INDIAN INSTITUTE OF SCIENCE

भारतीय विज्ञान संस्थान

Detecting AI Slop in Research and Beyond

Danish Pruthi

Natural Language Processing

The science and engineering of building computational models to comprehend language

Natural Language Processing

The science and engineering of building computational models to comprehend language

Text Classification

"Lots of epic shows feel a little underpopulated towards the end but there's really no excuse for something as mythic, huge and mesmerizing to end as disappointingly as this."

→ Negative

Machine Translation

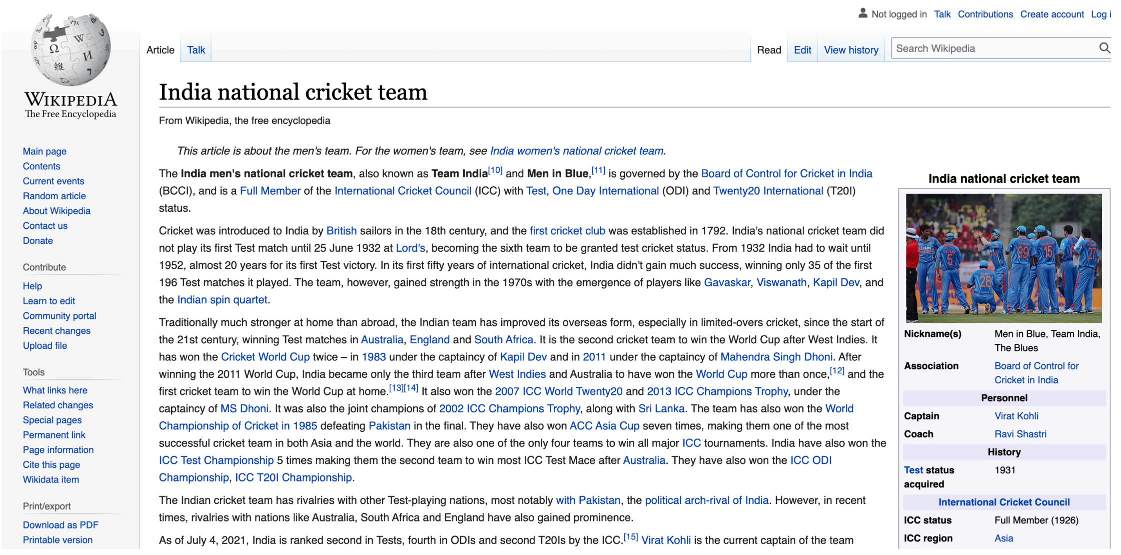
"India recorded their first Test victory, in their 24th match, against England at Madras in 1952. Later in the same year, they won their first Test series, which was against Pakistan."

→ भारत ने 1952 में मद्रास में इंग्लैंड के खिलाफ अपने 24वें मैच में अपनी पहली टेस्ट जीत दर्ज की। बाद में उसी वर्ष, उन्होंने अपनी पहली टेस्ट श्रृंखला जीती, जो पाकिस्तान के खिलाफ थी।

Question answering

"When did India win their first test match?"

+



- Main page
- Contents
- Current events
- Random article
- About Wikipedia
- Contact us
- Donate

Contribute

- Help
- Learn to edit
- Community portal
- Recent changes
- Upload file

Tools

- What links here
- Related changes
- Special pages
- Permanent link
- Page information
- Cite this page
- Wikidata item

Print/export

- Download as PDF
- Printable version

India national cricket team

From Wikipedia, the free encyclopedia

This article is about the men's team. For the women's team, see India women's national cricket team.

The **India men's national cricket team**, also known as **Team India**[10] and **Men in Blue**,[11] is governed by the Board of Control for Cricket in India (BCCI), and is a Full Member of the International Cricket Council (ICC) with Test, One Day International (ODI) and Twenty20 International (T20I) status.

Cricket was introduced to India by British sailors in the 18th century, and the first cricket club was established in 1792. India's national cricket team did not play its first Test match until 25 June 1932 at Lord's, becoming the sixth team to be granted test cricket status. From 1932 India had to wait until 1952, almost 20 years for its first Test victory. In its first fifty years of international cricket, India didn't gain much success, winning only 35 of the first 196 Test matches it played. The team, however, gained strength in the 1970s with the emergence of players like Gavaskar, Viswanath, Kapil Dev, and the Indian spin quartet.

Traditionally much stronger at home than abroad, the Indian team has improved its overseas form, especially in limited-overs cricket, since the start of the 21st century, winning Test matches in Australia, England and South Africa. It is the second cricket team to win the World Cup after West Indies. It has won the Cricket World Cup twice – in 1983 under the captaincy of Kapil Dev and in 2011 under the captaincy of Mahendra Singh Dhoni. After winning the 2011 World Cup, India became only the third team after West Indies and Australia to have won the World Cup more than once,[12] and the first cricket team to win the World Cup at home.[13][14] It also won the 2007 ICC World Twenty20 and 2013 ICC Champions Trophy, under the captaincy of MS Dhoni. It was also the joint champions of 2002 ICC Champions Trophy, along with Sri Lanka. The team has also won the World Championship of Cricket in 1985 defeating Pakistan in the final. They have also won ACC Asia Cup seven times, making them one of the most successful cricket team in both Asia and the world. They are also one of the only four teams to win all major ICC tournaments. India have also won the ICC Test Championship 5 times making them the second team to win most ICC Test Mace after Australia. They have also won the ICC ODI Championship, ICC T20I Championship.

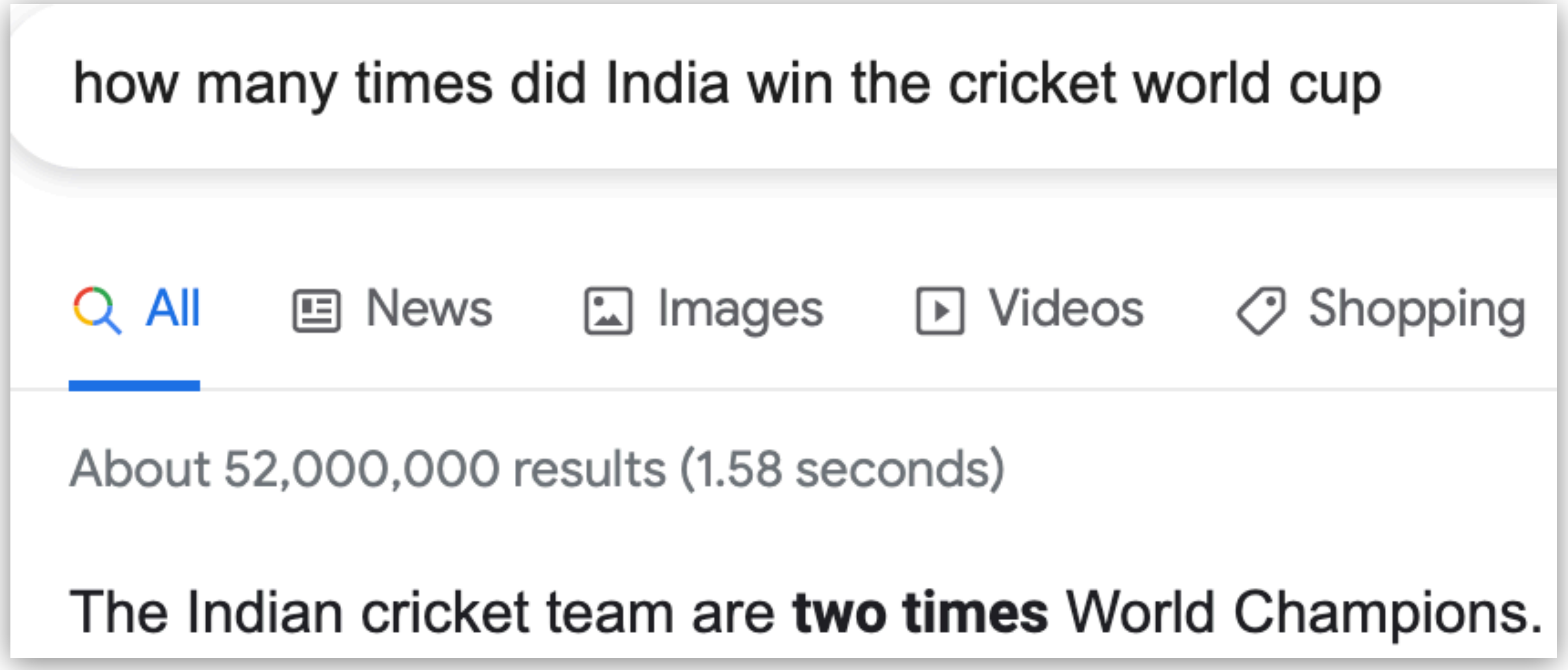
The Indian cricket team has rivalries with other Test-playing nations, most notably with Pakistan, the political arch-rival of India. However, in recent times, rivalries with nations like Australia, South Africa and England have also gained prominence.

