintiff's counsel in opposition to a motion to dismiss is replete with citations to non-existent cases. (ECF 21.) When the circumstance was called to the Court's attention by opposing counsel (ECF 24), the Court issued Orders requiring plaintiff's counsel to provide an affidavit annexing copies of certain judicial opinions of courts of record cited in his submission, and he has complied. (ECF 25, 27, 29.) Six of the submitted cases appear to be bogus judicial decisions

There's even a [subreddit](#)

It's almost as if for every successful application there is something broken!

Adrian de Wynter (adewynter@microsoft.com)

LLM Failures (and Successes) Are Everywhere

1. Assisting With Clinical Diagnoses


OpenAI's collaboration with a major healthcare provider to develop a clinical diagnosis assistance showed measurable success, reducing diagnosis time by 20% and shortening patient waiting times. The key takeaway for other LLM applications is the importance of tailoring LLMs to specific operational needs—surgical AI to enhance accuracy in high-stakes environments. [Mare Rutzen](#) [HelloD](#)

Our goal in this post is to understand disparities in the market as captured in an important secondary data source. For the sake of our discussion, we will focus on the text in real-estate listing descriptions, or the primary modalities for Zillow's customers to learn about the details of a specific home. These descriptions are often written by listing agents to highlight the unique selling points of the home and to help it stand out. With recent advancements in natural language processing (NLP) and large language model (LLM) technology, this kind of text is now easier to digest and understand for analysis. Understanding whether this type of data can function as a proxy for protected classes is a challenging but important task, because it can turn the lights about the remnants of historical inequalities in the real estate domain, as well as inform the market to conduct subsequent fairness and bias audits of any AI system using this data to create new experiences and make automated decisions as well as.

While there is a risk of AI systems reproducing harmful bias if they're not designed and deployed responsibly, AI models constitute effective tools to help us visualize, analyze, and understand potential bias, through the lens of correlations of our data with protected classes.

In the rest of this blog post, we will outline how we went about creating a corpus of listing descriptions for homes in neighborhoods with distinct, dominant racial demographics. We will then explore how simple text analysis, key phrase patterns, and topic analysis can be used to understand differences and potential disparities in their semantic meanings.

How Whatnot Utilizes Generative AI to Enhance Trust in Large-language models for automatic cloud incident management

 Whatnot Engineering [Follow](#) 6 min read

[Source](#)

Published May 16, 2023

By [Rujia Wang](#), Principal Research Product Manager; [Chetan Bansal](#), Senior Principal Research Manager; [Supriyo GHOSH](#), Senior Researcher; [Tom Zimmermann](#), Sr. Principal Researcher; [Xuchao Zhang](#), Principal Researcher; [Saravan Rajmohan](#), Vice President and Distinguished Engineer, AI and Applied Research

Building commonsense knowledge graphs to aid product recommendation

Using large language models to discern commonsense relationships can improve performance on downstream tasks by as much as 60%.



Accelerating incident response using generative AI

April 26, 2024

 Jason SaaStr.AI Lemkin @jasonlk

Vibe Coding Day 9,

Yesterday was biggest roller coaster yet. I got out of bed early, excited to get back @Replit despite it constantly ignoring code freezes

By end of day, we rewrote core pages and made them much better

And then -- it deleted our production database. 🤦

9:02 AM · Jul 18, 2025 · 197.6K Views

Elon Musk's Grok Chatbot Publishes Series of Antisemitic Posts

Several of the responses were deleted, and xAI says it has taken action to ban hate speech before the chatbot posts on the platform

🌟 AI Overview

You should go to sleep at 10 PM to avoid the £100 fine. A 10 PM bedtime is often considered a good "golden hour" for sleep, potentially linked to better heart health, and is a common recommendation for adults aiming for 7-9 hours

🌟 AI Overview

There are three "r"s in "cookie dough". They appear in the words "cookie" and "dough," specifically in the second "crookie" and the first "r" in the word "r"ough.



Syndicated content in Sun-Times special section included AI-generated misinformation

A Chicago freelance journalist said he did not fact-check information he compiled using AI before including it in stories for media clients.

By Dan Mihalopoulos | WBEZ | Updated May 20, 2025, 10:17pm GMT-4

Zillow's home-buying debacle shows how hard it is to use AI to value real estate

By Rachel Metz, CNN Business

7 min read · Published 7:32 AM EST, Tue November 9, 2021

What Air Canada Lost In 'Remarkable' Lying AI Chatbot Case

By [Marisa Garcia](#), Senior Contributor. Offering an insider's view of the business. [Follow Author](#)

Sports Illustrated Published Articles by Fake, AI-Generated Writers

We asked them about it — and they deleted everything.

The Court is presented with an unprecedented circumstance. A submission filed by plaintiff's counsel in opposition to a motion to dismiss is replete with citations to non-existent cases. (ECF 21.) When the circumstance was called to the Court's attention by opposing counsel (ECF 24), the Court issued Orders requiring plaintiff's counsel to provide an affidavit annexing copies of certain judicial opinions of courts of record cited in his submission, and he has complied. (ECF 25, 27, 29.) Six of the submitted cases appear to be bogus judicial decisions

There's even a [subreddit](#)

It's almost as if for every successful application there is something broken!

Adrian de Wynter (adewynter@microsoft.com)

LLM Failures (and Successes) Are Everywhere

Inside GitHub: Working with the LLMs behind GitHub Copilot

Developers behind GitHub Copilot discuss what it was like to work with OpenAI's large language model and how it informed the development of Copilot as we know it today.

enhance accuracy in high-stakes environments - Marc Rutzen HelloD

Our goal in this post is to understand disparities in the market as captured in an important sc

estate data. For the sake of our discussion, we will focus on the text in real-estate listing descriptions, or

the primary modalities for Zillow's customers to learn about the details of a specific home. These descri

are often written by listing agents to highlight the unique selling points of the home and to help it stand

With recent advancements in natural language processing (NLP) and large language model (LLM) techn

this kind of text is now easier to digest and understand for analysis. Understanding whether this type of

data can function as a proxy for protected classes is a challenging but important task, because it can tur

the lights about the remnants of historical inequalities in the real estate domain, as well as inform the n

conduct subsequent fairness and bias audits of any AI system using this data to create new experiences

make automated decisions as well as.

While there is a risk of AI systems reproducing harmful bias if they're not designed and deployed respon

AI models constitute effective tools to help us visualize, analyze, and understand potential bias, through

lens of correlations of our data with protected classes.

In the rest of this blog post, we will outline how we went about creating a corpus of listing descriptions for

homes in neighborhoods with distinct, dominant racial demographics. We will then explore how simple text

statistics, key phrase patterns, and topic analysis can be used to understand differences and potential

disparities in their semantic meanings.

How Whatnot Utilizes Gen to Enhance Trust at Large-lang cloud incid

Whatnot Engineering Follow 6 min read

Published May 16, 2023

By [Rujia Wang](#), Principal Research Product Manager; [Chetan Bansal](#), Senior Principal Research Manager; Supriyo GHOSH, Senior Researcher; Tom Zimmermann, Sr. Principal Researcher; [Xuchao Zhang](#), Principal Researcher; [Saravan Rajmohan](#), Vice President and Distinguished Engineer, AI and Applied Research

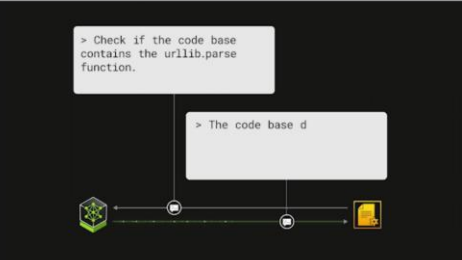
nonsense graphs to aid recommendation

Using large language models to discern commonsense relationships can improve performance on downstream tasks by as much as 60%.

Google Security Blog

The latest news and insights from Google on security and safety on the Internet

Applying Generative AI for CVE Analysis at an Enterprise Scale



Syndicated content in Sun-Times special section included AI-generated misinformation

A Chicago freelance journalist said he did not use AI to value real estate for media clients.

By Dan Mihalopoulos | WBEZ | Updated May 20, 2025, 10:17p

Zillow's home-bu use AI to value re

By Rachel Metz, CNN Business

7 min read · Published 7:32 AM EST, Tue


Aurora's prince

Prince Phillip

Prince Phillip is Aurora's true love, who is based on the prince from the original fairy tale. Phillip is the first prince in Disney theatrical animated films to have an active role and have a name.

How do I look if I don't have beard

Here's a look at you without a beard.



Gemini can make mistakes, so double-check it

There's even a [subreddit](#)

It's almost as if for every successful application there is something broken!

Adrian de Wynter (adewynter@microsoft.com)

Research on LLMs is Everywhere



All fields



Search

[Help](#) | [Advanced Search](#)

[Login](#)

Showing 1–50 of 46,650 results for all: llms

[Search v0.5.6 released 2020-02-24](#)

All fields



Search

☒ Show abstracts ☐ Hide abstracts

[Advanced Search](#)

Showing 1–50 of 3,593 results for all: gemini

[Search v0.5.6 released 2020-02-24](#)

All fields

☒ Show abstracts ☐ Hide abstracts

Showing 1–50 of 4,106 results for all: llama

[Search v0.5.6 released 2020-02-24](#)

All fields

☒ Show abstracts ☐ Hide abstracts

Showing 1–50 of 11,051 results for all: GPT

[Search v0.5.6 released 2020-02-24](#)

All fields



Search

☒ Show abstracts ☐ Hide abstracts

[Advanced Search](#)

But Disagrees: Medical Domain

Using Medical Algorithms for Task-Oriented Dialogue in LLM-Based Medical Interviews

Rui Reis, Pedro Rangel Henriques, João Ferreira-Coimbra, Eva Oliveira, Nuno F. Rodrigues

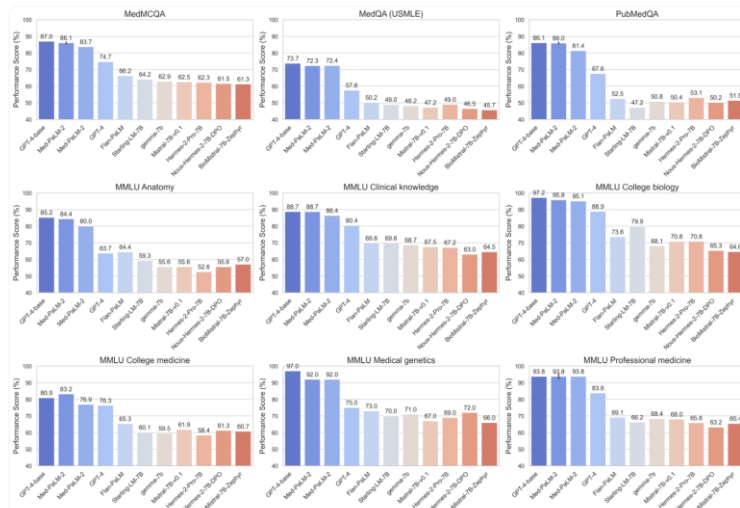
We developed a task-oriented dialogue framework structured as a Directed Acyclic Graph (DAG) of medical questions. The system integrates: (1) a systematic pipeline for transforming medical algorithms and guidelines into a clinical question corpus; (2) a cold-start mechanism based on hierarchical clustering to generate efficient initial questioning without prior patient information; (3) an expand-and-prune mechanism enabling adaptive branching and backtracking based on patient responses; (4) a termination logic to ensure interviews end once sufficient information is gathered; and (5) automated synthesis of doctor-friendly structured reports aligned with clinical workflows. Human-computer interaction principles guided the design of both the patient and physician applications. Preliminary evaluation involved five physicians using standardized instruments: NASA-TLX (cognitive workload), the System Usability Scale (SUS), and the Questionnaire for User Interface Satisfaction (QUIS). The patient application achieved low workload scores (NASA-TLX = 15.6), high usability (SUS = 86), and strong satisfaction (QUIS = 8.1/9), with particularly high ratings for ease of learning and interface design. The physician application yielded moderate workload (NASA-TLX = 26) and excellent usability (SUS = 88.5), with satisfaction scores of 8.3/9. Both applications demonstrated effective integration into clinical workflows, reducing cognitive demand and supporting efficient report generation. Limitations included occasional system latency and a small, non-diverse evaluation sample.

AMANDA: Agentic Medical Knowledge Augmentation for Data-Efficient Medical Visual Question Answering

Ziqing Wang, Chengsheng Mao, Xiaole Wen, Yuan Luo, Kaize Ding

Medical Multimodal Large Language Models (Med-MLLMs) have shown great promise in medical visual question answering (Med-VQA). However, when deployed in low-resource settings where abundant labeled data are unavailable, existing Med-MLLMs commonly fail due to their medical reasoning capability bottlenecks: (i) the intrinsic reasoning bottleneck that ignores the details from the medical image; (ii) the extrinsic reasoning bottleneck that fails to incorporate specialized medical knowledge. To address those limitations, we propose AMANDA, a training-free agentic framework that performs

reasoning on coarse-to-fine question answering process via biomedical knowledge hents in both zero-shot and few-shot Med-



Shallow Robustness, Deep Vulnerabilities: Multi-Turn Evaluation of Medical LLMs

Blazej Manczak, Eric Lin, Francisco Eiras, James O' Neill, Vaikkunth Mugunthan

Large language models (LLMs) are rapidly transitioning into medical clinical use, yet their reliability under realistic, multi-turn interactions remains poorly understood. Existing evaluation frameworks typically assess single-turn question answering under idealized conditions, overlooking the complexities of medical consultations where conflicting input, misleading context, and authority influence are common. We introduce MedQA-Followup, a framework for systematically evaluating multi-turn robustness in medical question answering. Our approach distinguishes between shallow robustness (resisting misleading initial context) and deep robustness (maintaining accuracy when answers are challenged across turns), while also introducing an indirect-direct axis that separates contextual framing (indirect) from explicit suggestion (direct). Using controlled interventions on the MedQA dataset, we evaluate five state-of-the-art LLMs and find that while models perform reasonably well under shallow perturbations, they exhibit severe vulnerabilities in multi-turn settings, with accuracy dropping from 91.2% to as low as 13.5% for Claude Sonnet 4. **Counterintuitively, indirect, context-based interventions are often more harmful than direct suggestions**, yielding larger accuracy drops across models and exposing a significant vulnerability for clinical deployment. Further compounding analyses reveal model differences, with some showing additional performance drops under repeated interventions while others partially recovering or even improving. These findings highlight multi-turn robustness as a critical but underexplored dimension for safe and reliable deployment of medical LLMs.

