

Evaluating Language Models for Mathematics Through Interactions

Katie Collins

Work Co-Led with Albert Jiang

Sept 8, 2023

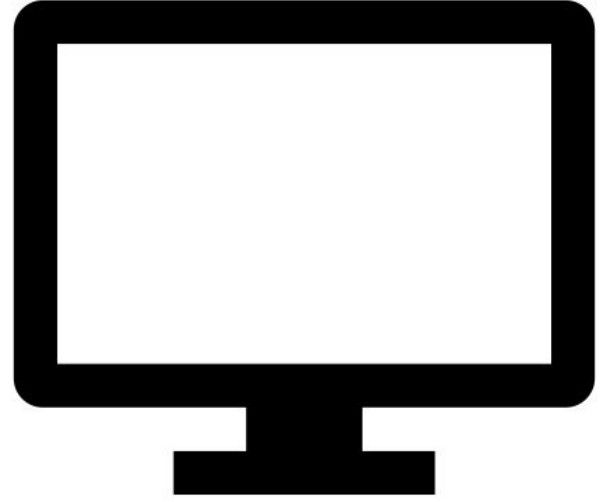
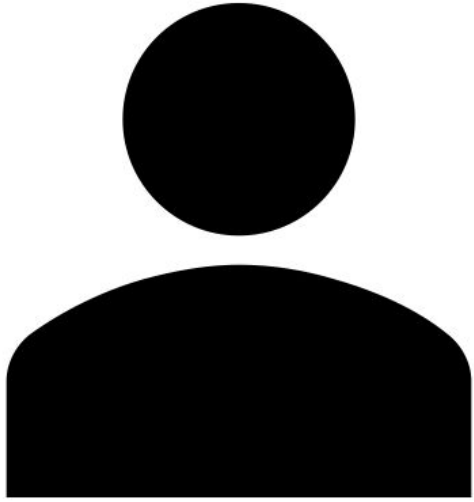
Jointly with a fantastic team!

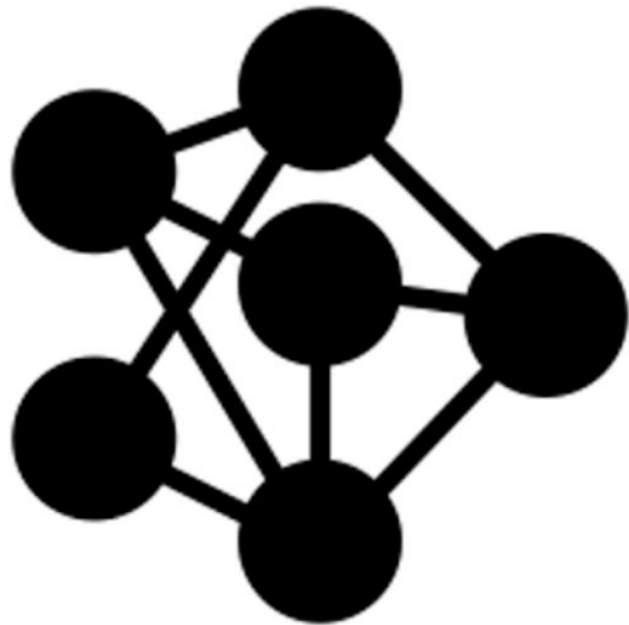
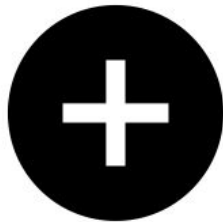
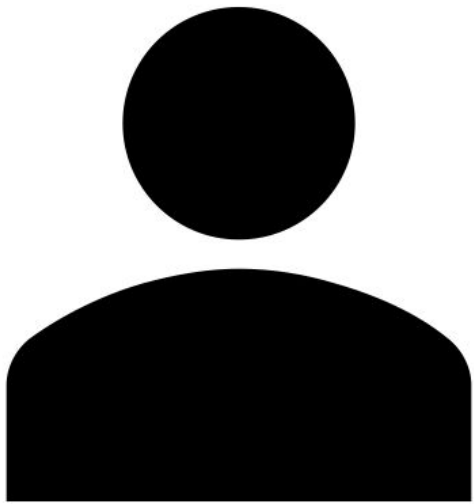
Albert Q. Jiang*, Simon Frieder, Lionel Wong,
Miri Zilka, Umang Bhatt, Thomas Lukasiewicz,
Yuhuai Wu, Joshua B. Tenenbaum, William
Hart, Timothy Gowers, Wenda Li,
Adrian Weller[‡], and Mateja Jamnik[‡]