As of July 4, 2021, India is ranked second in Tests, fourth in ODIs and second T20Is by the ICC.[15] Virat Kohli is the current captain of the team

India national cricket team	
Nickname(s)	Men in Blue, Team India, The Blues
Association	Board of Control for Cricket in India
Personnel	
Captain	Virat Kohli
Coach	Ravi Shastri
History	
Test status acquired	1931
International Cricket Council	
ICC status	Full Member (1926)
ICC region	Asia

→ 1952

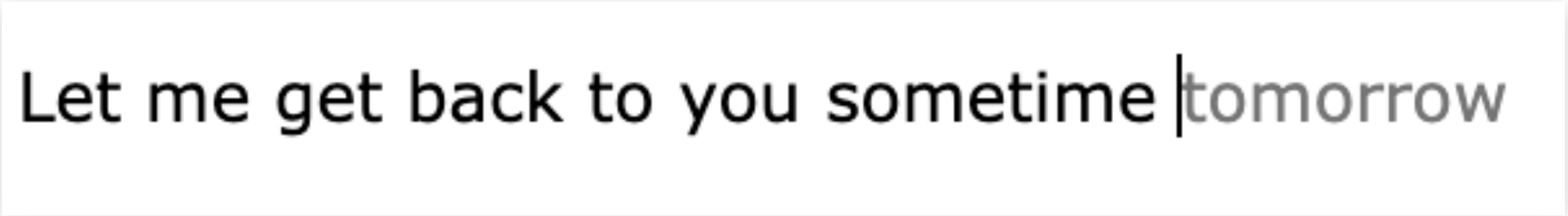
Impact of Language Technologies



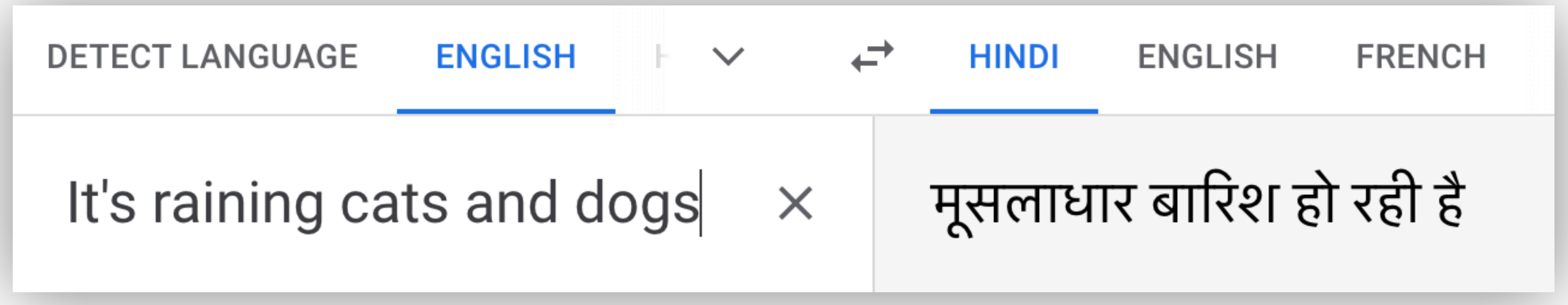
how many times did India win the cricket world cup

About 52,000,000 results (1.58 seconds)

The Indian cricket team are **two times** World Champions.