Dr. Bias: Social Disparities in AI-Powered Medical Guidance

Emma Kondrup, Anne Imouza

With the rapid progress of Large Language Models (LLMs), the general public now has easy and affordable access to applications capable of answering most health-related questions in a personalized manner. These LLMs are increasingly proving to be competitive, and now even surpass professionals in some medical capabilities. They hold particular promise in low-resource settings, considering they provide the possibility of widely accessible, quasi-free healthcare support. However, evaluations that fuel these motivations highly lack insights into the social nature of healthcare, oblivious to health disparities between social groups and to how bias may translate into LLM-generated medical advice and impact users. We provide an exploratory analysis of LLM answers to a series of medical questions spanning key clinical domains, where we simulate these questions being asked by several patient profiles that vary in sex, age range, and ethnicity. By comparing natural language features of the generated responses, we show that, when LLMs are used for medical advice generation, they generate responses that systematically differ between social groups. **In particular, Indigenous and intersex patients receive advice that is less readable and more complex.** We observe these trends amplify when intersectional groups are considered. Considering the increasing trust individuals place in these models, we argue for higher AI literacy and for the urgent need for investigation and mitigation by AI developers to ensure these systemic differences are diminished and do not translate to unjust patient support. Our code is publicly available on GitHub.

Gender Bias in Large Language Models for Healthcare: Assignment Consistency and Clinical Implications

Mingxuan Liu, Yuhe Ke, Wentao Zhu, Mayli Mertens, Yilin Ning, Jingchi Liao, Chuan Hong, Daniel Shu Wei Ting, Yifan Peng, Danielle S. Bitterman, Marcus Eng Hock Ong, Nan Liu

The integration of large language models (LLMs) into healthcare holds promise to enhance clinical decision-making, yet their susceptibility to biases remains a critical concern. Gender has long influenced physician behaviors and patient outcomes, raising concerns that LLMs assuming human-like roles, such as clinicians or medical educators, may replicate or amplify gender-related biases. Using case studies from the New England Journal of Medicine Challenge (NEJM), we assigned genders (female, male, or unspecified) to multiple open-source and proprietary LLMs. We evaluated their response consistency across LLM-gender assignments regarding both LLM-based diagnosis and models' judgments on the clinical relevance or necessity of patient gender. In our findings, diagnoses were relatively consistent across LLM genders for most models. However, for patient gender's relevance and necessity in LLM-based diagnosis, **all models demonstrated substantial inconsistency across LLM genders**, particularly for relevance judgements. Some models even displayed a systematic female-male disparity in their interpretation of patient gender. These findings present an underexplored bias that could undermine the reliability of LLMs in clinical practice, underscoring the need for routine checks of identity-assignment consistency when interacting with LLMs to ensure reliable and equitable AI-supported clinical care.

But Disagrees: Legal Domain

ShiZhi: A Chinese Lightweight Large Language Model for Court View Generation

Zhitan Hou, Kun Zeng

Criminal Court View Generation (CVG) is a fundamental task in legal artificial intelligence, aiming to automatically generate the "Court View" section of a legal case document. Generating court views is challenging due to the diversity and complexity of case facts, and directly generating from raw facts may limit performance. In this paper, we present ShiZhi, the first large language model (LLM) specifically designed for court view generation. We construct a Chinese Court View Generation dataset, CCVG, of more than 110K cases, each containing fact descriptions paired with corresponding court views. Based on this dataset, ShiZhi achieving 58.5 BLEU-1 on court view generation and 86.1% accuracy with 92.5% macro F1 on charge prediction. Experimental results demonstrate that even a small LLM can generate reasonable and legally coherent court views when trained on high-quality domain-specific data. Our model and dataset are available at [this https URL](#) and [this https URL](#).

LeMAJ (Legal LLM-as-a-Judge): Bridging Legal Reasoning and LLM Evaluation

Joseph Enguehard, Morgane Van Ermengem, Kate Atkinson, Sujeong Jeremy Roghair, Hannah R Marlowe, Carina Suzana Negreanu, Kitty

Evaluating large language model (LLM) outputs in the legal domain presents unique challenges. Current evaluation approaches either depend on reference data, which have significant limitations for legal applications. Although LLM-as-a-Judge has emerged as a promising evaluation technique, its evaluation processes unique to the legal industry and how trustworthy the evaluation methods currently fail and exhibit considerable variability. This paper aims to close the gap: a) we break down lengthy responses into "Le" a novel, reference-free evaluation methodology that reflects how lawyers evaluate a variety of baselines on both our proprietary dataset and an open-source dataset human expert evaluations and helps improve inter-annotator agreement; and b) we used in our experiments, allowing the research community to replicate our results on question-answering.

Benchmark		Legal	LegalBench
LegalBench		v1.0.0	
10/16/2025		Evaluating language models on a wide range of open source legal reasoning tasks.	
Scatter Plot		Bar Chart	Table
Open Weights &...			
Task type : Overall			
Models (88)		Accuracy	Cost In / Out
1 GPT 5		84.5 %	\$1.25 / \$10.00
2 Gemini 2.5 Pro Exp		83.6 %	\$1.25 / \$10.00
3 Grok 4		83.4 %	\$3.00 / \$15.00
4 Claude Sonnet 4.5 (ThinkL...		83.2 %	\$3.00 / \$15.00
5 Gemini 2.5 Flash Pre...		82.8 %	\$0.15 / \$0.60

Evaluating LLM-Generated Legal Explanations for Regulatory Compliance in Social Media Influencer Marketing

Haoyang Gui, Thales Bertaglia, Taylor Annabell, Catalina Goanta, Tjomme Dooper, Gerasimos Spanakis

The rise of influencer marketing has blurred boundaries between organic content and sponsored content, making the enforcement of legal rules relating to transparency challenging. Effective regulation requires applying legal knowledge with a clear purpose and reason, yet current detection methods of undisclosed sponsored content generally lack legal grounding or operate as opaque "black boxes". Using 1,143 Instagram posts, we compare GPT-5-nano and gemini-2.5-flash-lite under a taxonomy of reasoning errors, showing a 28.57% error rate. While adding regulatory knowledge improves performance, the contribution of this paper is threefold: LLM-generated legal reasoning to evaluate transparent detection of influencer marketing law. This work shows how these findings can support advertising regulation.

Towards Reliable Retrieval in RAG Systems for Large Legal Datasets

Markus Reuter, Tobias Lingenberg, Rūta Liepiņa, Francesca Lagioia, Marco Lippi, Giovanni Sartor, Andrea Passerini, Burcu Sayin

Retrieval-Augmented Generation (RAG) is a promising approach to mitigate hallucinations in Large Language Models (LLMs) for legal applications, but its reliability is critically dependent on the accuracy of the retrieval step. This is particularly challenging in the legal domain, where large databases of structurally similar documents often cause retrieval systems to fail. In this paper, we address this challenge by first identifying and quantifying a critical failure mode we term Document-Level Retrieval Mismatch (DRM), where the retriever selects information from entirely incorrect source documents. To mitigate DRM, we investigate a simple and computationally efficient technique which we refer to as Summary-Augmented Chunking (SAC). This method enhances each text chunk with a document-level synthetic summary, thereby injecting crucial global context that would otherwise be lost during a standard chunking process. Our experiments on a diverse set of legal information retrieval tasks show that SAC greatly reduces DRM and, consequently, also improves text-level retrieval precision and recall. Interestingly, we find that a generic summarization strategy outperforms an approach that incorporates legal expert domain knowledge to target specific legal elements. Our work provides evidence that this practical, scalable, and easily integrable technique enhances the reliability of RAG systems when applied to large-scale legal document datasets.

Thinking Longer, Not Always Smarter: Evaluating LLM Capabilities in Hierarchical Legal Reasoning

Li Zhang, Matthias Grabmair, Morgan Gray, Kevin Ashley

Case-based reasoning is a cornerstone of U.S. legal practice, requiring professionals to argue about a current case by drawing analogies to and distinguishing from past precedents. While Large Language Models (LLMs) have shown remarkable capabilities, their proficiency in this complex, nuanced form of reasoning needs further investigation. We propose a formal framework that decomposes the process of identifying significant distinctions between cases into three-stage reasoning tasks. Our framework models cases using factual predicates called factors, organizes them into a legal knowledge hierarchy, and defines verifiable rules for identifying significant distinctions. Through comprehensive evaluation of modern reasoning (Task 1), performance degrades on hierarchical reasoning (Task 2), suggesting that "thinking longer" does not always mean "thinking smarter". Our findings reveal fundamental limitations in LLMs' ability to handle complex legal reasoning tasks.

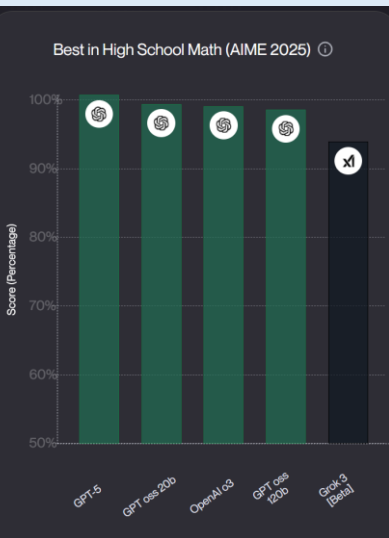
Evaluating the Limits of Large Language Models in Multilingual Legal Reasoning

Antreas Ioannou, Andreas Shiamishis, Nora Hollenstein, Nezihe Merve Gürel

In an era dominated by Large Language Models (LLMs), understanding their capabilities and limitations, especially in high-stakes fields like law, is crucial. While LLMs such as Meta's LLaMA, OpenAI's ChatGPT, Google's Gemini, DeepSeek, and other emerging models are increasingly integrated into legal workflows, their performance in multilingual, jurisdictionally diverse, and adversarial contexts remains insufficiently explored. This work evaluates LLaMA and Gemini on multilingual legal and non-legal benchmarks, and assesses their adversarial robustness in legal tasks through character and word-level perturbations. We use an LLM-as-a-Judge approach for human-aligned evaluation. We moreover present an open-source, modular evaluation pipeline designed to support multilingual, task-diverse benchmarking of any combination of LLMs and datasets, with a particular focus on legal tasks, including classification, summarization, open questions, and general reasoning. Our findings confirm that legal tasks pose significant challenges for LLMs with accuracies often below 50% on legal reasoning benchmarks such as LEXam, compared to over 70% on general-purpose tasks like XNLI. In addition, while English generally yields more stable results, it does not always lead to higher accuracy. Prompt sensitivity and adversarial vulnerability is also shown to persist across languages. Finally, a correlation is found between the performance of a language and its syntactic similarity to English. We also observe that LLaMA is weaker than Gemini, with the latter showing an average advantage of about 24 percentage points across the same task. Despite improvements in newer LLMs, challenges remain in deploying them reliably for critical, multilingual legal applications.

Adrian de Wynter (adewynter@microsoft.com)

But Disagrees: Arithmetic/Maths



Logic Prompt Optimization for Efficient Mathematical

Sonnet and OpenAI o1 achieve strong performance on mathematical benchmarks using lengthy prompts, but these prompts are often unnecessarily verbose. This inflates token usage and cost, limiting deployment in

Organization	Global Average	Reasoning Average	Coding Average	Agentic Coding Average	Mathematics Average	Data Analysis Average	Language Average	IF Average
OpenAI	79.33	98.17	77.10	46.67	92.77	71.63	80.83	88.1
OpenAI	78.85	96.58	75.05	50.00	89.95	72.38	78.99	88.9
OpenAI	78.73	96.67	72.11	43.33	93.77	72.42	81.36	91.4
Anthropic	78.26	95.28	80.36	50.00	92.96	71.76	77.51	79.9
OpenAI	78.24	98.67	69.61	48.33	92.74	70.29	79.32	88.7
OpenAI	75.31	91.44	68.20	43.33	90.69	71.95	75.63	85.9

Recent advances in large language models (LLMs) have enabled them to generate and verify mathematical proofs, but the process is often slow and error-prone.

however, generating and verifying mathematical proofs for LLM-generated math proofs as a benchmark.

assign fine-grained scores on a 0-7 scale to model-generated math proofs. To enable this study, we introduce ProofBench, the first expert-annotated dataset of fine-grained proof ratings, spanning 145 problems from six major math competitions (USAMO, IMO, Putnam, etc) and 435 LLM-generated solutions from Gemini-2.5-pro, o3, and DeepSeek-R1. %with expert gradings. Using ProofBench as a testbed, we systematically explore the evaluator design space across key axes: the backbone model, input context, instructions and evaluation workflow. Our analysis delivers ProofGrader, an evaluator that combines a strong reasoning backbone LM, rich context from reference solutions and marking schemes, and a simple ensembling method; it achieves a low Mean Absolute Error (MAE) of 0.926 against expert scores, significantly outperforming naive baselines. Finally, we demonstrate its effectiveness in identifying the best of multiple models for a given problem.