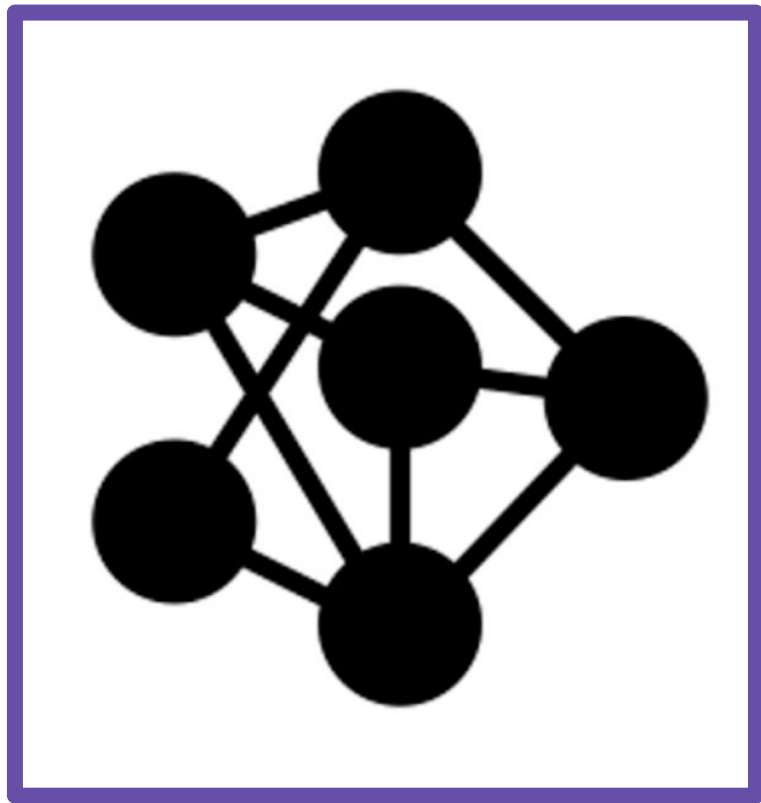
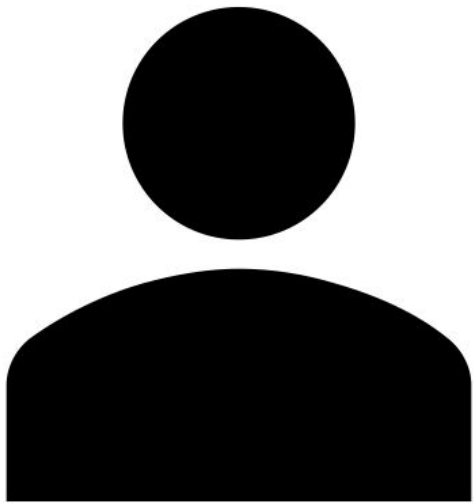
Jointly with a fantastic team!

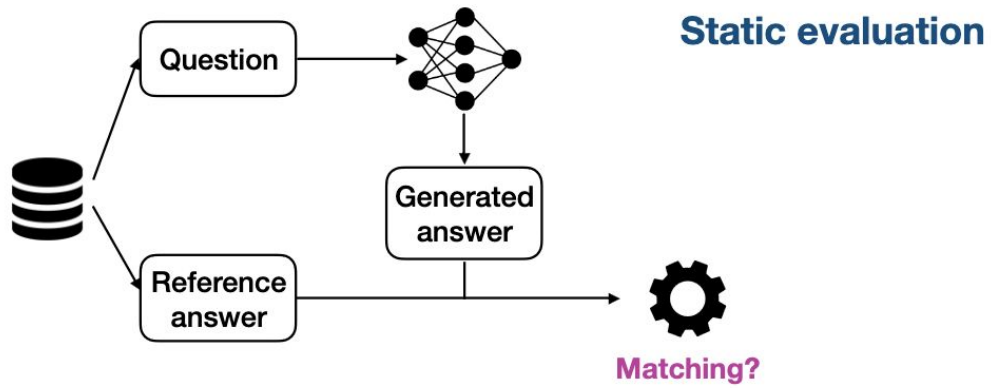
Albert Q. Jiang*, Simon Frieder, Lionel Wong,
Miri Zilka, Umang Bhatt, Thomas Lukasiewicz,
Yuhuai Wu, Joshua B. Tenenbaum, William
Hart, Timothy Gowers, Wenda Li,
Adrian Weller[‡], and Mateja Jamnik[‡]