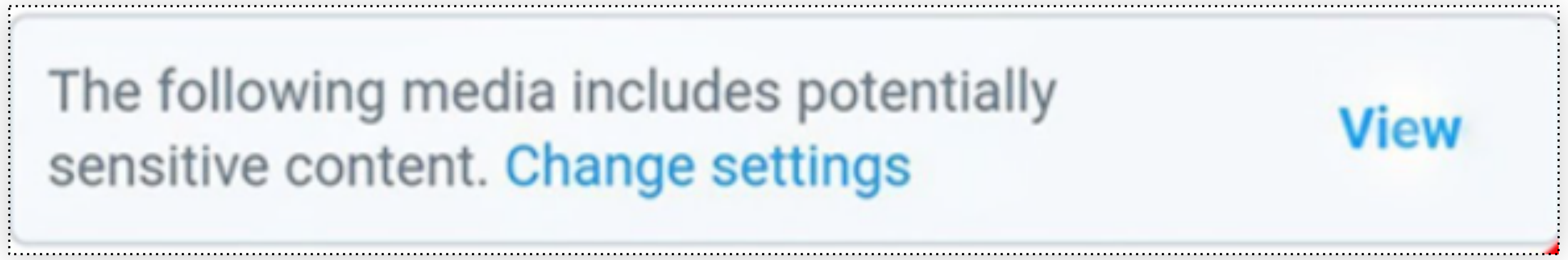


Let me get back to you sometime tomorrow



It's raining cats and dogs

मूसलाधार बारिश हो रही है



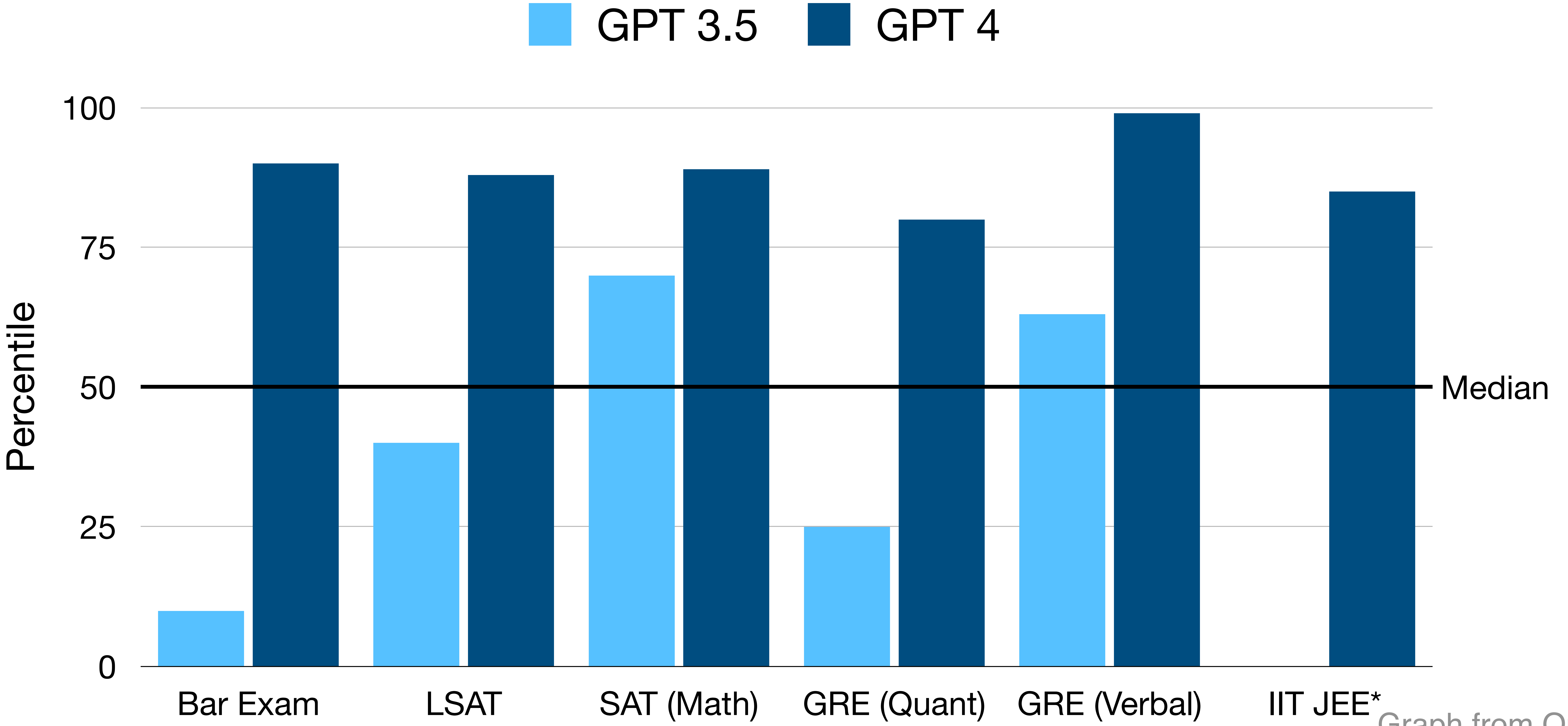
The following media includes potentially sensitive content. Change settings

Language Models

Models that assign probabilities to a sequence of words

- I am sorry for the inconvenience ___
- Let me get back to you sometime ___
- I work at IISc, I live in ___
- The Prime Minister of India is ___

Recent Performance Trends



Graph from OpenAI

* Arora et al. 2024

Promise of LLMs

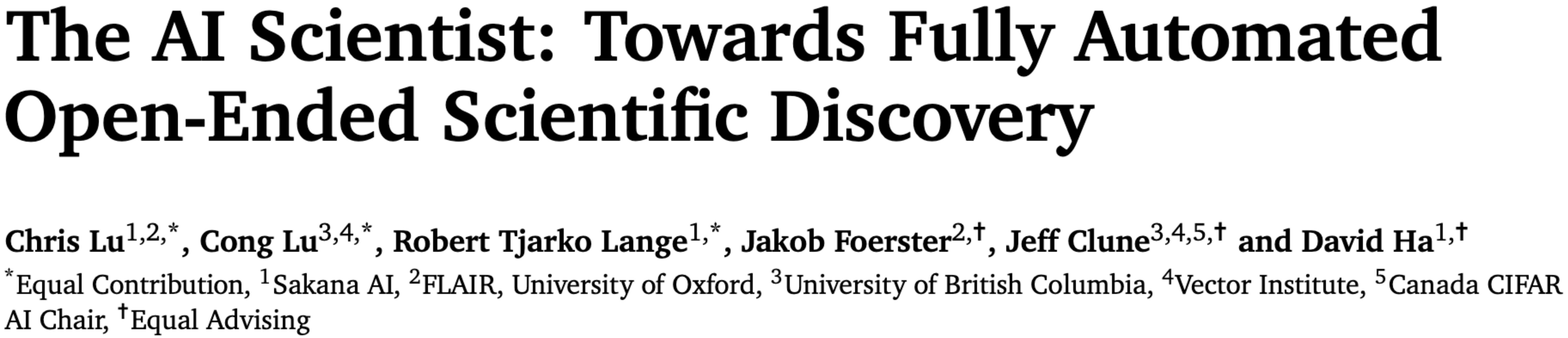
- In advancing science (e.g., Alphafold)

Promise of LLMs

- In advancing science (e.g., Alphafold)
- In automating scientific research?

Promise of LLMs

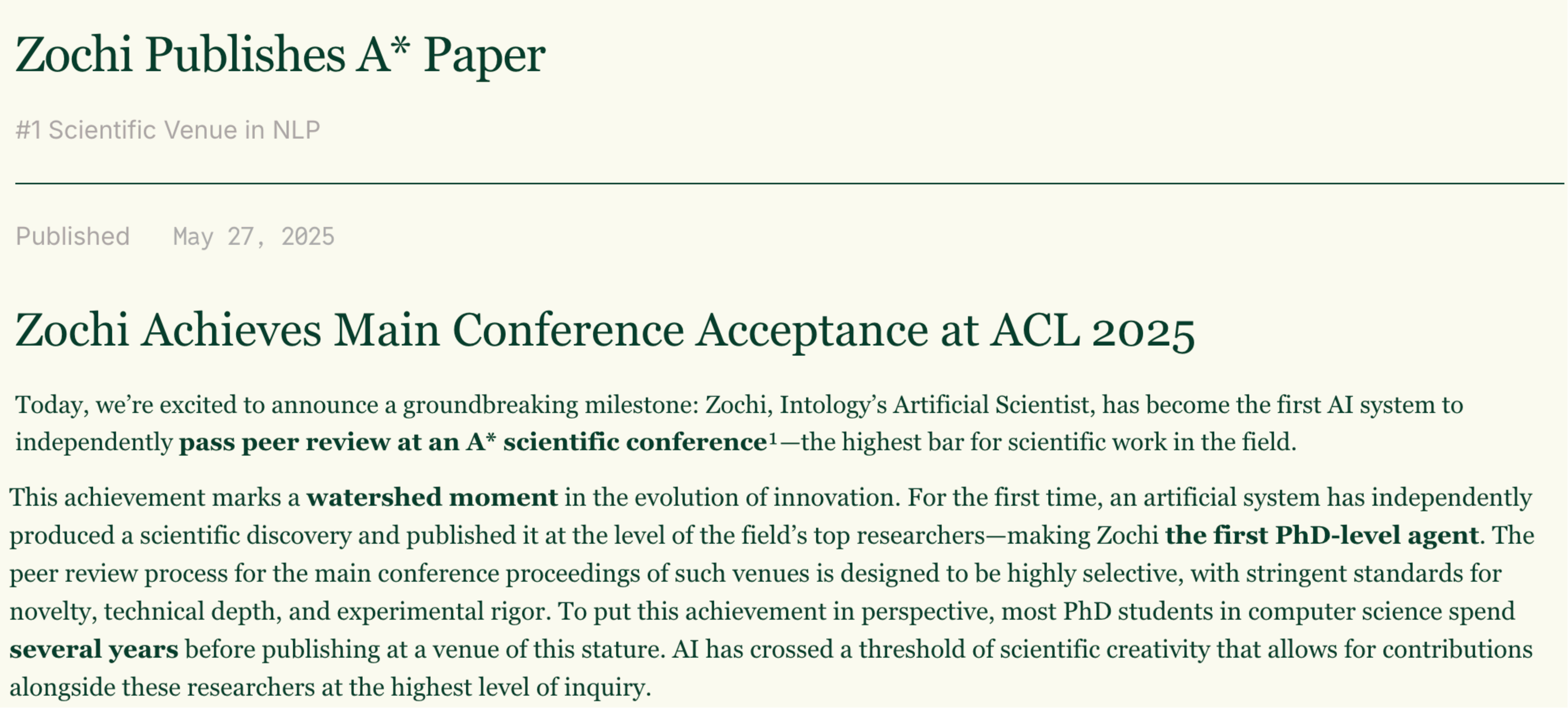
- In advancing science (e.g., Alphafold)
- In automating scientific research?