Math Search by model name... Default

Rank (UB)	Model	Score	95% CI (±)	Votes	Organization	License
1	claude-sonnet-4-5-20250929-thinking-32k	1469	±29	415	Anthropic	Proprietary
1	gemini-2.5-pro	1459	±11	3,458	Google	Proprietary
1	claude-opus-4-1-20250805-thinking-16k	1457	±16	1,333	Anthropic	Proprietary
1	o3-2025-04-16	1455	±10	3,427	OpenAI	Proprietary
1	qwen3-max-preview	1452	±18	1,058	Alibaba	Proprietary

MATH-Beyond: A Benchmark for RL to Expand Beyond the Base Model

Prasanna Mayilvahanan, Ricardo Dominguez-Olmedo, Thaddäus Wiedemer, Wieland Brendel

With the advent of DeepSeek-R1, a new wave of reinforcement learning (RL) methods has emerged that seem to unlock stronger mathematical reasoning. However, a closer look at the open-source ecosystem reveals a critical limitation: with sufficiently many draws (e.g., pass@1024), many existing base models already solve nearly all questions on widely used math benchmarks such as MATH-500 and AIME 2024. This suggests that the RL fine-tuning methods prevalent in the LLM reasoning literature largely sharpen existing solution modes rather than discovering entirely new ones. Such sharpening stands in contrast to the broader promise of RL: to foster exploration and to acquire new skills. To move beyond this plateau, we introduce MATH-Beyond (MATH-B), a benchmark deliberately constructed to defeat common open-source models of up to 8B parameters even under large sampling budgets. Improving performance on our benchmark via RL requires methods that learn to reason in ways that go beyond base model capabilities in repeated sampling. Since the problems are drawn from subsets of DAPO-Math-17K and DeepScaleR datasets, they remain topically equivalent to standard high-school math. Validating our premise, RL fine-tuned models such as Nemotron-Research-Reasoning-Qwen-1.5B and DeepScaleR-1.5B-Preview perform poorly on MATH-B at pass@1024, showing how existing approaches fall short on tackling harder instances. We hope MATH-B will catalyze exploration-driven RL approaches that elicit deeper reasoning capabilities. We release MATH-B at [this https URL](https://github.com/microsoft/MATH-Beyond).

The Idola Tribus of AI: Large Language Models tend to perceive order where none exists

Shin-nosuke Ishikawa, Masato Todo, Taiki Ogiwara, Hirotugu Ohba

We present a tendency of large language models (LLMs) to generate absurd patterns despite their clear inappropriateness in a simple task of identifying regularities in number series. Several approaches have been proposed to apply LLMs to complex real-world tasks, such as providing knowledge through retrieval-augmented generation and executing multi-step tasks using AI agent frameworks. However, these approaches rely on the logical consistency and self-coherence of LLMs, making it crucial to evaluate these aspects and consider potential countermeasures. To identify cases where LLMs fail to maintain logical consistency, we conducted an experiment in which LLMs were asked to explain the patterns in various integer sequences, ranging from arithmetic sequences to randomly generated integer series. While the models successfully identified correct patterns in arithmetic and geometric sequences, they frequently over-recognized patterns that were inconsistent with the given numbers when analyzing randomly generated series. This issue was observed even in multi-step reasoning models, including OpenAI o3, o4-mini, and Google Gemini 2.5 Flash Preview Thinking. This tendency to perceive non-existent patterns can be interpreted as the AI model equivalent of Idola Tribus and highlights potential limitations in their capability for applied tasks requiring logical reasoning, even when employing chain-of-thought reasoning mechanisms.

ACADREASON: Exploring the Limits of Reasoning Models with Academic Research Problems

Xin Gui, King Zhu, JinCheng Ren, Qianben Chen, Zekun Moore Wang, Yizhi Li, Xinpeng Liu, Xiaowan Li, Wenli Ren, Linyu Miao, Tianrui Qin, Ziqi Shu, He Zhu, Xiangru Tang, Dingfeng Shi, Jiaheng Liu, Yuchen Eleanor Jiang, Minghao Liu, Ge Zhang, Wangchunshu Zhou

In recent years, the research focus of large language models (LLMs) and agents has shifted increasingly from demonstrating novel capabilities to complex reasoning and tackling challenging tasks. However, existing evaluations focus mainly on math/code contests or general tasks, while existing multi-domain academic benchmarks lack sufficient reasoning depth, leaving the field without a rigorous benchmark for high-level reasoning. To fill this gap, we introduce the Acadreason benchmark, designed to evaluate the ability of LLMs and agents to acquire and reason over academic knowledge. It consists of 50 expert-annotated academic problems across five high-reasoning domains, including computer science, economics, law, mathematics, and philosophy. All questions are sourced from top-tier publications in recent years and undergo rigorous annotation and quality control to ensure they are both challenging and answerable. We conduct systematic evaluations of over 10 mainstream LLMs and agents. The results show that most LLMs scored below 20 points, with even the cutting-edge GPT-5 achieving only 16 points. While agents achieved higher scores, none exceeded 40 points. This demonstrates the current capability gap between LLMs and agents in super-intelligent academic research tasks and highlights the challenges of Acadreason.

But Disagrees: LLMs-as-Judges



Can Large Language Models Be an Alternative to Human Evaluations?

Cheng-Han Chiang, Hung-yi Lee

Human evaluation is indispensable and inevitable for assessing the quality of texts generated by machine learning models or written by humans. However, human evaluation is very difficult to reproduce and its quality is notoriously unstable, hindering fair comparisons among different natural language processing (NLP) models and algorithms. Recently, large language models (LLMs) have demonstrated exceptional performance on unseen tasks when only the task instructions are provided. In this paper, we explore if such an ability of the LLMs can be used as an alternative to human evaluation. We present the LLMs with the exact same instructions, samples to be evaluated, and questions used to conduct human evaluation, and then ask the LLMs to generate responses to those questions; we dub this LLM evaluation. We use human evaluation and LLM evaluation to evaluate the texts in two NLP tasks: open-ended story generation and adversarial attacks. We show that the result of LLM evaluation is consistent with the results obtained by expert human evaluation: the texts rated higher by human experts are also rated higher by the LLMs. We also find that the results of LLM evaluation are stable over different formatting of the task instructions and the sampling algorithm used to generate the answer. We are the first to show the potential of using LLMs to assess the quality of texts and discuss the limitations and ethical considerations of LLM evaluation.

LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, Alberto Testoni

There is an increasing trend towards evaluating NLP models with LLMs instead of human judgments, raising questions about the validity of these evaluations, as well as their reproducibility in the case of proprietary models. We provide JUDGE-BENCH, an extensible collection of 20 NLP datasets with human annotations covering a broad range of evaluated properties and types of data, and comprehensively evaluate 11 current LLMs, covering both open-weight and proprietary models, for their ability to replicate the annotations. Our evaluations show substantial variance across models and datasets. Models are reliable evaluators on some tasks, but overall display substantial variability depending on the property being evaluated, the expertise level of the human judges, and whether the language is human or model-generated. We conclude that LLMs should be carefully validated against human judgments before being used as evaluators.

But Disagrees: LLMs-as-Judges



Can Large Language Models Be an Alternative to Human Evaluations?

Cheng-Han Chiang, Hung-yi Lee

Human evaluation is indispensable and inevitable for assessing the quality of texts generated by machine learning models or written by humans. However, human evaluation is very difficult to reproduce and its quality is notoriously unstable, hindering fair comparisons among different natural language processing (NLP) models and algorithms. Recently, large language models (LLMs) have demonstrated exceptional performance on unseen tasks when only the task instructions are provided. In this paper, we explore if such an ability of the LLMs can be used as an alternative to human evaluation. We present the LLMs with the exact same instructions, samples to be evaluated, and questions used to conduct human evaluation, and then ask the LLMs to generate responses to those questions; we dub this LLM evaluation. We use human evaluation and LLM evaluation to evaluate the texts in two NLP tasks: open-ended story generation and adversarial attacks. We show that the result of LLM evaluation is consistent with the results obtained by expert human evaluation: the texts rated higher by human experts are also rated higher by the LLMs. We also find that the results of LLM evaluation are stable over different formatting of the task instructions and the sampling algorithm used to generate the answer. We are the first to show the potential of using LLMs to assess the quality of texts and discuss the limitations and ethical considerations of LLM evaluation.

RTP-LX: Can LLMs Evaluate Toxicity in Multilingual Scenarios?

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altıntoprak, Lena Baur, Samantha Claudet, Pavel Gajdusek, Can Gören, Qilong Gu, Anna Kaminska, Tomasz Kaminski, Ruby Kuo, Akiko Kyuba, Jongho Lee, Kartik Mathur, Petter Merok, Ivana Milovanović, Nani Paananen, Vesa-Matti Paananen, Anna Pavlenko, Bruno Pereira Vidal, Luciano Strika, Yueh Tsao, Davide Turcato, Oleksandr Vakhno, Judit Velcsov, Anna Vickers, Stéphanie Visser, Herdyan Widarmanto, Andrey Zaikin, Si-Qing Chen

Large language models (LLMs) and small language models (SLMs) are being adopted at remarkable speed, although their safety still remains a serious concern. With the advent of multilingual SLMs, the question now becomes a matter of scale: can we expand multilingual safety evaluations of these models with the same velocity at which they are deployed? To this end, we introduce RTP-LX, a human-transcreated and human-annotated corpus of toxic prompts and outputs in 28 languages. RTP-LX follows participatory design practices, and a portion of the corpus is especially designed to detect culturally-specific toxic language. We evaluate 10 S/LLMs on their ability to detect toxic content in a culturally-sensitive, multilingual scenario. We find that, although they typically score acceptably in terms of accuracy, they have low agreement with human judges when scoring holistically the toxicity of a prompt; and have difficulty discerning harm in context-dependent scenarios, particularly with subtle-yet-harmful content (e.g. microaggressions, bias). We release this dataset to contribute to further reduce harmful uses of these models and improve their safe deployment.

LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, Alberto Testoni

There is an increasing trend towards evaluating NLP models with LLMs instead of human judgments, raising questions about the validity of these evaluations, as well as their reproducibility in the case of proprietary models. We provide JUDGE-BENCH, an extensible collection of 20 NLP datasets with human annotations covering a broad range of evaluated properties and types of data, and comprehensively evaluate 11 current LLMs, covering both open-weight and proprietary models, for their ability to replicate the annotations. Our evaluations show substantial variance across models and datasets. Models are reliable evaluators on some tasks, but overall display substantial variability depending on the property being evaluated, the expertise level of the human judges, and whether the language is human or model-generated. We conclude that LLMs should be carefully validated against human judgments before being used as evaluators.

Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation?

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, Sunayana Sitaram

Large Language Models (LLMs) excel in various Natural Language Processing (NLP) tasks, yet their evaluation, particularly in languages beyond the top 20, remains inadequate due to existing benchmarks and metrics limitations. Employing LLMs as evaluators to rank or score other models' outputs emerges as a viable solution, addressing the constraints tied to human annotators and established benchmarks. In this study, we explore the potential of LLM-based evaluators, specifically GPT-4 in enhancing multilingual evaluation by calibrating them against 20K human judgments across three text-generation tasks, five metrics, and eight languages. Our analysis reveals a bias in GPT4-based evaluators towards higher scores, underscoring the necessity of calibration with native speaker judgments, especially in low-resource and non-Latin script languages, to ensure accurate evaluation of LLM performance across diverse languages.

But Disagrees: LLMs-as-Judges

Can Large Language Models Be an Alternative to Human Evaluations?

Cheng-Han Chiang, Hung-yi Lee

Human evaluation is indispensable and inevitable for assessing the quality of texts generated by machine learning models or written by humans. However, human evaluation is very difficult to reproduce and its quality is notoriously unstable, hindering fair comparisons among different natural language processing (NLP) models and algorithms. Recently, large language models (LLMs) have demonstrated exceptional performance on unseen tasks when only the task instructions are provided. In this paper, we explore if such an ability of the LLMs can be used as an alternative to human evaluation. We present the LLMs with the exact same instructions, samples to be evaluated, and questions used to conduct human evaluation, and then ask the LLMs to generate responses to those questions; we dub this LLM evaluation. We use human evaluation and LLM evaluation to evaluate the texts in two NLP tasks: open-ended story generation and adversarial attacks. We show that the result of LLM evaluation is consistent with the results obtained by expert human evaluation: the texts rated higher by human experts are also rated higher by the LLMs. We also find that the results of LLM evaluation are stable over different formatting of the task instructions and the sampling algorithm used to generate the answer. We are the first to show the potential of using LLMs to assess the quality of texts and discuss the limitations and ethical considerations of LLM evaluation.