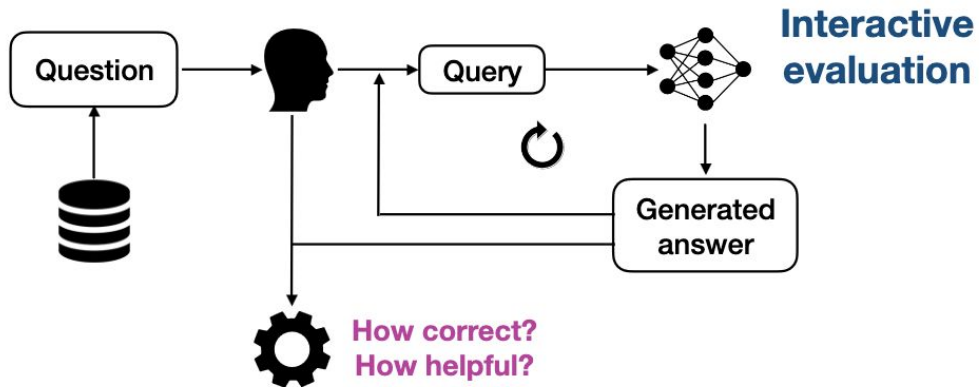
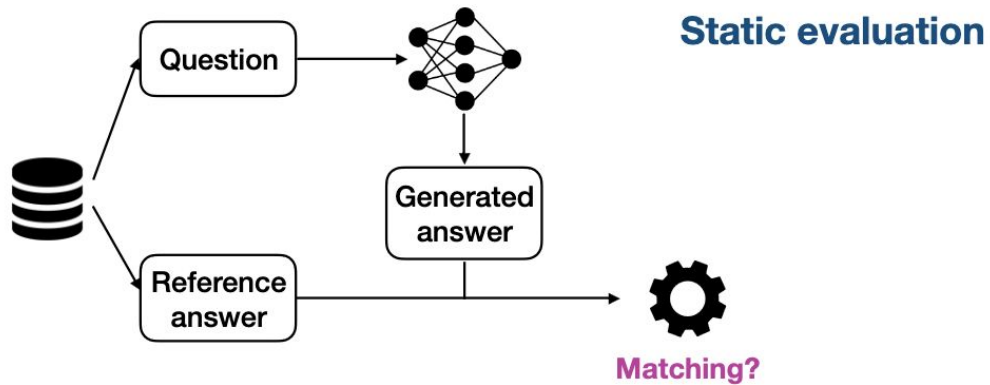












Rethink reporting of evaluation results in AI

Aggregate metrics and lack of access to results limit understanding

By **Ryan Burnell¹, Wout Schellaert², John Burden^{1,3}, Tomer D. Ullman⁴, Fernando Martinez-Plumed², Joshua B. Tenenbaum⁵, Danaja Rutar¹, Lucy G. Cheke^{1,6}, Jascha Sohl-Dickstein⁷, Melanie Mitchell⁸, Douwe Kiela⁹, Murray Shanahan^{10,11}, Ellen M. Voorhees¹², Anthony G. Cohn^{13,14,15,16}, Joel Z. Leibo¹⁰, Jose Hernandez-Orallo^{1,2,3}**

was incorrect. For other systems, the score for each instance might be based on how quickly the system completed its task, the quality of its outputs, or the total reward it obtained. Finally, performance across the various instances and tasks is usually aggregated to a small number of metrics that summarize how well the system performed, such as percentage accuracy

example, the Beyond the Imitation Benchmark (BIG-bench) for language models includes over 200 tasks that aggregate everything from language modeling to causal reasoning (4). ... across these disparate tasks—the BIG-bench leaderboard does—reduces information in the benchmark to a single all score that is hard to interpret

Rethink reporting of evaluation results in AI

Aggregate metrics and lack of access to results limit understanding

By **Ryan Burnell¹, Wout Schellaert², John Burden^{1,3}, Tomer D. Ullman⁴, Fernando Martinez-Plumed², Joshua B. Tenenbaum⁵, Danaja Rutar¹, Lucy G. Cheke^{1,6}, Jascha Sohl-Dickstein⁷, Melanie Mitchell⁸, Douwe Kiela⁹, Murray Shanahan^{10,11}, Ellen M. Voorhees¹², Anthony G. Cohn^{13,14,15,16}, Joel Z. Leibo¹⁰, Jose Hernandez-Orallo^{1,2,3}**

Mathematical Capabilities of ChatGPT

Simon Frieder^{*,1,5}, Luca Pinchetti¹, Alexis Chevalier³, Ryan-Rhys Griffiths⁴, Tommaso Salvatori^{2,7}, Thomas Lukasiewicz^{2,1}, Philipp Christian Petersen^{5,6}, and Julius Berner⁵

¹Department of Computer Science, University of Oxford, Oxford, UK

²Institute of Logic and Computation, Vienna University of Technology, Vienna, Austria

³School of Mathematics, Institute for Advanced Study, Princeton, US

⁴Department of Physics, University of Cambridge, Cambridge, UK

⁵Faculty of Mathematics, University of Vienna, Vienna, Austria

⁶Research Network Data Science, University of Vienna, Vienna, Austria

⁷VERSES Research Lab, Los Angeles, CA 90016, USA

This Talk

Part I: Systematic Interactive Eval

Part II: Expert Case Studies

Part III: Synthesized Takeaways and Looking Ahead

This Talk

Part I: Systematic Interactive Eval

Part II: Expert Case Studies

Part III: Synthesized Takeaways and Looking Ahead

CheckMate

CheckMate

- Tool for humans to interact and rate LLMs/conversation AI systems
- Permits both per-model and cross-model evaluation

CheckMate

- Tool for humans to interact and rate LLMs/conversation AI systems
- Permits both per-model and cross-model evaluation
- **Extendible**

CheckMate

Model 1 Model 2 Model 3 Final preference

As a reminder, the problem is:

Let $a, b \in \mathbb{Z}$.

Let r and s be coprime integers.