The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery

Chris Lu[1,2,*], **Cong Lu**[3,4,*], **Robert Tjarko Lange**[1,*], **Jakob Foerster**[2,†], **Jeff Clune**[3,4,5,†] **and David Ha**[1,†]

[*]Equal Contribution, [1]Sakana AI, [2]FLAIR, University of Oxford, [3]University of British Columbia, [4]Vector Institute, [5]Canada CIFAR AI Chair, [†]Equal Advising



Zochi Publishes A* Paper

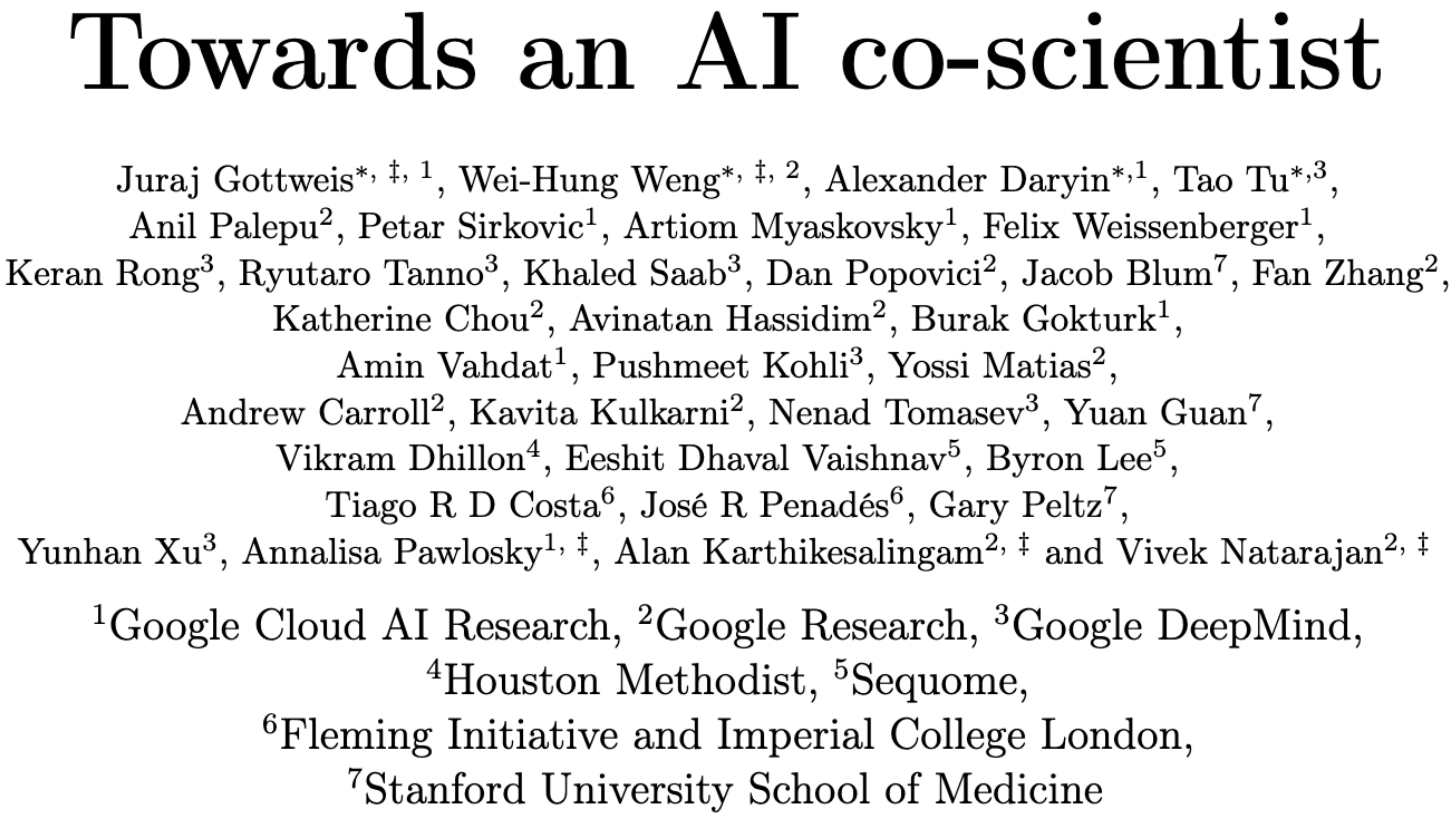
#1 Scientific Venue in NLP

Published May 27, 2025

Zochi Achieves Main Conference Acceptance at ACL 2025

Today, we're excited to announce a groundbreaking milestone: Zochi, Intology's Artificial Scientist, has become the first AI system to independently **pass peer review at an A* scientific conference**[1]—the highest bar for scientific work in the field.

This achievement marks a **watershed moment** in the evolution of innovation. For the first time, an artificial system has independently produced a scientific discovery and published it at the level of the field's top researchers—making Zochi **the first PhD-level agent**. The peer review process for the main conference proceedings of such venues is designed to be highly selective, with stringent standards for novelty, technical depth, and experimental rigor. To put this achievement in perspective, most PhD students in computer science spend **several years** before publishing at a venue of this stature. AI has crossed a threshold of scientific creativity that allows for contributions alongside these researchers at the highest level of inquiry.