RTP-LX: Can LLMs Evaluate Toxic

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroorn, Claudet, Pavel Gajdusek, Can Gören, Qilong Gu, Anr Petter Merok, Ivana Milovanović, Nani Paananen, Ve Davide Turcato, Oleksandr Vakhno, Judit Velcsov, An

Large language models (LLMs) and small language models concern. With the advent of multilingual S/LLMs, the quest models with the same velocity at which they are deployed? toxic prompts and outputs in 28 languages. RTP-LX follows culturally-specific toxic language. We evaluate 10 S/LLMs c that, although they typically score acceptably in terms of ao prompt; and have difficulty discerning harm in context-depe We release this dataset to contribute to further reduce harm

Beyond Consensus: Mitigating the Agreeableness Bias in LLM Judge Evaluations

Suryaansh Jain, Umair Z. Ahmed, Shubham Sahai, Ben Leong

New Large Language Models (LLMs) become available every few weeks, and modern application developers confronted with the unenviable task of having to decide if they should switch to a new model. While human evaluation remains the gold standard, it is costly and unscalable. The state-of-the-art approach is to use LLMs as evaluators (LLM-as-a-judge), but this suffers from a critical flaw: LLMs exhibit a strong positive bias. We provide empirical evidence showing that while LLMs can identify valid outputs with high accuracy (i.e., True Positive Rate 96%), they are remarkably poor at identifying invalid ones (i.e., True Negative Rate <25%). This systematic bias, coupled with class imbalance, often leads to inflated reliability scores. While ensemble-based methods like majority voting can help, we show that they are not good enough. We introduce an optimal minority-veto strategy that is resilient to missing data and mitigates this bias to a large extent. For scenarios requiring even higher precision, we propose a novel regression-based framework that directly models the validator bias using a small set of human-annotated ground truth data. On a challenging code feedback task over 366 high-school Python programs, our regression approach reduces the maximum absolute error to just 1.2%, achieving a 2x improvement over the best-performing ensemble of 14 state-of-the-art LLMs.

LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, Alberto Testoni

There is an increasing trend towards evaluating NLP models with LLMs instead of human judgments, raising questions about the validity of these evaluations, as well as their reproducibility in the case of proprietary models. We provide JUDGE-BENCH, an extensible collection of 20 NLP datasets with human annotations covering a broad range of evaluated properties and types of data, and comprehensively evaluate 11 current LLMs, covering both open-weight and proprietary models, for their ability to replicate the annotations. Our evaluations show substantial variance across models and datasets. Models are reliable evaluators on some tasks, but overall display substantial variability depending on the property being evaluated, the expertise level of the human judges, and whether the language is human or model-generated. We conclude that LLMs should be carefully validated against human judgments before being used as evaluators.

Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation?

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Didee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, Sunayana Sitaram

Large Language Models (LLMs) excel in various Natural Language Processing (NLP) tasks, yet their evaluation, particularly in languages beyond the top 20, remains inadequate due to existing benchmarks and metrics limitations. Employing LLMs as evaluators to rank or score other models' outputs emerges as a viable solution, addressing the constraints tied to human annotators and established benchmarks. In this study, we explore the potential of LLM-based evaluators, specifically GPT-4 in enhancing multilingual evaluation by calibrating them against 20K human judgments across three text-generation tasks, five metrics, and eight languages. Our analysis reveals a bias in GPT4-based evaluators towards higher scores, underscoring the necessity of calibration with native speaker judgments, especially in low-resource and non-Latin script languages, to ensure accurate evaluation of LLM performance across diverse languages.

Dual-stage and Lightweight Patient Chart Summarization for Emergency Physicians

Jiajun Wu, Swaleh Zaidi, Braden Teitge, Henry Leung, Jiayu Zhou, Jessalyn Holodinsky, Steve Drew

Electronic health records (EHRs) contain extensive unstructured clinical data that can overwhelm emergency physicians trying to identify critical information. We present a two-stage summarization system that runs entirely on embedded devices, enabling offline clinical summarization while preserving patient privacy. In our approach, a dual-device architecture first retrieves relevant patient record sections using the Jetson Nano-R (Retrieve), then generates a structured summary on another Jetson Nano-S (Summarize), communicating via a lightweight socket link. The summarization output is two-fold: (1) a fixed-format list of critical findings, and (2) a context-specific narrative focused on the clinician's query. The retrieval stage uses locally stored EHRs, splits long notes into semantically coherent sections, and searches for the most relevant sections per query. The generation stage uses a locally hosted small language model (SLM) to produce the summary from the retrieved text, operating within the constraints of two NVIDIA Jetson devices. We first benchmarked six open-source SLMs under 7B parameters to identify viable models. We incorporated an LLM-as-Judge evaluation mechanism to assess summary quality in terms of factual accuracy, completeness, and clarity. Preliminary results on MIMIC-IV and de-identified real EHRs demonstrate that our fully offline system can effectively produce useful summaries in under 30 seconds.

Adrian de Wynter (adewynter@microsoft.com)

But Disagrees: LLMs-as-Judges



Can Large Language Models Be an Alternative to Human Evaluations?

Cheng-Han Chiang, Hung-yi Lee

Human evaluation is indispensable and inevitable for assessing the quality of texts generated by machine learning models or written by humans. However, human evaluation is very difficult to reproduce and its quality is notoriously unstable, hindering fair comparisons among different natural language processing (NLP) models and algorithms. Recently, large language models (LLMs) have demonstrated exceptional performance on unseen tasks when only the task instructions are provided. In this paper, we explore if such an ability of the LLMs can be used as an alternative to human evaluation. We present the LLMs with the exact same instructions, samples to be evaluated, and questions used to conduct human evaluation, and then ask the LLMs to generate responses to those questions; we dub this LLM evaluation. We use human evaluation and LLM evaluation to evaluate the texts in two NLP tasks: open-ended story generation and adversarial attacks. We show that the result of LLM evaluation is consistent with the results obtained by expert human evaluation; the texts rated higher by human experts are also rated higher by the LLMs. We also find that the results of LLM evaluation are stable over different formatting of the task instructions and the sampling algorithm used to generate the answer. We are the first to show the potential of using LLMs to assess the quality of texts and discuss the limitations and ethical considerations of LLM evaluation.

LeMAJ (Legal LLM-as-a-Judge): Bridging Legal Reasoning and LLM Evaluation

Joseph Enguehard, Morgane Van Ermengem, Kate Atkinson, Sujeong Cha, Arijit Ghosh Chowdhury, Prashanth Kallur Ramaswamy, Jeremy Roghair, Hannah R Marlowe, Carina Suzana Negreanu, Kitty Boxall, Diana Mincu

Evaluating large language model (LLM) outputs in the legal domain presents unique challenges due to the complex and nuanced nature of legal analysis. Current evaluation approaches either depend on reference data, which is costly to produce, or use standardized assessment methods, both of which have significant limitations for legal applications. Although LLM-as-a-Judge has emerged as a promising evaluation technique, its reliability and effectiveness in legal contexts depend heavily on evaluation processes unique to the legal industry and how trustworthy the evaluation appears to the human legal expert. This is where existing evaluation methods currently fail and exhibit considerable variability. This paper aims to close the gap: a) we break down lengthy responses into 'Legal Data Points' (LDPs), self-contained units of information, and introduce a novel, reference-free evaluation methodology that reflects how lawyers evaluate legal answers; b) we demonstrate that our method outperforms a variety of baselines on both our proprietary dataset and an open-source dataset (LegalBench); c) we show how our method correlates more closely with human expert evaluations and helps improve inter-annotator agreement; and finally d) we open source our Legal Data Points for a subset of LegalBench used in our experiments, allowing the research community to replicate our results and advance research in this vital area of LLM evaluation on legal question-answering.

Dual-stage and Lightweight Patient Chart Summarization for Emergency Physicians

Jiajun Wu, Swaleh Zaidi, Braden Teitge, Henry Leung, Jiayu Zhou, Jessalyn Holodinsky, Steve Drew

Electronic health records (EHRs) contain extensive unstructured clinical data that can overwhelm emergency physicians trying to identify critical information. We present a two-stage summarization system that runs entirely on embedded devices, enabling offline clinical summarization while preserving patient privacy. In our approach, a dual-device architecture first retrieves relevant patient record sections using the Jetson Nano-R (Retrieve), then generates a structured summary on another Jetson Nano-S (Summarize), communicating via a lightweight socket link. The summarization output is two-fold: (1) a fixed-format list of critical findings, and (2) a context-specific narrative focused on the clinician's query. The retrieval stage uses locally stored EHRs, splits long notes into semantically coherent sections, and searches for the most relevant sections per query. The generation stage uses a locally hosted small language model (SLM) to produce the summary from the retrieved text, operating within the constraints of two NVIDIA Jetson devices. We first benchmarked six open-source SLMs under 7B parameters to identify viable models. We incorporated an LLM-as-Judge evaluation mechanism to assess summary quality in terms of factual accuracy, completeness, and clarity. Preliminary results on MIMIC-IV and de-identified real EHRs demonstrate that our fully offline system can effectively produce useful summaries in under 30 seconds.

RTP-LX: Can LLMs Evaluate Toxic

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroorn, Claudet, Pavel Gajdusek, Can Gören, Qilong Gu, Anr Petter Merok, Ivana Milovanović, Nani Paananen, Ve Davide Turcato, Oleksandr Vakhno, Judit Velcsov, An

Large language models (LLMs) and small language models concern. With the advent of multilingual SLLMs, the quest models with the same velocity at which they are deployed? toxic prompts and outputs in 28 languages. RTP-LX follows culturally-specific toxic language. We evaluate 10 SLLMs c that, although they typically score acceptably in terms of ao prompt; and have difficulty discerning harm in context-depe We release this dataset to contribute to further reduce harm

Beyond Consensus: Mitigating the Agreeableness Bias in LLM Judge Evaluations

Suryaansh Jain, Umair Z. Ahmed, Shubham Sahai, Ben Leong

New Large Language Models (LLMs) become available every few weeks, and modern application developers confronted with the unenviable task of having to decide if they should switch to a new model. While human evaluation remains the gold standard, it is costly and unscalable. The state-of-the-art approach is to use LLMs as evaluators (LLM-as-a-judge), but this suffers from a critical flaw: LLMs exhibit a strong positive bias. We provide empirical evidence showing that while LLMs can identify valid outputs with high accuracy (i.e., True Positive Rate 96%), they are remarkably poor at identifying invalid ones (i.e., True Negative Rate <25%). This systematic bias, coupled with class imbalance, often leads to inflated reliability scores. While ensemble-based methods like majority voting can help, we show that they are not good enough. We introduce an optimal minority-veto strategy that is resilient to missing data and mitigates this bias to a large extent. For scenarios requiring even higher precision, we propose a novel regression-based framework that directly models the validator bias using a small set of human-annotated ground truth data. On a challenging code feedback task over 366 high-school Python programs, our regression approach reduces the maximum absolute error to just 1.2%, achieving a 2x improvement over the best-performing ensemble of 14 state-of-the-art LLMs.

LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, Alberto Testi

There is an increasing trend towards evaluating NLP models with LLMs instead of human judges. evaluations, as well as their reproducibility in the case of proprietary models. We provide JUDGE with human annotations covering a broad range of evaluated properties and types of data, and both open-weight and proprietary models, for their ability to replicate the annotations. Our evaluation datasets. Models are reliable evaluators on some tasks, but overall display substantial variability: expertise level of the human judges, and whether the language is human or model-generated. against human judgments before being used as evaluators.

Humans or LLMs as the Judge? A Study on Judgement Biases

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, Benyou Wang

Adopting human and large language models (LLM) as judges (a.k.a human- and LLM-as-a-judge) for evaluating the performance of LLMs has recently gained attention. Nonetheless, this approach concurrently introduces potential biases from human and LLMs, questioning the reliability of the evaluation results. In this paper, we propose a novel framework that is free from referencing groundtruth annotations for investigating Misinformation Oversight Bias, Gender Bias, Authority Bias and Beauty Bias on LLM and human judges. We curate a dataset referring to the revised Bloom's Taxonomy and conduct thousands of evaluations. Results show that human and LLM judges are vulnerable to perturbations to various degrees, and that even the cutting-edge judges possess considerable biases. We further exploit these biases to conduct attacks on LLM judges. We hope that our work can notify the community of the bias and vulnerability of human- and LLM-as-a-judge, as well as the urgency of developing robust evaluation systems.

Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation?

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, Sunayana Sitaram

Large Language Models (LLMs) excel in various Natural Language Processing (NLP) tasks, yet their evaluation, particularly in languages beyond the top 20, remains inadequate due to existing benchmarks and metrics limitations. Employing LLMs as evaluators to rank or score other models' outputs emerges as a viable solution, addressing the constraints tied to human annotators and established benchmarks. In this study, we explore the potential of LLM-based evaluators, specifically GPT-4 in enhancing multilingual evaluation by calibrating them against 20K human judgments across three text-generation tasks, five metrics, and eight languages. Our analysis reveals a bias in GPT4-based evaluators towards higher scores, underscoring the necessity of calibration with native speaker judgments, especially in low-resource and non-Latin script languages, to ensure accurate evaluation of LLM performance across diverse languages.

Adrian de Wynter (adewynter@microsoft.com)

But Disagrees: LLMs-as-Judges

Can Large Language Models Be an Alternative to Human Evaluations?

Cheng-Han Chiang, Hung-yi Lee

Human evaluation is indispensable and inevitable for assessing the quality of texts generated by machine learning models or written by humans. However, human evaluation is very difficult to reproduce and its quality is notoriously unstable, hindering fair comparisons among different natural language processing (NLP) models and algorithms. Recently, large language models (LLMs) have demonstrated exceptional performance on unseen tasks when only the task instructions are provided. In this paper, we explore if such an ability of the LLMs can be used as an alternative to human evaluation. We present the LLMs with the exact same instructions, samples to be evaluated, and questions used to conduct human evaluation, and then ask the LLMs to generate responses to those questions; we dub this LLM evaluation. We use human evaluation and LLM evaluation to evaluate the texts in two NLP tasks: open-ended story generation and adversarial attacks. We show that the result of LLM evaluation is consistent with the results obtained by expert human evaluation: the texts rated higher by human experts are also rated higher by the LLMs. We also find that the results of LLM evaluation are stable over different formatting of the task instructions and the sampling algorithm used to generate the answer. We are the first to show the potential of using LLMs to assess the quality of texts and discuss the limitations and ethical considerations of LLM evaluation.