Then show that $a \equiv b \pmod{rs}$ iff $a \equiv b \pmod{r}$ and $a \equiv b \pmod{s}$, where $a \equiv b \pmod{r}$ denotes that a is congruent modulo r to b .

Note, the problem is NOT automatically provided to the model. You will need to provide it, or part of the problem, as desired. You can copy and paste from the problem above. You can optionally render your text in markdown before entering by pressing the --> button (note: the set of LaTeX symbols is restricted).

After many interactions, you may also need to SCROLL to see new model generations.

Chatbot

User: What are coprime integers in \mathbb{Z} ?

AI: Two integers a and b are said to be coprime (or relatively prime) if their greatest common divisor (GCD) is 1. In other words, there is no integer greater than 1 that divides both a and b . For example, 4 and 9 are coprime because their GCD is 1, while 6 and 9 are not coprime because their GCD is 3.

What are coprime integers in \mathbb{Z} ? --> What are coprime integers in \mathbb{Z} ?

Interact

Done with interaction

Done with interaction

User: What are coprime integers in \mathbb{Z} ?

AI: Two integers a and b are said to be coprime (or relatively prime) if their greatest common divisor (GCD) is 1. In other words, there is no integer greater than 1 that divides both a and b . For example, 4 and 9 are coprime because their GCD is 1, while 6 and 9 are not coprime because their GCD is 3.

How helpful would this AI generated response be towards helping someone solve this problem? If you already know how to solve the problem, evaluate this as if you were an undergraduate mathematics student encountering this problem for the first time.

(0) Actively harmful (1) Very harmful (2) Somewhat harmful

(3) Unlikely to help, but unlikely to hurt (4) Somewhat helpful (5) Very helpful

(6) Definitely helpful

How correct (i.e., mathematically sound) is the generation?

(0) N/A - this response does not contain any mathematical information

(1) Completely incorrect or nonsensical (2) Multiple critical maths errors

(3) At least one critical math error or multiple small errors

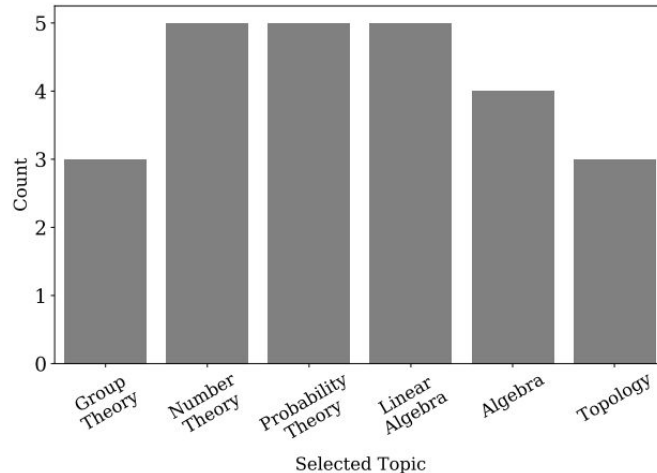
(4) One or more minor errors, but otherwise mostly correct

(5) One or two minor errors, but almost entirely correct (6) Completely correct

Finish rating

MathConverse

- 25 entries over 261 human-model interactions
 - Mixed cohort of mathematical expertise
- 3 models (InstructGPT, ChatGPT-3, GPT-4)
 - Blind to model!
- ProofWiki problems drawn from 6 topics



MathConverse

- 25 entries over 261 human-model interactions
 - Mixed cohort of mathematical expertise
- 3 models (InstructGPT, ChatGPT-3, GPT-4)
 - Blind to model!
- ProofWiki problems drawn from 6 topics

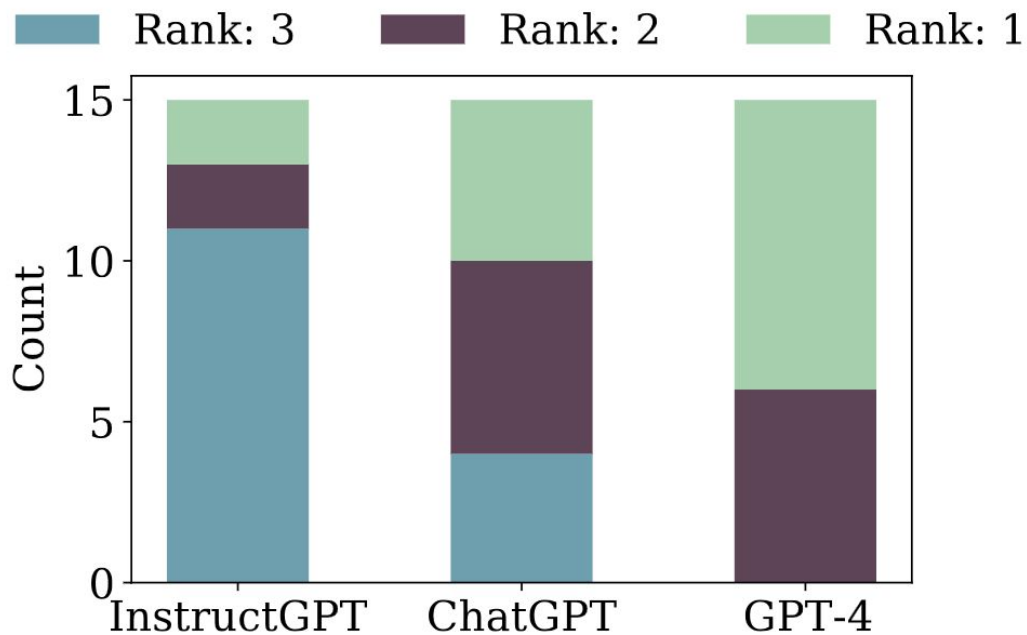
Data and Code!