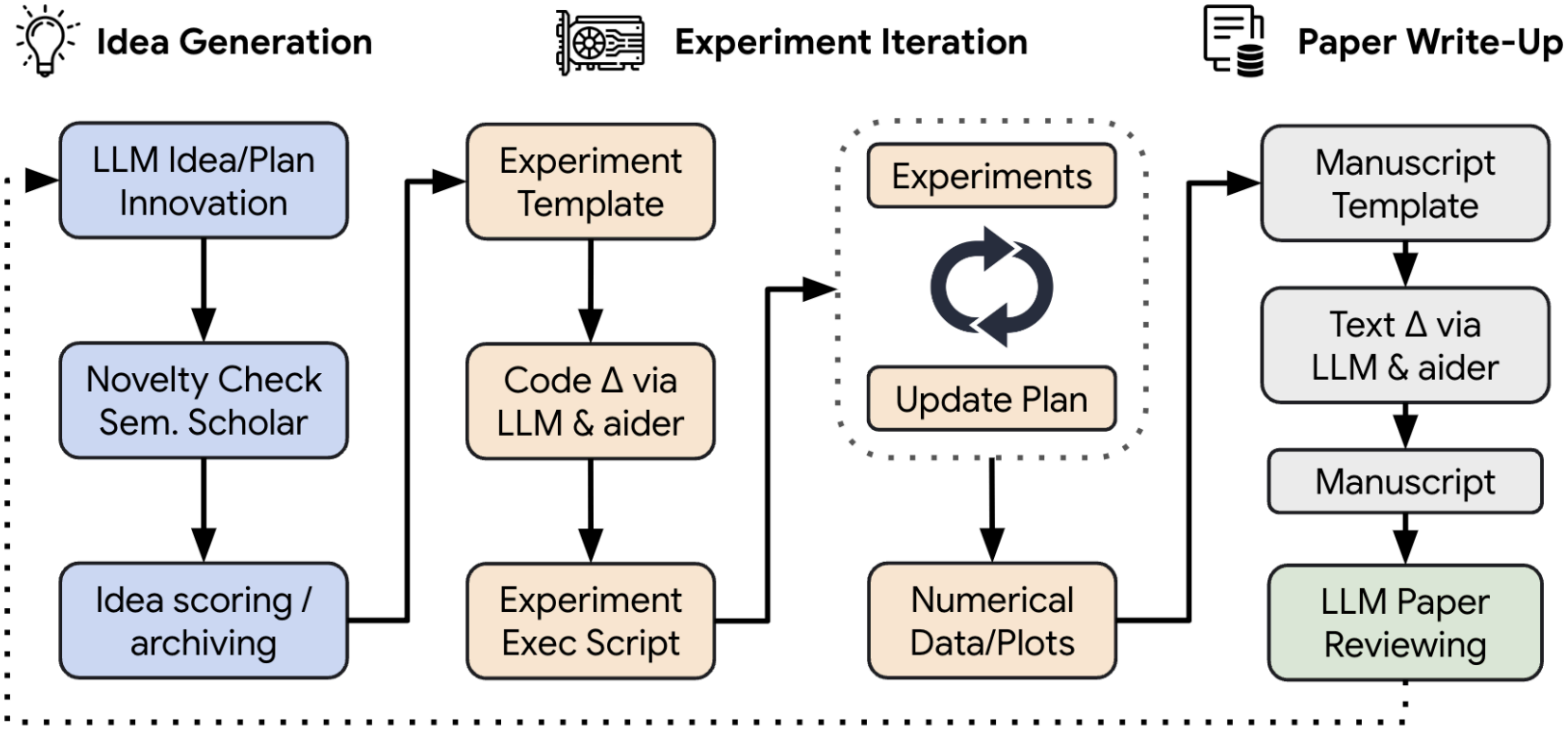
Towards an AI co-scientist

Juraj Gottweis[*, ‡, 1], Wei-Hung Weng[*, ‡, 2], Alexander Daryin[*,1], Tao Tu[*,3],
Anil Palepu[2], Petar Sirkovic[1], Artiom Myaskovsky[1], Felix Weissenberger[1],
Keran Rong[3], Ryutaro Tanno[3], Khaled Saab[3], Dan Popovici[2], Jacob Blum[7], Fan Zhang[2],
Katherine Chou[2], Avinatan Hassidim[2], Burak Gokturk[1],
Amin Vahdat[1], Pushmeet Kohli[3], Yossi Matias[2],
Andrew Carroll[2], Kavita Kulkarni[2], Nenad Tomasev[3], Yuan Guan[7],
Vikram Dhillon[4], Eeshit Dhaval Vaishnav[5], Byron Lee[5],
Tiago R D Costa[6], José R Penadés[6], Gary Peltz[7],
Yunhan Xu[3], Annalisa Pawlosky[1, ‡], Alan Karthikesalingam[2, ‡] and Vivek Natarajan[2, ‡]

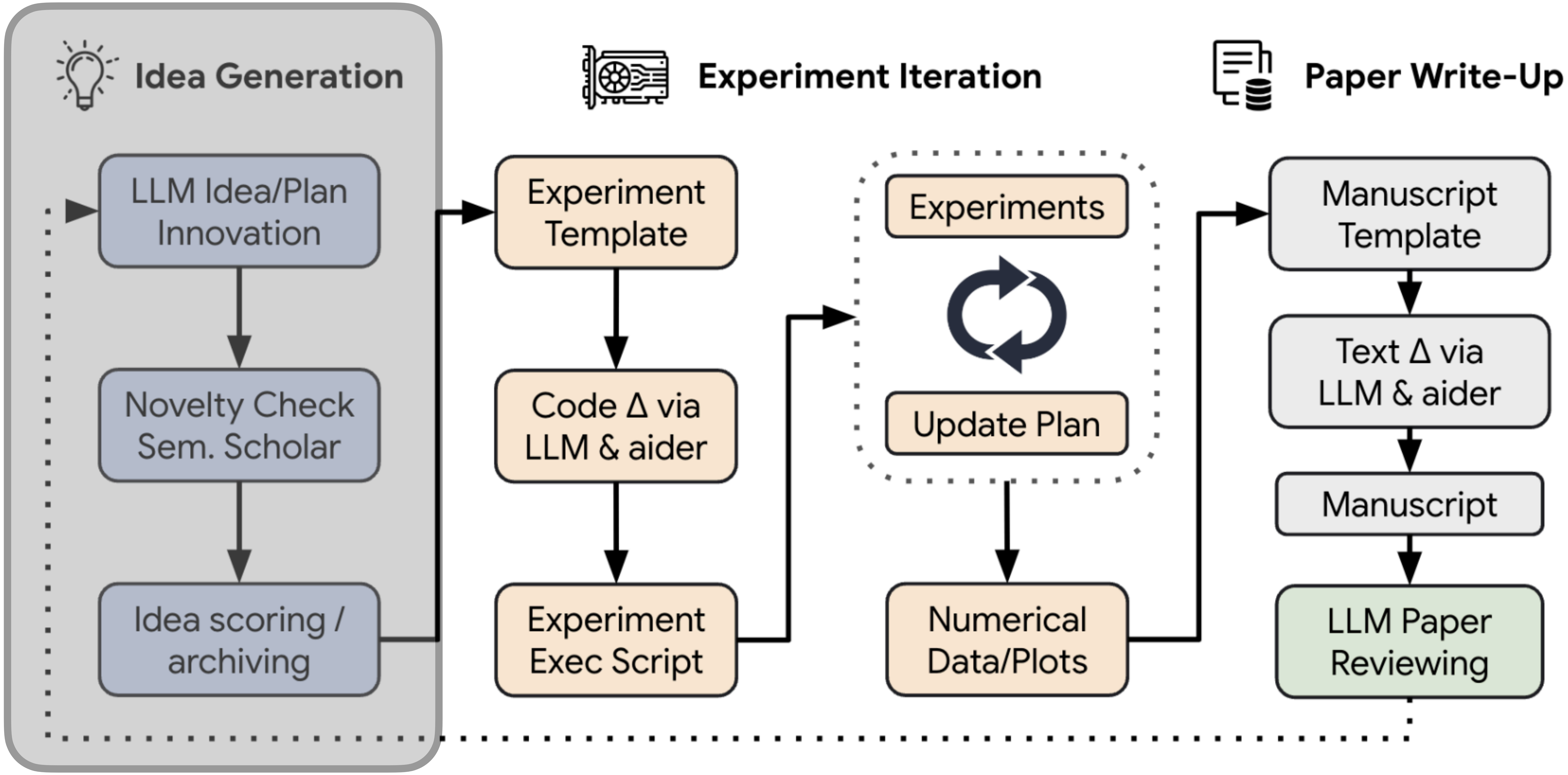
[1]Google Cloud AI Research, [2]Google Research, [3]Google DeepMind,
[4]Houston Methodist, [5]Sequome,
[6]Fleming Initiative and Imperial College London,
[7]Stanford University School of Medicine