JudgeBench: A Benchmark for Evaluating LLM-Based Judges

Evaluating LLM-based judges for factual and logical correctness

[\[Paper\]](#) • [\[Github\]](#) • [\[Dataset\]](#) • [\[Leaderboard\]](#)

[GPT-4o Dataset](#) Claude-3.5-Sonnet Dataset

	Category	Knowledge Sc...	Reasoning Sc...	Math Sco...	Coding Sc...	Overall Sc...
11 (high))	Prompted Judge	67.5	89.8	87.5	100	80.9
GenRM + voting@32	Fine-Tuned Jud	70.8	83.7	87.5	83.3	78.6
11 (medium))	Prompted Judge	62.3	86.7	85.7	92.9	76.6
GenRM-Multilingual + vo	Fine-Tuned Jud	65.6	82.7	87.5	85.7	76.3
9-12)	Prompted Judge	66.2	79.6	85.7	85.7	75.4
GenRM	Fine-Tuned Jud	71.4	73.5	87.5	76.2	75.1
1	Reward Model	70.8	76.5	82.1	66.7	73.7
9)	Prompted Judge	65.6	82.7	88.4	97.0	79.1

Dual-stage and Lightweight Patient Chart Summarization for Emergency Physicians

Jiajun Wu, Swaleh Zaidi, Braden Teitge, Henry Leung, Jiayu Zhou, Jessalyn Holodinsky, Steve Drew

Electronic health records (EHRs) contain extensive unstructured clinical data that can overwhelm emergency physicians trying to identify critical information. We present a two-stage summarization system that runs entirely on embedded devices, enabling offline clinical summarization while preserving patient privacy. In our approach, a dual-device architecture first retrieves relevant patient record sections using the Jetson Nano-R (Retrieve), then generates a structured summary on another Jetson Nano-S (Summarize), communicating via a lightweight socket link. The summarization output is two-fold: (1) a fixed-format list of critical findings, and (2) a context-specific narrative focused on the clinician's query. The retrieval stage uses locally stored EHRs, splits long notes into semantically coherent sections, and searches for the most relevant sections per query. The generation stage uses a locally hosted small language model (SLM) to produce the summary from the retrieved text, operating within the constraints of two NVIDIA Jetson devices. We first benchmarked six open-source SLMs under 7B parameters to identify viable models. We incorporated an LLM-as-Judge evaluation mechanism to assess summary quality in terms of factual accuracy, completeness, and clarity. Preliminary results on MIMIC-IV and de-identified real EHRs demonstrate that our fully offline system can effectively produce useful summaries in under 30 seconds.

RTP-LX: Can LLMs Evaluate Toxic

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroorn, Claudet, Pavel Gajdusek, Can Gören, Qilong Gu, Anr Petter Merok, Ivana Milovanović, Nani Paananen, Ve Davide Turcato, Oleksandr Vakhno, Judit Velcsov, An

Large language models (LLMs) and small language models concern. With the advent of multilingual SLLMs, the questi models with the same velocity at which they are deployed? toxic prompts and outputs in 28 languages. RTF culturally-specific toxic language. We evaluate 1 that, although they typically score acceptably in prompt; and have difficulty discerning harm in o We release this dataset to contribute to further r

LLMs instead of Human Judge Evaluation Tasks

Anna Bavaresco, Raffaella Bernardi, Leonardo F Giulianelli, Michael Hanna, Alexander Koller, An David Schlangen, Alessandro Suglia, Aditya K S

There is an increasing trend towards evaluating NLP evaluations, as well as their reproducibility in the cas with human annotations covering a broad range of ev both open-weight and proprietary models, for their ab datasets. Models are reliable evaluators on some tas expertise level of the human judges, and whether the language is human or model-generated. against human judgments before being used as evaluators.

Beyond Consensus: Mitigating the Agreeableness Bias in LLM Judge Evaluations

Suryaansh Jain, Umair Z. Ahmed, Shubham Sahai, Ben Leong

New Large Language Models (LLMs) become available every few weeks, and modern application developers confronted with the unenviable task of having to decide if they should switch to a new model. While human evaluation remains the gold standard, it is costly and unscalable. The state-of-the-art approach is to use LLMs as evaluators (LLM-as-a-judge), but this suffers from a critical flaw: LLMs exhibit a strong positive bias. We provide empirical evidence showing that while LLMs can identify valid outputs with high accuracy (i.e., True Positive Rate 96%), they are remarkably poor at

Neither Valid nor Reliable? Investigating the Use of LLMs as Judges

Khaoula Chehbouni, Mohammed Haddou, Jackie Chi Kit Cheung, Golnoosh Farnadi

Evaluating natural language generation (NLG) systems remains a core challenge of natural language processing (NLP), further complicated by the rise of large language models (LLMs) that aims to be general-purpose. Recently, large language models as judges (LLJs) have emerged as a promising alternative to traditional metrics, but their validity remains underexplored. This position paper argues that the current enthusiasm around LLJs may be premature, as their adoption has outpaced rigorous scrutiny of their reliability and validity as evaluators. Drawing on measurement theory from the social sciences, we identify and critically assess four core assumptions underlying the use of LLJs: their ability to act as proxies for human judgment, their capabilities as evaluators, their scalability, and their cost-effectiveness. We examine how each of these assumptions may be challenged by the inherent limitations of LLMs, LLJs, or current practices in NLG evaluation. To ground our analysis, we explore three applications of LLJs: text summarization, data annotation, and safety alignment. Finally, we highlight the need for more responsible evaluation practices in LLJs evaluation, to ensure that their growing role in the field supports, rather than undermines, progress in NLG.

results. In this paper, we propose a novel framework that is free from referencing groundtruth annotations for investigating Misinformation Oversight Bias, Gender Bias, Authority Bias and Beauty Bias on LLM and human judges. We curate a dataset referring to the revised Bloom's Taxonomy and conduct thousands of evaluations. Results show that human and LLM judges are vulnerable to perturbations to various degrees, and that even the cutting-edge judges possess considerable biases. We further exploit these biases to conduct attacks on LLM judges. We hope that our work can notify the community of the bias and vulnerability of human- and LLM-as-a-judge, as well as the urgency of developing robust evaluation systems.

Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation?

Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, Sunayana Sitaram

TrustJudge: Inconsistencies of LLM-as-a-Judge and How to Alleviate Them in Bias in Communication Systems

Yidong Wang, Yunze Song, Tingyuan Zhu, Xuanwang Zhang, Zhuohao Yu, Hao Chen, Chiyu Song, Qiufeng Wang, Cunxiang Wang, Zhen Qian Wang, Wuxu Dai, Yue Zhang, Wei Ye, Shikun Zhang

The adoption of Large Language Models (LLMs) as automated evaluators (LLM-as-a-judge) has revealed critical inconsistencies in current evaluation frameworks. We identify two fundamental types of inconsistencies: (1) Score-Comparison Inconsistency, where lower-rated responses outperform higher-scored ones in pairwise comparisons, and (2) Pairwise Transitivity Inconsistency, manifested through circular preference chains (A>B>C>A) and equivalence contradictions (A=B<C≠q A). We argue that these issues come from information loss in discrete rating systems and ambiguous tie judgments during pairwise evaluation. We propose TrustJudge, a probabilistic framework that addresses these limitations through two key innovations: 1) distribution-sensitive scoring that computes continuous expectations from discrete rating probabilities, preserving information entropy for more precise scoring, and 2) likelihood-aware aggregation that resolves transitivity violations using bidirectional preference probabilities or perplexity. We also formalize the theoretical limitations of current LLM-as-a-judge frameworks and demonstrate how TrustJudge's components overcome them. When evaluated with Llama-3.1-70B-Instruct as judge using our dataset, TrustJudge reduces Score-Comparison inconsistency by 8.43% (from 23.32% to 14.89%) and Pairwise Transitivity inconsistency by 10.82% (from 15.22% to 4.40%), while maintaining higher evaluation accuracy. Our work provides the first systematic analysis of evaluation framework inconsistencies in LLM-as-a-judge paradigms, offering both theoretical insights and practical solutions for reliable automated assessment. The framework demonstrates consistent improvements across various model architectures and scales, enabling more trustworthy LLM evaluation without requiring additional training or human annotations. The codes can be found at [this https URL](#).

Adrian de Wynter (adewynter@microsoft.com)

LLM Evaluation is Flawed

- Benchmarks are contaminated [1] or gamed [2]
- Evaluation of LLMs leads to unfair comparisons between model sizes [3]

◆ AI Overview

Yes, companies can "cheat" in LLM benchmarks through various methods, including data leakage, where models are trained on benchmark data, and overfitting, where models are optimized for the benchmark rather than general performance. Other techniques involve "gaming" automated benchmarks by manipulating output style or length, or even using simple, consistent "null models" that can achieve high scores on certain automated evaluations by exploiting the benchmark's design. [🔗](#)

Types of cheating and manipulation

- **Data leakage:** Models are accidentally or intentionally trained on the data used for the benchmark, leading to inflated scores that don't reflect real-world performance. This can also happen if benchmark data is leaked publicly, and then becomes part of the training set. [🔗](#)
- **Overfitting to the benchmark:** The model is fine-tuned specifically to perform well on a particular benchmark, but it fails to generalize to real-world tasks. This is a common problem in machine learning, often seen as "teaching to the test," and is made worse by having real-time access to benchmark results during training, as seen in Kaggle contests, notes [this Substack post](#). [🔗](#)
- **Automated benchmark manipulation:** Developers can manipulate model outputs to increase their win rates on automated benchmarks, such as by controlling output length and style. Even a simple "null model" can sometimes cheat by giving a consistent answer, as demonstrated in a [study on arXiv](#). [🔗](#)

[1] Sainz et al., [NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark](#)

[2] Reisner, A. [Chatbots Are Cheating on Their Benchmark Tests](#)

[3] Yu, G. et al., [Rethinking the Evaluation of In-Context Learning for LLMs](#)

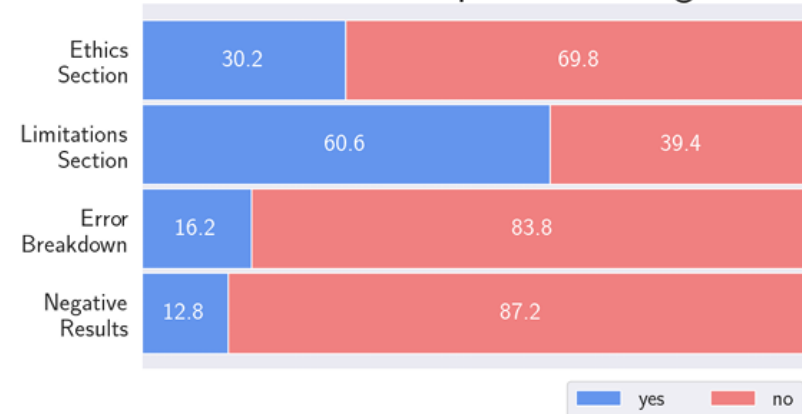
...And Lacks Rigour

Claims are noisy because:

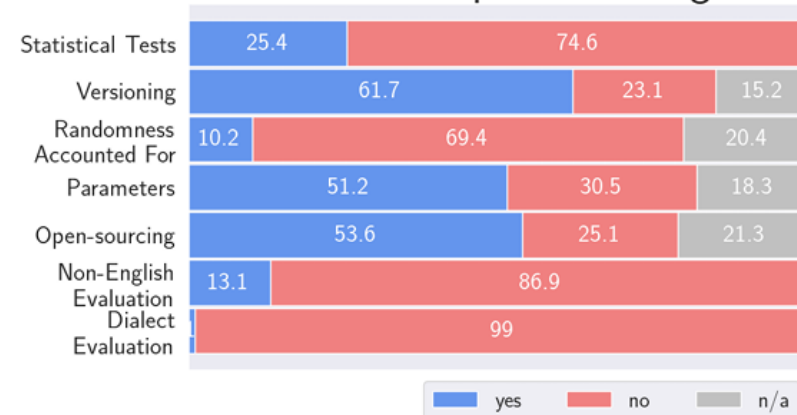
- Most studies lack statistical tests, disclosures of parameters, or error breakdowns.
- This is actually an *increasing* trend [1]

The definitions of ‘reasoning’ and ‘emergence’ are all over the place [2].

Structural Features for Papers Claiming SOTA



Research Features for Papers Claiming SOTA



[1] De Wynter, A. [Awes, Laws, and Flaws From Today's LLM Research](#)

[2] Huang and Chang, [Towards Reasoning in Large Language Models: A Survey](#)

Including in FLT :<

- Supposedly a less-realistic, better controlled evaluation
- The transformer has difficulties with languages like PARITY
 - Although it *can* learn it
- Some context-free languages are too difficult for LLMs
 - Although they *can* learn them
- Predicting learnability of LMs is difficult, although correlated to some measurable quantities

[1] Borenstein, N. et al., [What Languages are Easy to Language-Model? A Perspective from Learning Probabilistic Regular Languages](#)

[2] Deletang, G. et al., [Neural Networks and the Chomsky Hierarchy](#)

[3] De Wynter, A. [Is In-Context Learning Learning?](#)

[4] Butoi, A. et al., [Training Neural Networks as Recognizers of Formal Languages](#)

Including in FLT :<

- Supposedly a less-realistic, better controlled evaluation
- The transformer has difficulties with languages like PARITY
 - Although it *can* learn it
- Some context-free languages are too difficult for LLMs
 - Although they *can* learn them
- Predicting learnability of LMs is difficult, although correlated to some measurable quantities

Empiricism appears to beat theory on this one

[1] Borenstein, N. et al., [What Languages are Easy to Language-Model? A Perspective from Learning Probabilistic Regular Languages](#)

[2] Deletang, G. et al., [Neural Networks and the Chomsky Hierarchy](#)

[3] De Wynter, A. [Is In-Context Learning Learning?](#)

[4] Butoi, A. et al., [Training Neural Networks as Recognizers of Formal Languages](#)

Including in FLT :<

HOWEVER, empirical studies have **even *more* problems**.