<https://github.com/collinskatie/checkmate>

Insights into MathConverse

Insights into MathConverse

- Preferred mathematical assistant?

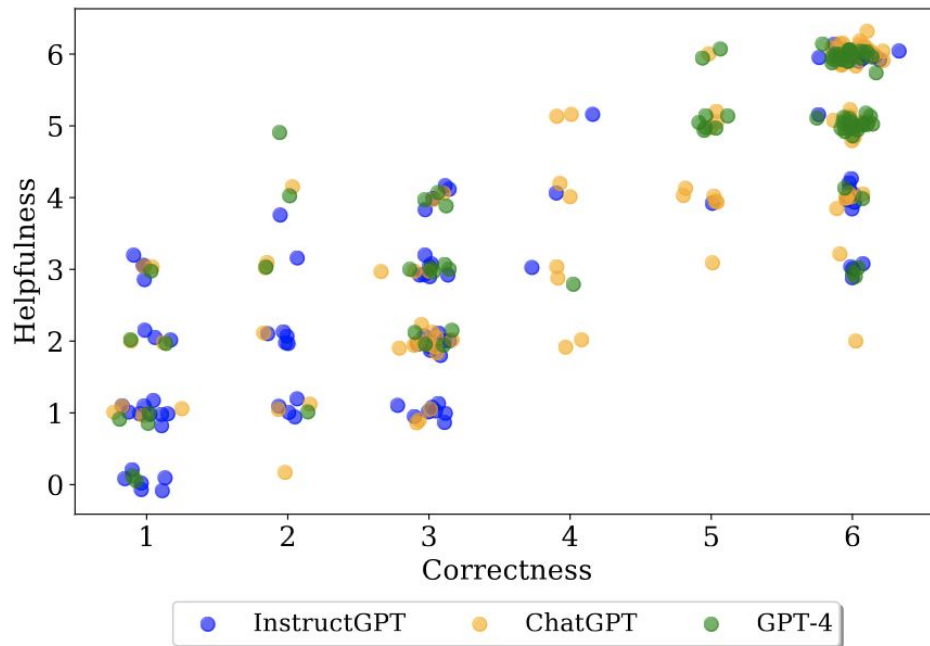


Insights into MathConverse

- Correctness and helpfulness

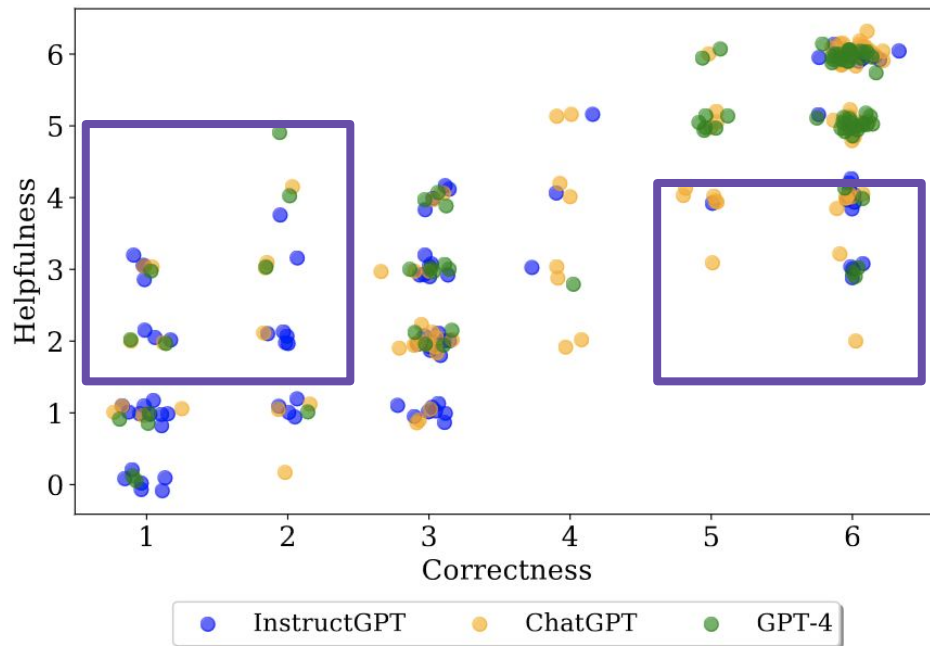
Insights into MathConverse

- Correctness and helpfulness related (pearson $r \sim 0.83$)



Insights into MathConverse

- Correctness and helpfulness related, but can diverge



Insights into MathConverse

But — is correct really correct??