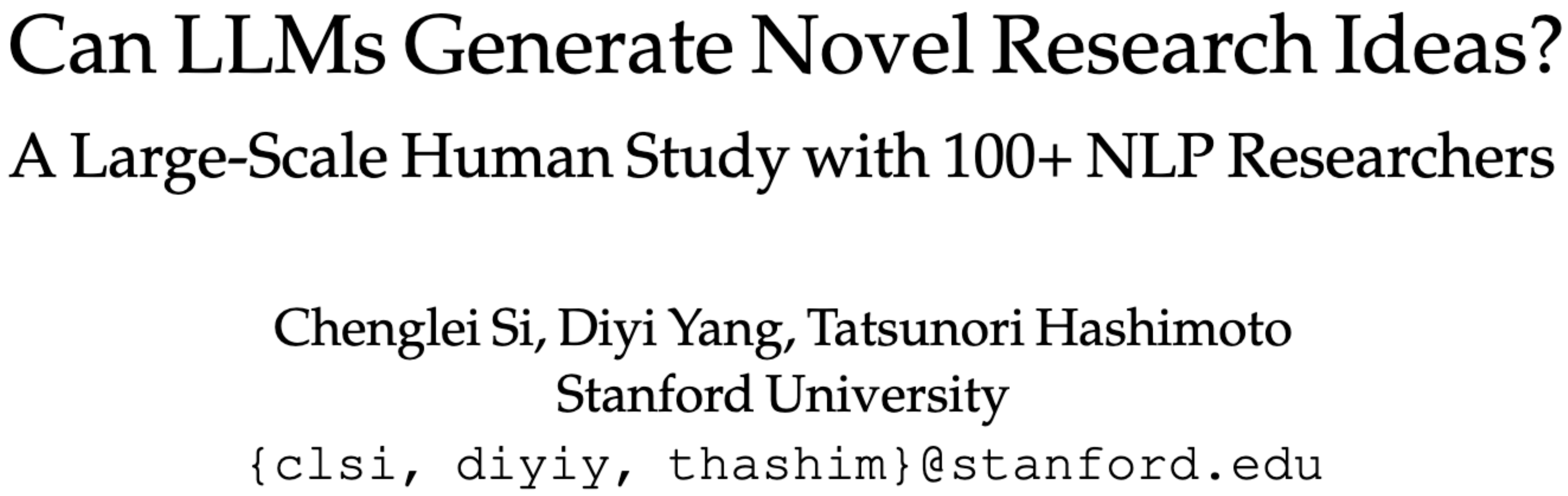
https://arxiv.org/pdf/2502.18864

Conceptual Illustration of the "AI Scientist"



Conceptual Illustration of the "AI Scientist"





Can LLMs Generate Novel Research Ideas?

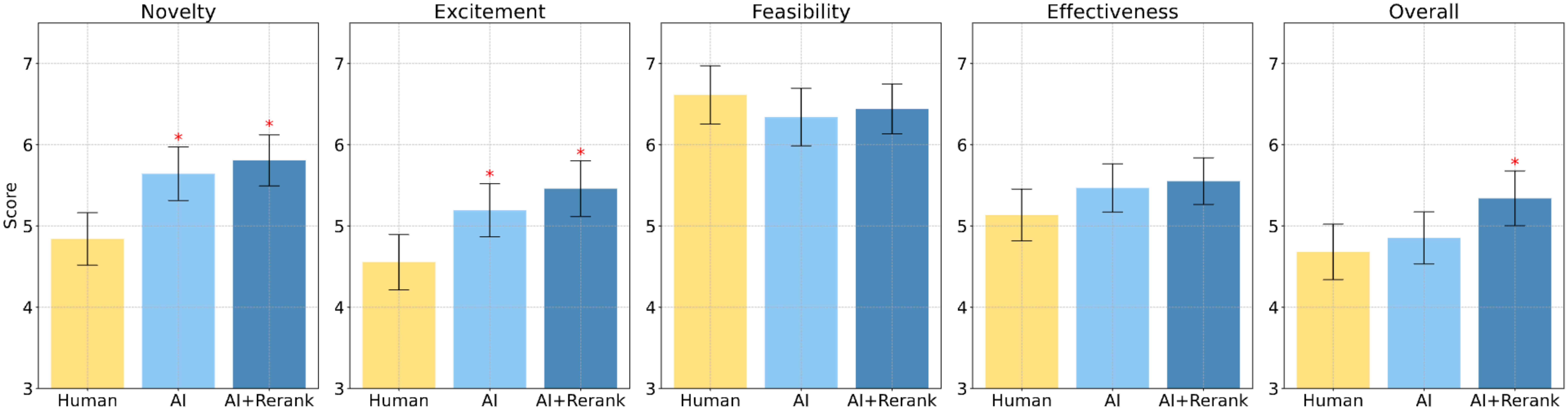
A Large-Scale Human Study with 100+ NLP Researchers

Chenglei Si, Diyi Yang, Tatsunori Hashimoto
Stanford University
{clsi, diyiy, thashim}@stanford.edu

Can LLMs Generate Novel Research Ideas?

A Large-Scale Human Study with 100+ NLP Researchers

Chenglei Si, Diyi Yang, Tatsunori Hashimoto
Stanford University
{clsi, diyiy, thashim}@stanford.edu



Evaluation Philosophy

- **Prior work: Experts assess shuffled LLM/human documents for novelty, feasibility, interestingness, etc.**

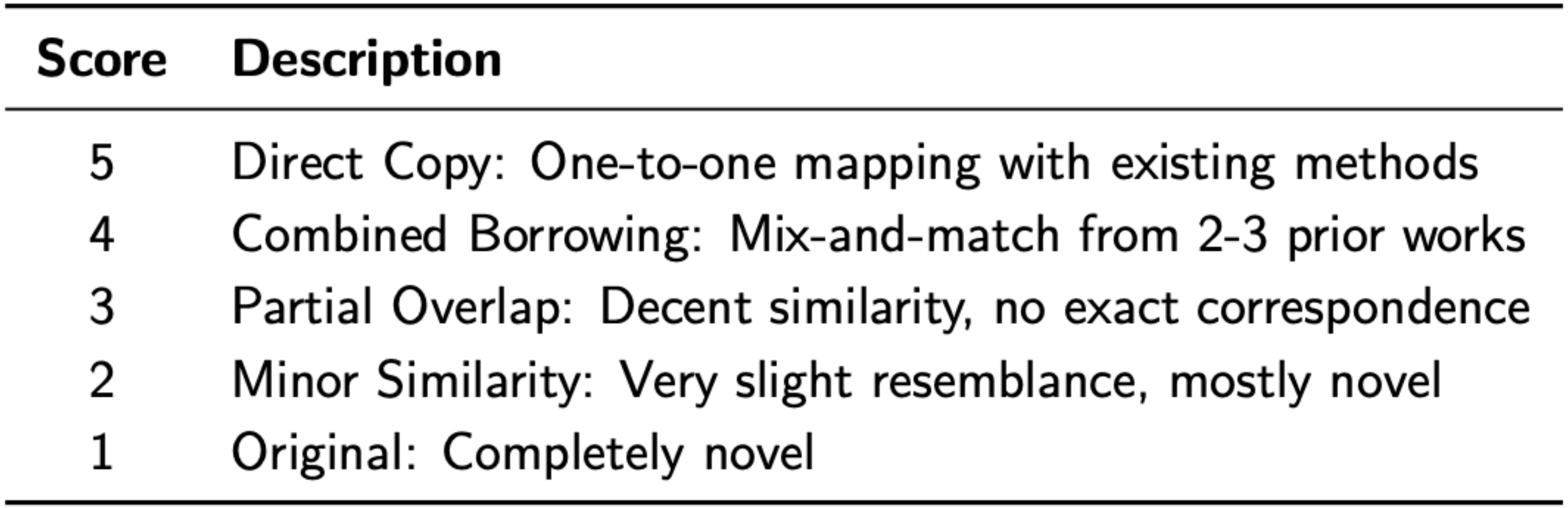
- **Our work: Experts actively search for plagiarism**
  - Different situational logic (Popper, 2013)
  - Presume plagiarism
  - Actively search for overlap in methodology in existing work

Dataset

- **50 LLM-generated research documents**
  - 36 fresh proposals generated from Si et al. (2024)
  - 4 exemplar proposals from Si et al. (2024)
  - 10 exemplar papers from "The AI Scientist" paper (Lu et al. (2024))
- **12 NLP research topics**
  - Long context capabilities, abstention techniques, bias evaluation
  - Hallucination reduction, interpretability, speech processing
  - Formal proof generation, human evaluation, machine translation
  - Scaling laws, inference optimization, persona development