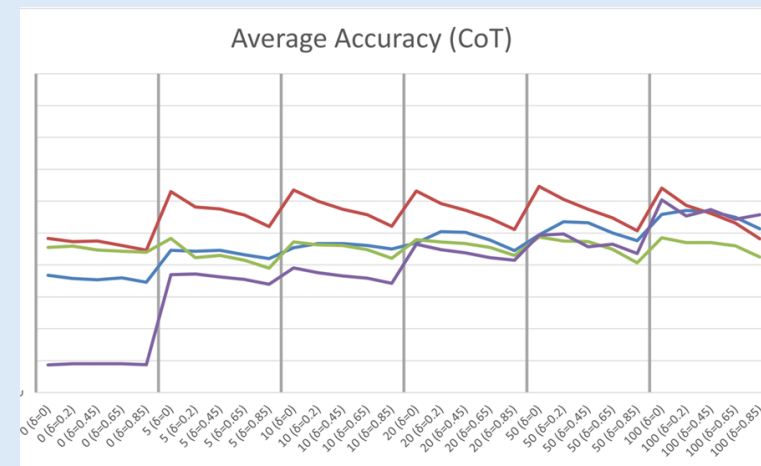
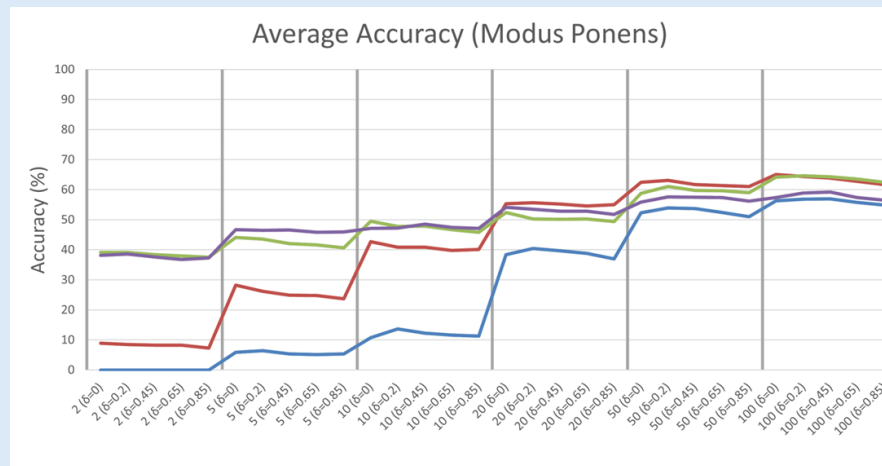
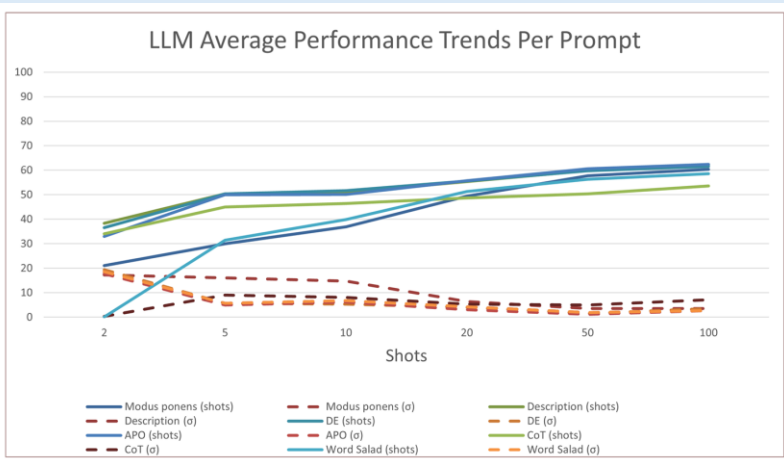
They typically lack one or more of the following ablations:

1. Sensitivity to natural language
2. Impact of exemplar distributions
3. Impact of exemplar *volume*
4. Robustness to distributional shifts

There is something there,
however

LLMs are **NOT** ‘Just’ Next-Token Prediction Engines

- With synthetic data scenarios, one can show that:
 - Almost all prompts learn the given (non-natural language) problem.
 - Almost all models learn the same way
 - They are insensitive to in-prompt exemplar distribution
 - They are all brittle to out-of-distribution



But Aren't Great

- They overfixate in statistical features of the prompt
- They are, however, **extremely convincing**

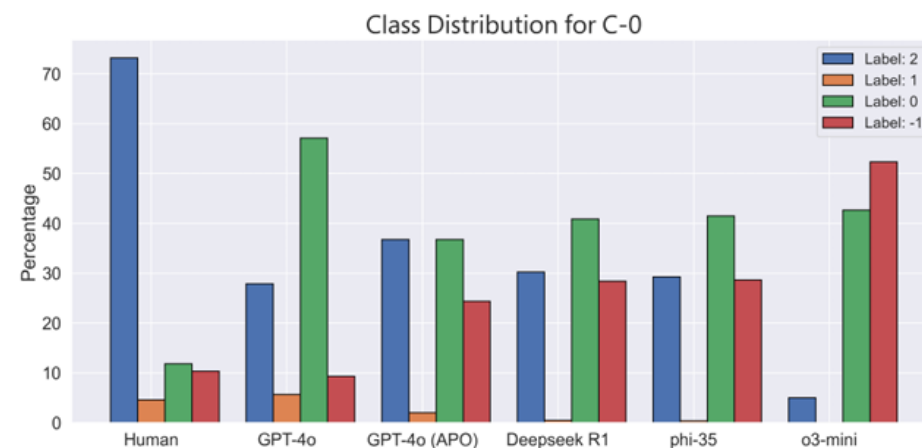


Figure 4: Class distribution for C-0 between humans and LLMs. All LLMs favour label 0 ('no reasons provided to support the thesis') more than others, and, for some, they have a high incidence of 'not an argument' outputs (red rightmost column on every group). In contrast, human annotators typically preferred label 2 ('there are reasons provided to support the thesis').

Evaluating supporting arguments for theses shows little agreement with LLMs, even under APO

Adrian de Wynter (adewynter@microsoft.com)

[1] De Wynter, A. [Is In-Context Learning Learning?](#)

[2] De Wynter, A. and Yuan, T. [The Thin Line Between Comprehension and Persuasion in LLMs](#)

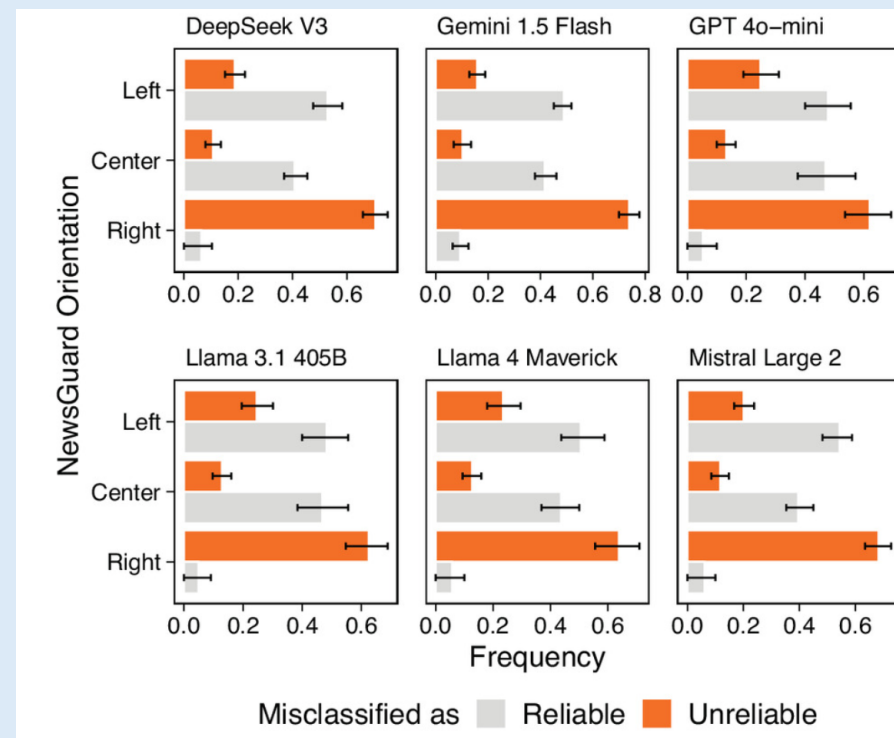
[3] Loru, E. et al., [The simulation of judgment in LLMs](#)

[4] Pletenev, S., et al. [How Much Knowledge Can You Pack into a LoRA Adapter without Harming LLM?](#)

[5] Xiao et al., [On the Role of Difficult Prompts in Self-Play Preference Optimization](#)

But Aren't Great

- They overfixate in statistical features of the prompt
- They are, however, **extremely convincing**



LLMs-as-judges are known to emphasise the natural-language representation of the prompt, but not the arguments

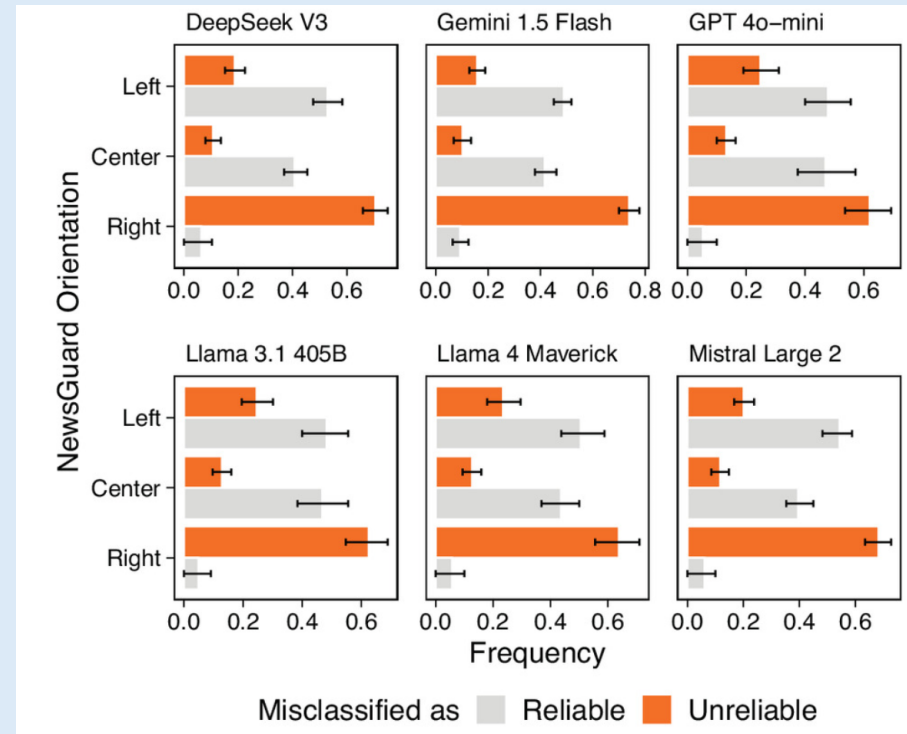
- [1] De Wynter, A. [Is In-Context Learning Learning?](#)
- [2] De Wynter, A. and Yuan, T. [The Thin Line Between Comprehension and Persuasion in LLMs](#)
- [3] Loru, E. et al., [The simulation of judgment in LLMs](#)
- [4] Pletenev, S., et al. [How Much Knowledge Can You Pack into a LoRA Adapter without Harming LLM?](#)
- [5] Xiao et al., [On the Role of Difficult Prompts in Self-Play Preference Optimization](#)

But Aren't Great

- They overfixate in statistical features of the prompt
- They are, however, **extremely convincing**

Specialisation is difficult!

- In many ways, ICL is inferior to SFT.
- Practical methods like LoRA are mostly meant as a behavioural alignment tool



LLMs-as-judges are known to emphasise the natural-language representation of the prompt, but not the arguments

- [1] De Wynter, A. [Is In-Context Learning Learning?](#)
[2] De Wynter, A. and Yuan, T. [The Thin Line Between Comprehension and Persuasion in LLMs](#)
[3] Loru, E. et al., [The simulation of judgment in LLMs](#)
[4] Pletenev, S., et al. [How Much Knowledge Can You Pack into a LoRA Adapter without Harming LLM?](#)
[5] Xiao et al., [On the Role of Difficult Prompts in Self-Play Preference Optimization](#)

Core Insight

Applied papers (and companies) typically show **good** LLMs performances, *especially* in LLMs-as-judges

Meta-research papers show the **opposite**, *especially* in LLMs-as-judges

Core Insight

ALL focus on per-task evaluation, and assume it is per-class generalization!

Interpreting LLM capabilities
requires **NUANCE!**

Interpreting LLM capabilities requires **NUANCE!**

But how do you scale?

1. Evaluating and interpreting LLM capabilities requires good measurements

Interpreting LLM capabilities requires **NUANCE!**

But how do you scale?

1. Evaluating and interpreting LLM capabilities requires good measurements
2. Measurements need to keep up with LLM capabilities

Interpreting LLM capabilities requires **NUANCE!**

But how do you scale?

1. Evaluating and interpreting LLM capabilities requires good measurements
2. Measurements need to keep up with LLM capabilities
3. Best tool so far that can keep up: an LLM

Interpreting LLM capabilities requires **NUANCE!**

But how do you scale?

1. Evaluating and interpreting LLM capabilities requires good measurements
2. Measurements need to keep up with LLM capabilities
3. Best tool so far that can keep up: an LLM
4. GOTO 1 🧠

So What Can We Do?

Ensure your judges are trustworthy!