Insights into MathConverse

But — is correct really correct??

- Had participants rate their self-confidence in being able to solve the problem before

Insights into MathConverse

But — is correct really correct??

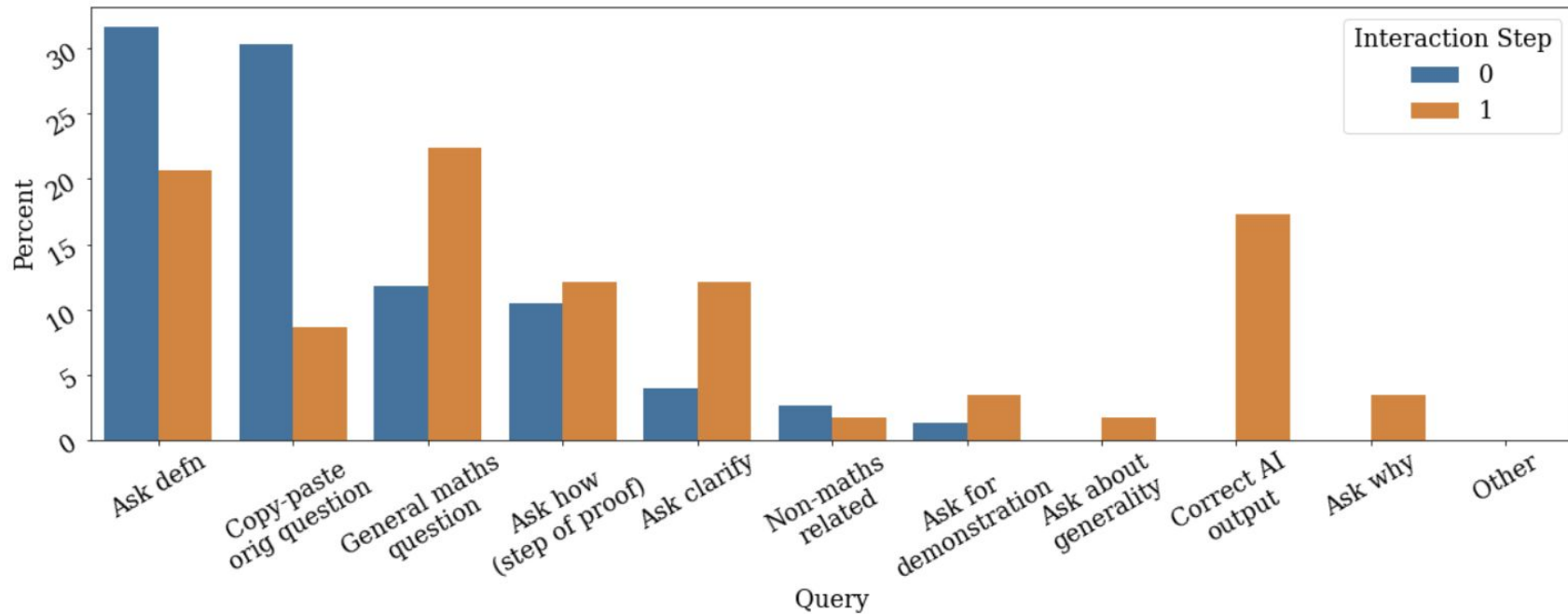
- Had participants rate their self-confidence in being able to solve the problem before
- Found instances where participants who were not confident rated responses that were incorrect as correct!