Expert-led Evaluation

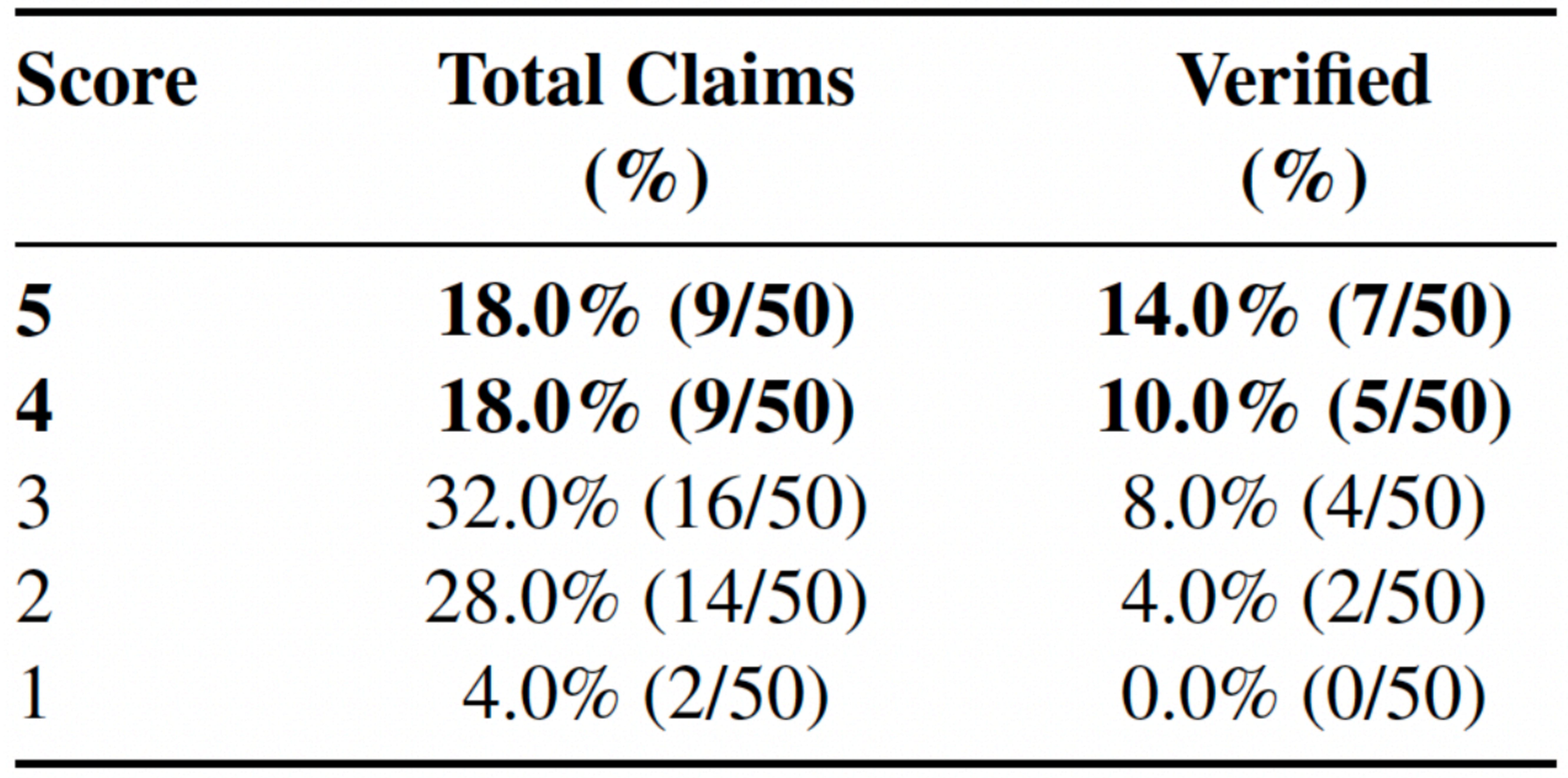
- **13 experts from 5 universities, 2 industrial labs**



Score	Description
5	Direct Copy: One-to-one mapping with existing methods
4	Combined Borrowing: Mix-and-match from 2-3 prior works
3	Partial Overlap: Decent similarity, no exact correspondence
2	Minor Similarity: Very slight resemblance, mostly novel
1	Original: Completely novel

Key result

- **Large number of proposals (> 24%) are plagiarized**



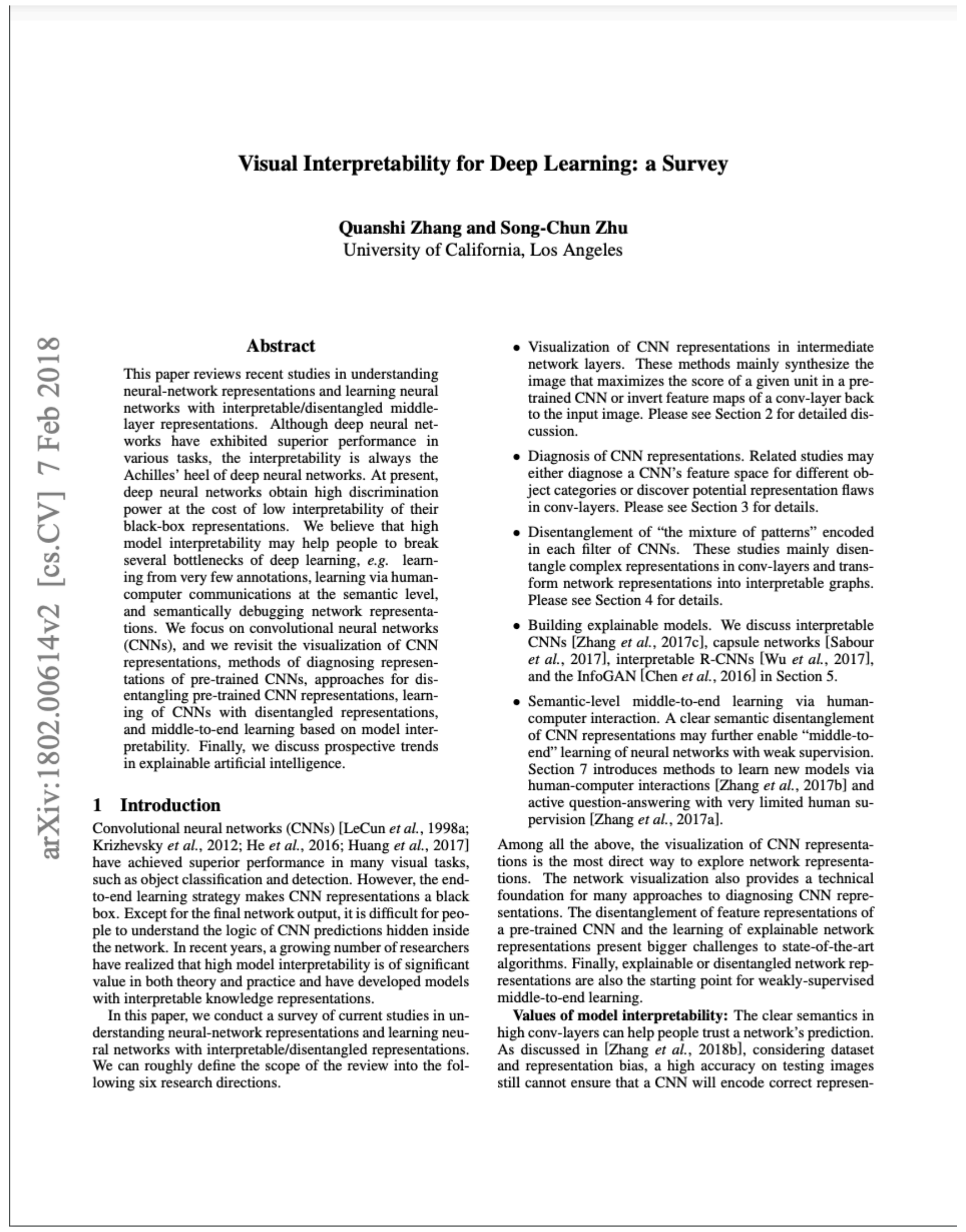
Score	Total Claims (%)	Verified (%)
5	**18.0% (9/50)**	**14.0% (7/50)**
4	**18.0% (9/50)**	**10.0% (5/50)**
3	32.0% (16/50)	8.0% (4/50)
2	28.0% (14/50)	4.0% (2/50)
1	4.0% (2/50)	0.0% (0/50)

The nature of plagiarism is sophisticated

- **Models learn to disguise existing work as novel**
- **Re-invent terminologies**
  - "Resonance graph" instead of "weighted adjacency matrix"
- **Several other case studies in the paper**

Can plagiarism be automatically detected?