So What Can We Do?

Ensure your judges are trustworthy!

... and stop measuring capabilities per problem class

The Ways to Evaluate LLMs

- LLMs tell you **something**
- It is up to you to decide to trust them

The Ways to Evaluate LLMs

- LLMs tell you **something**
- It is up to you to decide to trust them

There are three ways to establish this trust

The Ways to Evaluate LLMs

- LLMs tell you **something**
- It is up to you to decide to trust them

There are three ways to establish this trust

1. Statistics,

The Ways to Evaluate LLMs

- LLMs tell you **something**
- It is up to you to decide to trust them

There are three ways to establish this trust

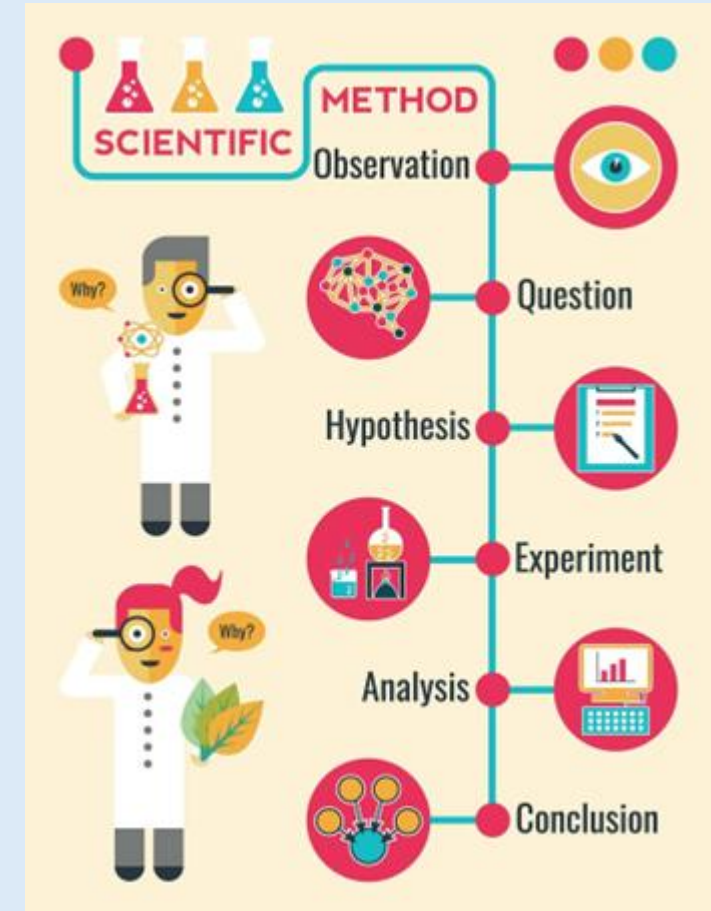
1. Statistics,
2. Formal proofs, or

The Ways to Evaluate LLMs

- LLMs tell you **something**
- It is up to **you** to decide to trust them

There are three ways to establish this trust

1. Statistics,
2. Formal proofs, or
3. Assuming the results are true :D

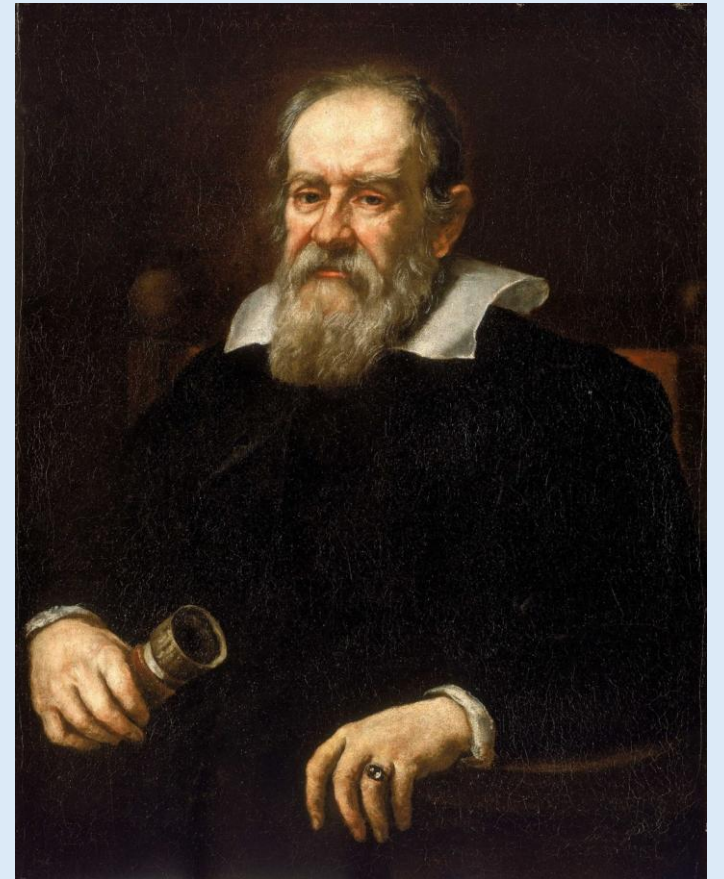


The Ways to Evaluate LLMs

- LLMs tell you ***something***
- It is up to **you** to decide to trust them

There are three ways to establish this trust

1. Statistics,
2. Formal proofs, or
3. Assuming the results are true :D



Galileo is very disappointed in us

Algorithmic Trust in LLMs-as-Judges

A case study

Algorithmic Trust in LLMs-as-Judges

- Whether you like it or not, LLMs lie to you.

Or rather, you have no way of knowing if they are lying to you if **you** do not know the truth.

Algorithmic Trust in LLMs-as-Judges

- Whether you like it or not, LLMs lie to you.

Or rather, you have no way of knowing if they are lying to you if **you** do not know the truth.

If you knew the truth then you wouldn't need to ask

Algorithmic Trust in LLMs-as-Judges

- Whether you like it or not, LLMs lie to you.

Or rather, you have no way of knowing if they are lying to you if **you** do not know the truth.

If you knew the truth then you wouldn't need to ask

- This algorithm is a cryptographically-secure way to enable trust in an LLM-as-a-judge.
- 'Judge the judges!'

Algorithmic Trust in LLMs-as-Judges

Example:

You want to label a West Frisian writing assistance dataset

1. You do not have **any** labels
2. You **do** know what the labels should be (**a rubric**), e.g.:
 1. Is the response in West Frisian?
 2. Is it culturally relevant?
 3. Is it correct?
3. You do not know which parts of the rubric are more important (because you **don't have any labels**)



Algorithmic Trust in LLMs-as-Judges

If you knew which parts of the rubric are important,
then you would know how to label the data.

But you don't.

Algorithmic Trust in LLMs-as-Judges

If you knew which parts of the rubric are important, then you would know how to label the data.

But you don't.

Someone comes and says: 'I know how to label the data' :D

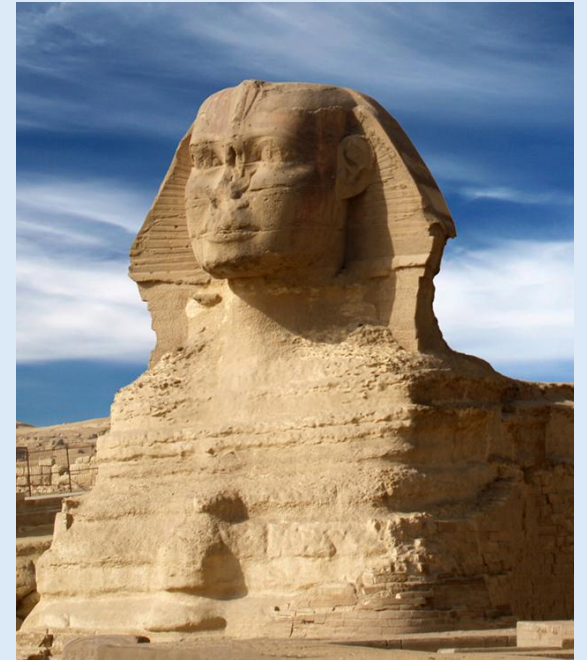
Algorithmic Trust in LLMs-as-Judges

If you knew which parts of the rubric are important, then you would know how to label the data.

But you don't.

Someone comes and says: 'I know how to label the data' :D

So you pose a bunch of challenges to decide if you can trust them

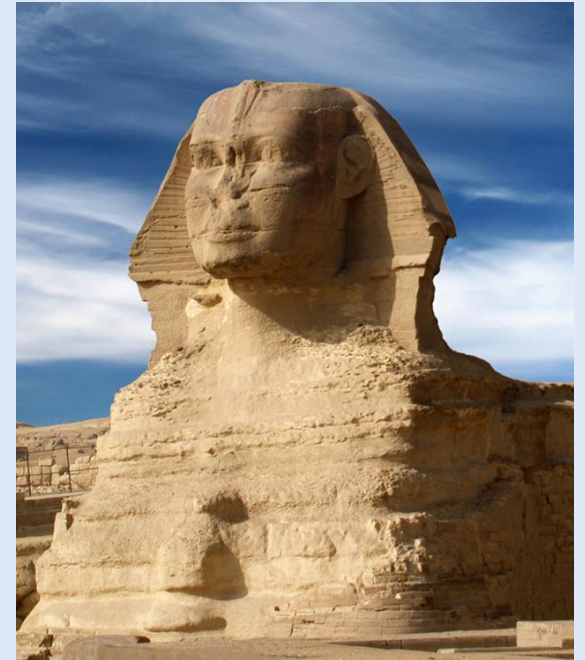


In this scenario, you are the sphynx

Algorithmic Trust in LLMs-as-Judges

If they pass the challenges, you
may trust their labels.

Otherwise, off to the bin with
them.



In this scenario, you are the sphynx

Algorithmic Trust in LLMs-as-Judges

- In short, you have a rubric that you use to annotate the data.

Algorithmic Trust in LLMs-as-Judges

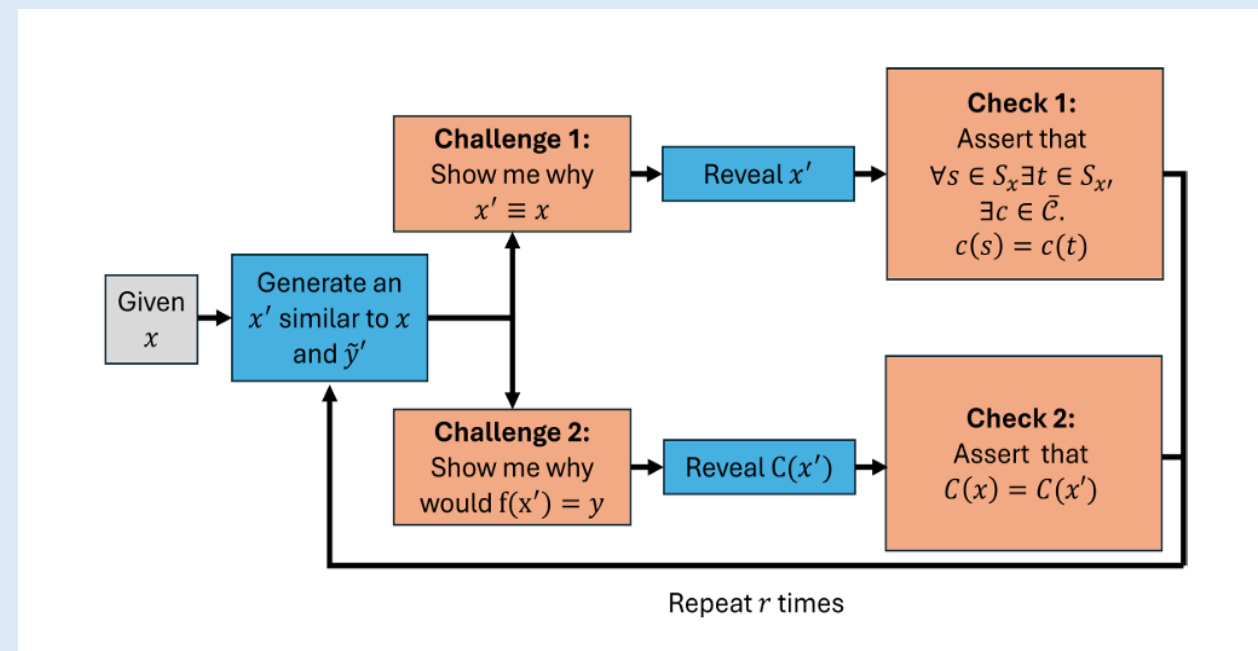
- In short, you have a rubric that you use to annotate the data.
- Given a datapoint, the LLM claims it knows the label.

Algorithmic Trust in LLMs-as-Judges

- In short, you have a rubric that you use to annotate the data.
- Given a datapoint, the LLM claims it knows the label.
- You pose two challenges (uniformly at random) r times to establish trust.