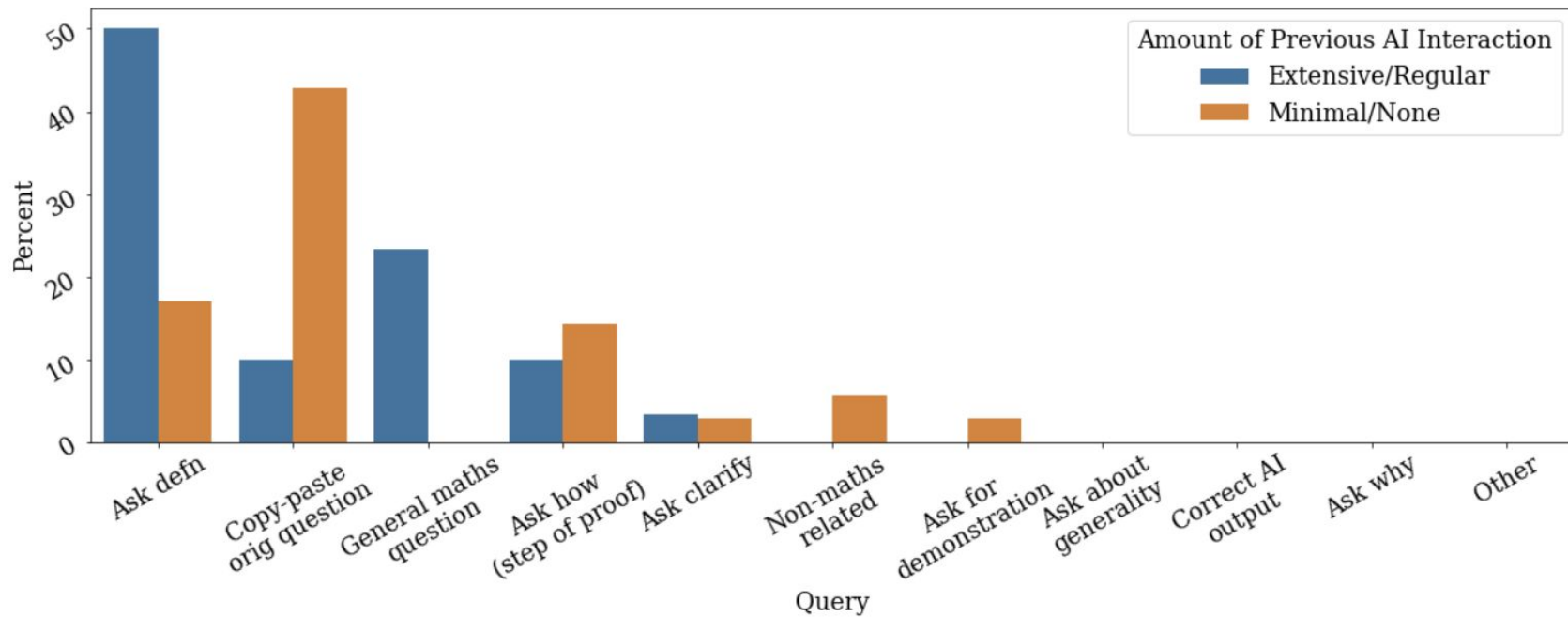
[New!] Annotated Taxonomy

- Categorize each human query
- Produce **query profiles** – signature of query types preferred by cohort subgroups, across time

[New!] Annotated Taxonomy



[New!] Annotated Taxonomy

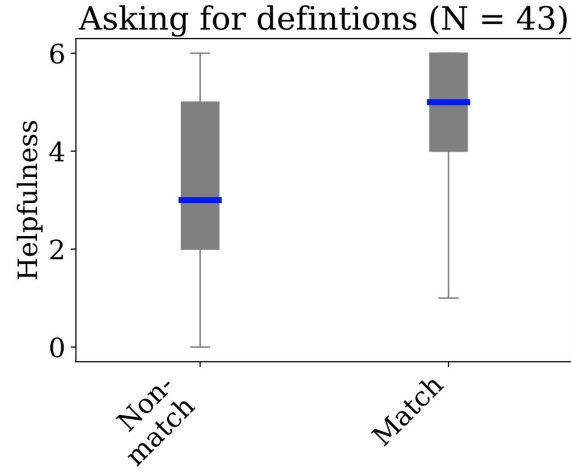
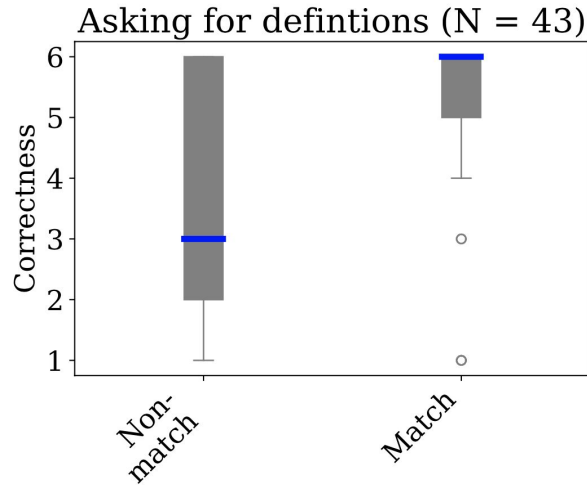


[New!] Annotated Taxonomy

- Relationship between type of query and satisfaction with model response

[New!] Annotated Taxonomy

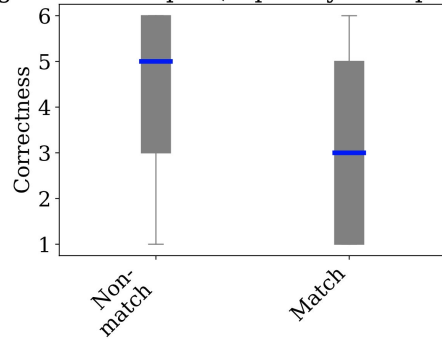
- Relationship between type of query and satisfaction with model response



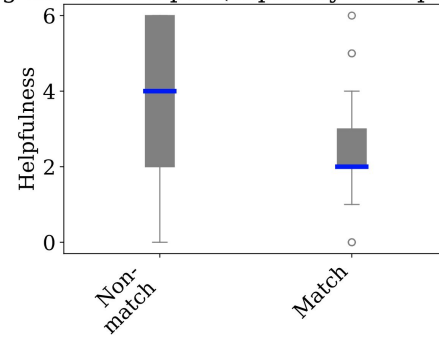
[New!] Annotated Taxonomy

- Relationship between type of query and satisfaction with model response

Correcting the AI's output (explicitly or implicitly) (N = 53)



Correcting the AI's output (explicitly or implicitly) (N = 53)



This Talk

Part I: Systematic Interactive Eval

Part II: Expert Case Studies

Part III: Synthesized Takeaways and Looking Ahead

Expert Case Studies



Bill Hart



Tim Gowers



Wenda Li

Expert Case Studies

(Some) Takeaways

- Good at producing definitions and can occasionally provide a helpful scaffold for solving a problem
- Seeming overreliance on memorised solutions
- Algebraic weakness is a real Achilles heel
- Challenges with complex planning
- Verbosity

This Talk

Part I: Systematic Interactive Eval

Part II: Expert Case Studies

Part III: Synthesized Takeaways and Looking Ahead

Synthesized Takeaways For...

ML Practitioners

**Mathematicians (Students,
Educators, Researchers)**

**LLM Evaluation, Development,
and Deployment**

Synthesized Takeaways For...

ML Practitioners

**Mathematicians (Students,
Educators, Researchers)**

**LLM Evaluation, Development,
and Deployment**

Synthesized Takeaways For...

ML Practitioners

- Enable models to communicate calibrated uncertainty and uptake user corrections
- Enable provisions of rationales
- Strive for conciseness

Mathematicians (Students, Educators, Researchers)

LLM Evaluation, Development, and Deployment

Synthesized Takeaways For...

ML Practitioners

- Enable models to communicate calibrated uncertainty and uptake user corrections
- Enable provisions of rationales
- Strive for conciseness

Mathematicians (Students, Educators, Researchers)

LLM Evaluation, Development, and Deployment

Synthesized Takeaways For...

ML Practitioners

- Enable models to communicate calibrated uncertainty and uptake user corrections
- Enable provisions of rationales
- Strive for conciseness

Mathematicians (Students, Educators, Researchers)

- Pay attention!!!!
- Take a nuanced view on when/where/how these models can help
- Be cautious when using LLMs (alone) for heavy algebra

LLM Evaluation, Development, and Deployment

Synthesized Takeaways For...

ML Practitioners

- Enable models to communicate calibrated uncertainty and uptake user corrections
- Enable provisions of rationales
- Strive for conciseness

Mathematicians (Students, Educators, Researchers)

- Pay attention!!!!
- Take a nuanced view on when/where/how these models can help
- Be cautious when using LLMs (alone) for heavy algebra

LLM Evaluation, Development, and Deployment

Synthesized Takeaways For...

ML Practitioners

- Enable models to communicate calibrated uncertainty and uptake user corrections
- Enable provisions of rationales
- Strive for conciseness

Mathematicians (Students, Educators, Researchers)

- Pay attention!!!!
- Take a nuanced view on when/where/how these models can help
- Be cautious when using LLMs (alone) for heavy algebra

LLM Evaluation, Development, and Deployment

- Carefully discern when assistance is needed (or even worth utilizing)
- Collaboration between ML practitioners + domain experts is valuable
- Incorporate interactivity into LLM capability assessments

Discussion

Big Picture: What is / will be the relationship between mathematicians and LLMs?
Across stages of mathematical development?

Discussion

Big Picture: What is / will be the relationship between mathematicians and LLMs?
Across stages of mathematical development?

(Some) Computational Questions:

- How can we build systems which are more preferable “teammates” *which express when unsure, can uptake user corrections, and provide rationales?*
- Where to integrate LLMs versus structured systems? Data-efficiency?

Discussion

Big Picture: What is / will be the relationship between mathematicians and LLMs?
Across stages of mathematical development?

(Some) Computational Questions:

- How can we build systems which are more preferable “teammates” *which express when unsure, can uptake user corrections, and provide rationales?*
- Where to integrate LLMs versus structured systems? Data-efficiency?

How do humans learn about and push the boundaries of maths?

Thanks!

Amazing Co-Authors :) Albert Q. Jiang*, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B. Tenenbaum, William Hart, Timothy Gowers, Wenda Li, Adrian Weller[‡], and Mateja Jamnik[‡]

The Human-Oriented ATP Team! <https://wtgowers.github.io/human-style-atp/team>

Especially Fabian Gloeckle, Anshula Gandhi, Jonas Bayer, Angeliki Koutsoukou-Argyraiki, Fredy Yip, Mariusz Mirek

And many more, incl:

Ced Zhang, Gabriel Poesia, Noah Goodman, Valerie Chen, Nina Grgić-Hlača, Ilija Sucholutsky, and Shoaib Ahmed Siddiqui for valuable discussions around assistive systems in AI and mathematics, and LLM evaluation

And thanks to the organizers!!!

Any Questions?

kmc61@cam.ac.uk and qj213@cam.ac.uk

Any Questions?

kmc61@cam.ac.uk and qj213@cam.ac.uk

Check out our code + data :)

<https://github.com/collinskatie/checkmate>