Algorithmic Trust in LLMs-as-Judges

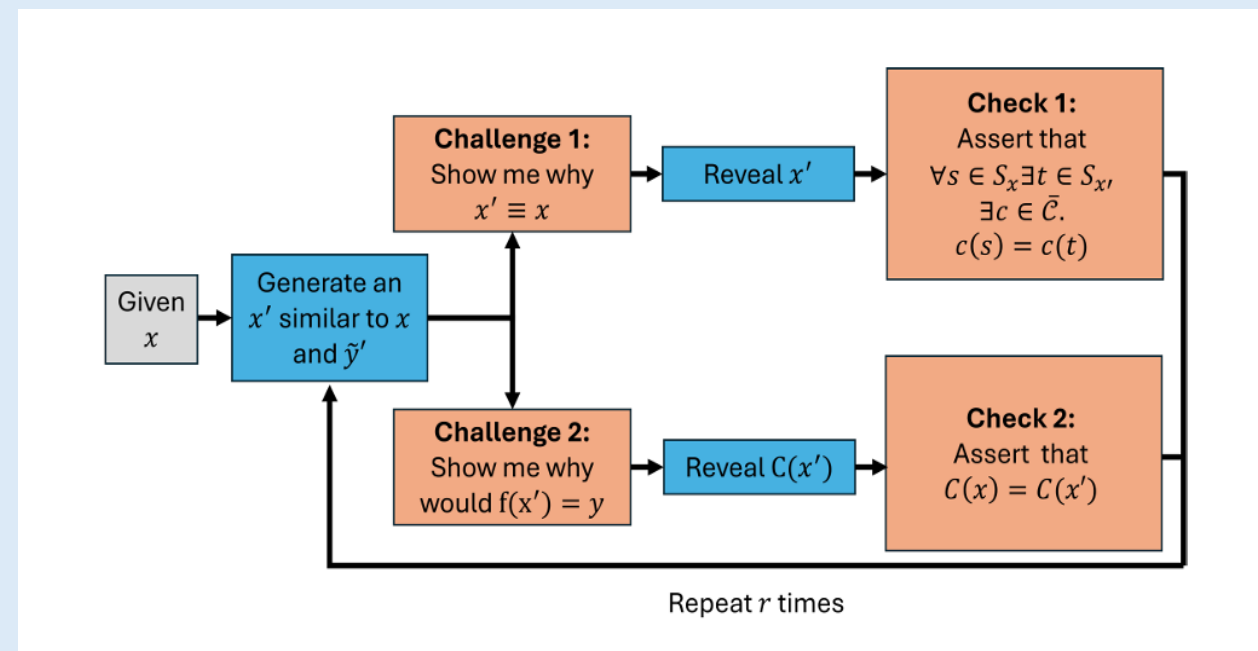
- In short, you have a rubric that you use to annotate the data.
- Given a datapoint, the LLM claims it knows the label.
- You pose two challenges (uniformly at random) r times to establish trust.
- Probability of catching a liar: $1 - \left(\frac{1}{4}\right)^r$



Algorithmic Trust in LLMs-as-Judges

- In short, you have a rubric that you use to annotate the data.
- Given a datapoint, the LLM claims it knows the label.
- You pose two challenges (uniformly at random) r times to establish trust.
- Probability of catching a liar: $1 - \left(\frac{1}{4}\right)^r$

It works!



Algorithmic Trust in LLMs-as-Judges

- West Frisian dataset
- Two LLMs as evaluators:
 - GPT-4.1
 - Qwen-2.5B[†]
- One verifier:
 - GPT-4.1

Algorithmic Trust in LLMs-as-Judges

- West Frisian dataset
- Two LLMs as evaluators:
 - GPT-4.1
 - Qwen-2.5B[†]
- One verifier:
 - GPT-4.1
- First:
 - (IP) Correctly-annotated data
- Second:
 - (OOP) Incorrectly-annotated data

Algorithmic Trust in LLMs-as-Judges

- West Frisian dataset
- Two LLMs as evaluators:
 - GPT-4.1
 - Qwen-2.5B[†]
- One verifier:
 - GPT-4.1
- First:
 - (IP) Correctly-annotated data
- Second:
 - (OOP) Incorrectly-annotated data

	IP Acc. / F1	OOP Acc. / F1	IP Succ. / Flips	OOP Succ. / Flips
Known				
Predicting y	76.3 / 80.3	51.5 / 67.2	—/—	—/—
Avg. per $c \in \bar{C}$	89.5 / 94.0	45.3 / 46.5	—/—	—/—
Unknown				
GPT-4.1				
$\phi = 0.7$	72.4 / 76.6	49.3 / 57.8	87.8 / 9.3	1.4 / 70.3
$\phi = 0.3$	74.4 / 78.1	49.1 / 39.1	86.8 / 4.9	1.2 / 33.6
$\phi = 0.1$	76.9 / 79.8	50.7 / 16.4	86.2 / 1.8	1.2 / 9.1
Qwen 2.5B				
$\phi = 0.7$	62.5 / 63.9	51.7 / 59.5	34.4 / 46.0	1.8 / 70.3
$\phi = 0.3$	55.9 / 64.8	51.5 / 43.2	34.2 / 21.4	2.1 / 34.4
$\phi = 0.1$	55.5 / 67.7	49.7 / 19.8	35.0 / 6.2	2.1 / 11.5

Table 6: Performance of the LLM on the natural-language version of the No-Data Algorithm with our ablation study including a model not proficient in the tasks. The success rate for this model was low in both the IP case (the model not lying; just wrong) and the OOP case (the model lying *and* wrong), although much lower in the latter.

Algorithmic Trust in LLMs-as-Judges

Takeaways:

- **Lying** models can be **caught** easily:
 - Successes are high / low when needed
- Models who do not know the task can also be caught easily
- **Algorithmic** performance is **close to known** performance

	IP Acc. / F1	OOP Acc. / F1	IP Succ. / Flips	OOP Succ. / Flips
Known				
Predicting y	76.3 / 80.3	51.5 / 67.2	—/—	—/—
Avg. per $c \in \bar{C}$	89.5 / 94.0	45.3 / 46.5	—/—	—/—
Unknown				
GPT-4.1				
$\phi = 0.7$	72.4 / 76.6	49.3 / 57.8	87.8 / 9.3	1.4 / 70.3
$\phi = 0.3$	74.4 / 78.1	49.1 / 39.1	86.8 / 4.9	1.2 / 33.6
$\phi = 0.1$	76.9 / 79.8	50.7 / 16.4	86.2 / 1.8	1.2 / 9.1
Qwen 2.5B				
$\phi = 0.7$	62.5 / 63.9	51.7 / 59.5	34.4 / 46.0	1.8 / 70.3
$\phi = 0.3$	55.9 / 64.8	51.5 / 43.2	34.2 / 21.4	2.1 / 34.4
$\phi = 0.1$	55.5 / 67.7	49.7 / 19.8	35.0 / 6.2	2.1 / 11.5

Table 6: Performance of the LLM on the natural-language version of the No-Data Algorithm with our ablation study including a model not proficient in the tasks. The success rate for this model was low in both the IP case (the model not lying; just wrong) and the OOP case (the model lying *and* wrong), although much lower in the latter.

Other Algorithmic Approaches

- (In)Formal verification
 - Chain-of-thought verification by constructing DAGs
 - These works have shortcomings!
- † Proxy Performance Prediction
 - Random-matrix theory (e.g., WeightWatcher)
 - Proxy-model prediction (e.g., ProxyLM)
- † Effective reinforcement learning
 - After all, all *PO methods are the same
 - They are also subject to the same shortcomings we have been discussing

[1] Zhao et al., [Verifying Chain-of-Thought Reasoning via Its Computational Graph](#)

[2] Wu et al., [It Takes Two: Your GRPO Is Secretly DPO](#)

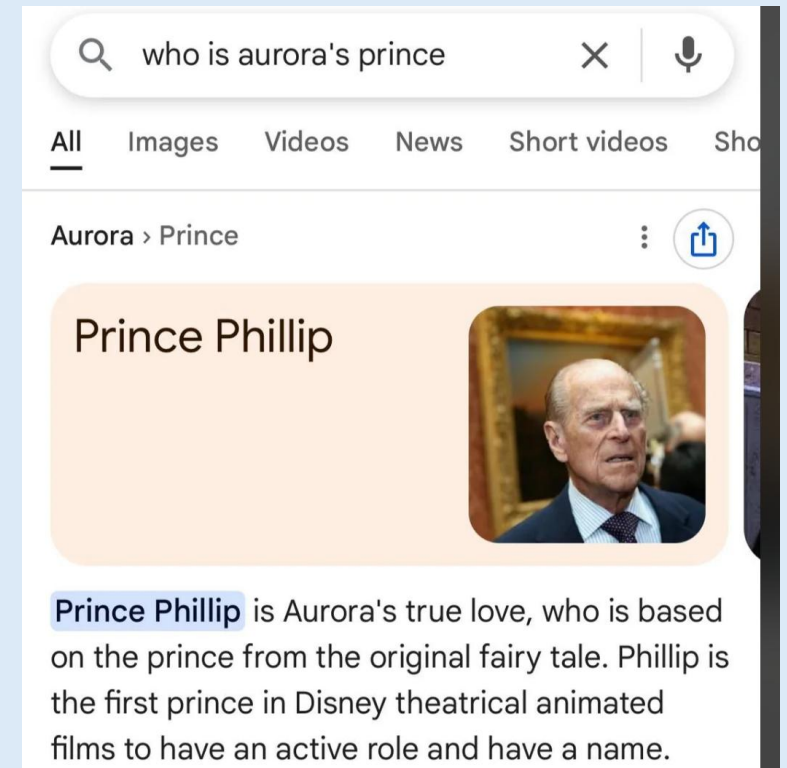
[3] Gehrmann, S. [Reward Models are Metrics in a Trench Coat](#)

Open Questions

1. Verification over cyclical graphs

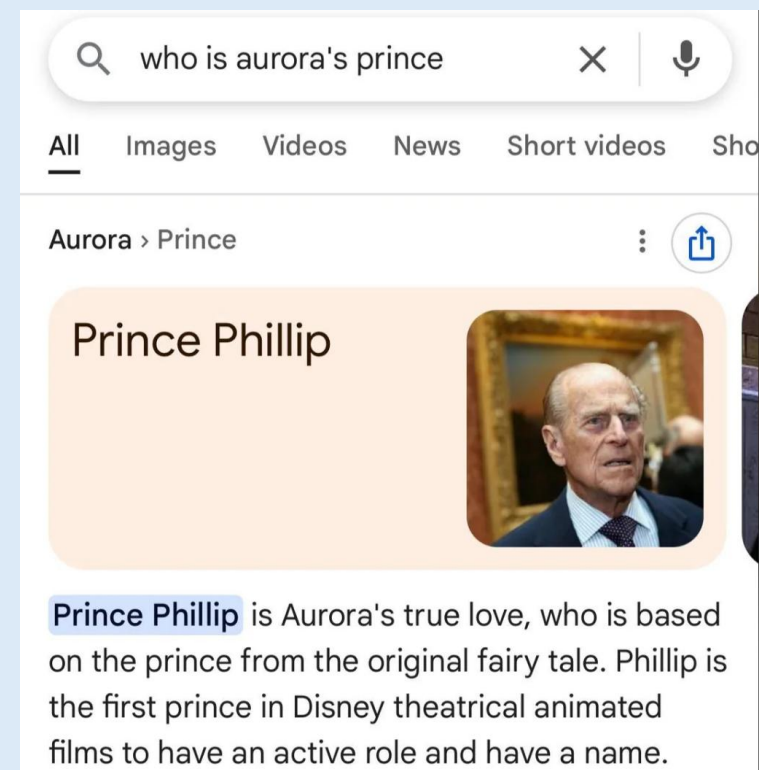
Open Questions

1. Verification over cyclical graphs (?)
2. How does one verify natural-language claims?
(Not fact-checking, but also that)



Open Questions

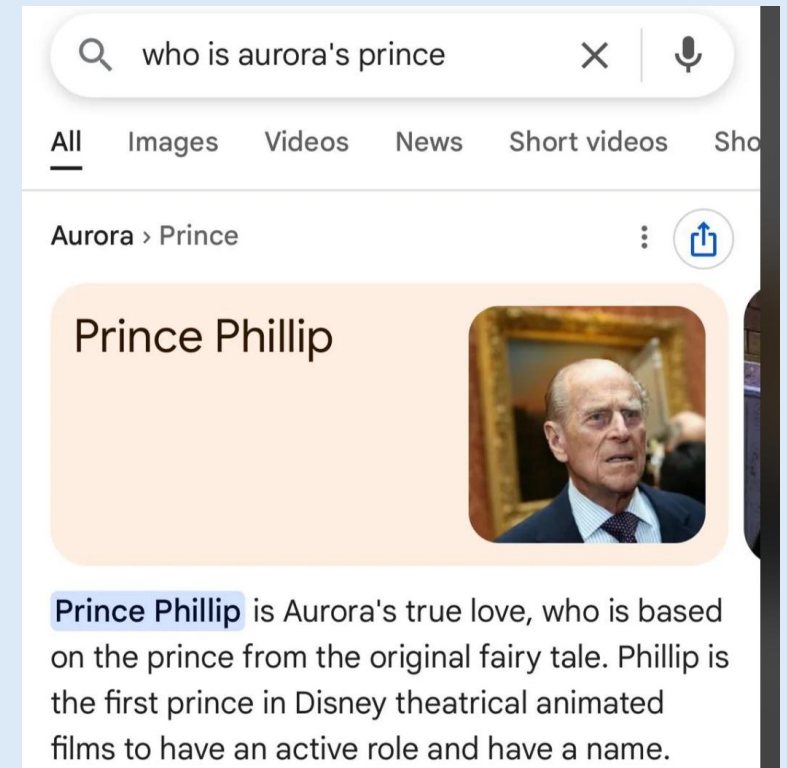
1. Verification over cyclical graphs (?)
2. How does one verify natural-language claims?
(Not fact-checking, but also that)
3. Is there a way to make PEFT methods robust to OOD?



Open Questions

1. Verification over cyclical graphs (?)
2. How does one verify natural-language claims?
(Not fact-checking, but also that)
3. Is there a way to make PEFT methods robust to OOD?
4. Related: can we marry the efficiency of ICL with the performance of ML?

Conjecture: weak learners



In Sum

1. We need **research** to better understand when a 'metric' could be used to **measure** things.
2. **Research** needs to get better
3. We need to rethink what it means to **measure** things