- We create a synthetic dataset of plagiarized ideas



Visual Interpretability for Deep Learning: a Survey

Quanshi Zhang and Song-Chun Zhu
University of California, Los Angeles

Abstract

This paper reviews recent studies in understanding neural-network representations and learning neural networks with interpretable/disentangled middle-layer representations. Although deep neural networks have exhibited superior performance in various tasks, the interpretability is always the Achilles' heel of deep neural networks. At present, deep neural networks obtain high discrimination power at the cost of low interpretability of their black-box representations. We believe that high model interpretability may help people to break several bottlenecks of deep learning, *e.g.* learning from very few annotations, learning via human-computer communications at the semantic level, and semantically debugging network representations. We focus on convolutional neural networks (CNNs), and we revisit the visualization of CNN representations, methods of diagnosing representations of pre-trained CNNs, approaches for disentangling pre-trained CNN representations, learning of CNNs with disentangled representations, and middle-to-end learning based on model interpretability. Finally, we discuss prospective trends in explainable artificial intelligence.

1 Introduction

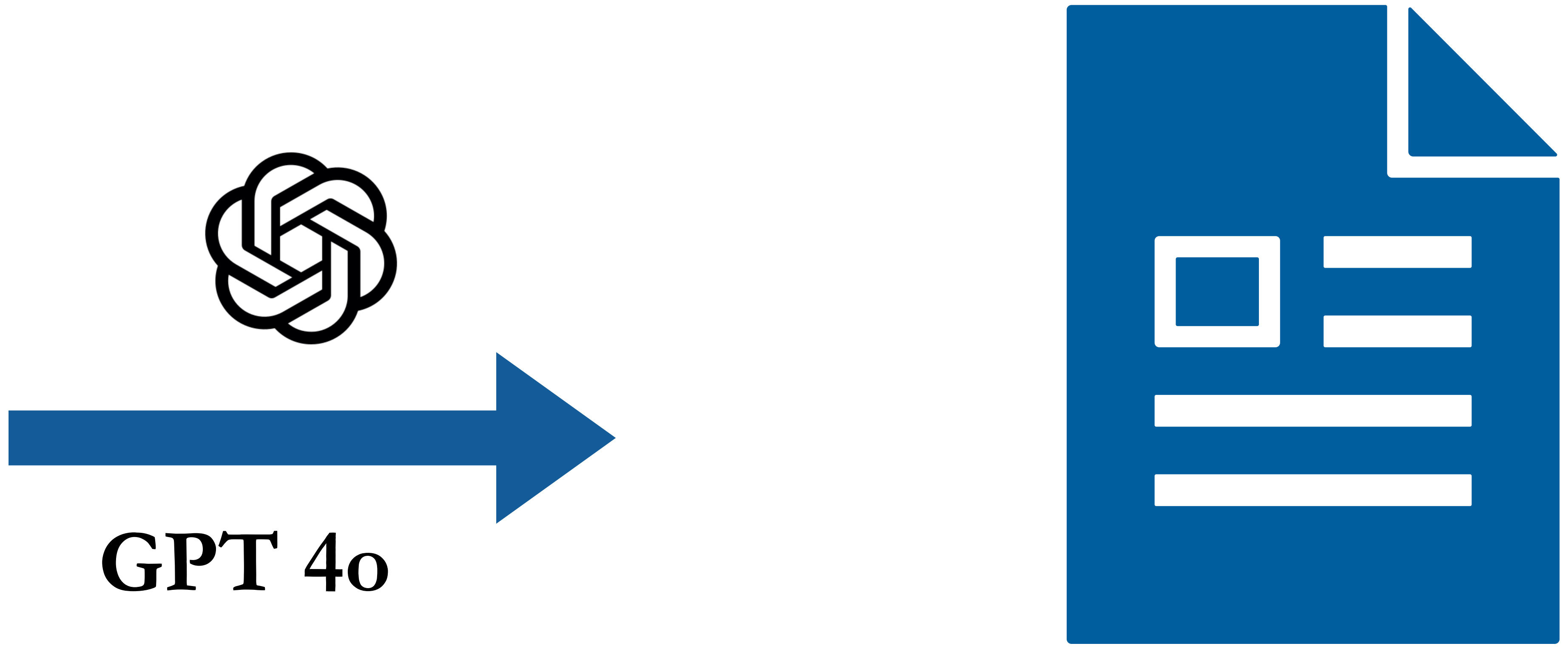
Convolutional neural networks (CNNs) [LeCun *et al.*, 1998a; Krizhevsky *et al.*, 2012; He *et al.*, 2016; Huang *et al.*, 2017] have achieved superior performance in many visual tasks, such as object classification and detection. However, the end-to-end learning strategy makes CNN representations a black box. Except for the final network output, it is difficult for people to understand the logic of CNN predictions hidden inside the network. In recent years, a growing number of researchers have realized that high model interpretability is of significant value in both theory and practice and have developed models with interpretable knowledge representations.

In this paper, we conduct a survey of current studies in understanding neural-network representations and learning neural networks with interpretable/disentangled representations. We can roughly define the scope of the review into the following six research directions.

- Visualization of CNN representations in intermediate network layers. These methods mainly synthesize the image that maximizes the score of a given unit in a pre-trained CNN or invert feature maps of a conv-layer back to the input image. Please see Section 2 for detailed discussion.
- Diagnosis of CNN representations. Related studies may either diagnose a CNN's feature space for different object categories or discover potential representation flaws in conv-layers. Please see Section 3 for details.
- Disentanglement of "the mixture of patterns" encoded in each filter of CNNs. These studies mainly disentangle complex representations in conv-layers and transform network representations into interpretable graphs. Please see Section 4 for details.
- Building explainable models. We discuss interpretable CNNs [Zhang *et al.*, 2017c], capsule networks [Sabour *et al.*, 2017], interpretable R-CNNs [Wu *et al.*, 2017], and the InfoGAN [Chen *et al.*, 2016] in Section 5.
- Semantic-level middle-to-end learning via human-computer interaction. A clear semantic disentanglement of CNN representations may further enable "middle-to-end" learning of neural networks with weak supervision. Section 7 introduces methods to learn new models via human-computer interactions [Zhang *et al.*, 2017b] and active question-answering with very limited human supervision [Zhang *et al.*, 2017a].

Among all the above, the visualization of CNN representations is the most direct way to explore network representations. The network visualization also provides a technical foundation for many approaches to diagnosing CNN representations. The disentanglement of feature representations of a pre-trained CNN and the learning of explainable network representations present bigger challenges to state-of-the-art algorithms. Finally, explainable or disentangled network representations are also the starting point for weakly-supervised middle-to-end learning.

Values of model interpretability: The clear semantics in high conv-layers can help people trust a network's prediction. As discussed in [Zhang *et al.*, 2018b], considering dataset and representation bias, a high accuracy on testing images still cannot ensure that a CNN will encode correct represen-



GPT 4o

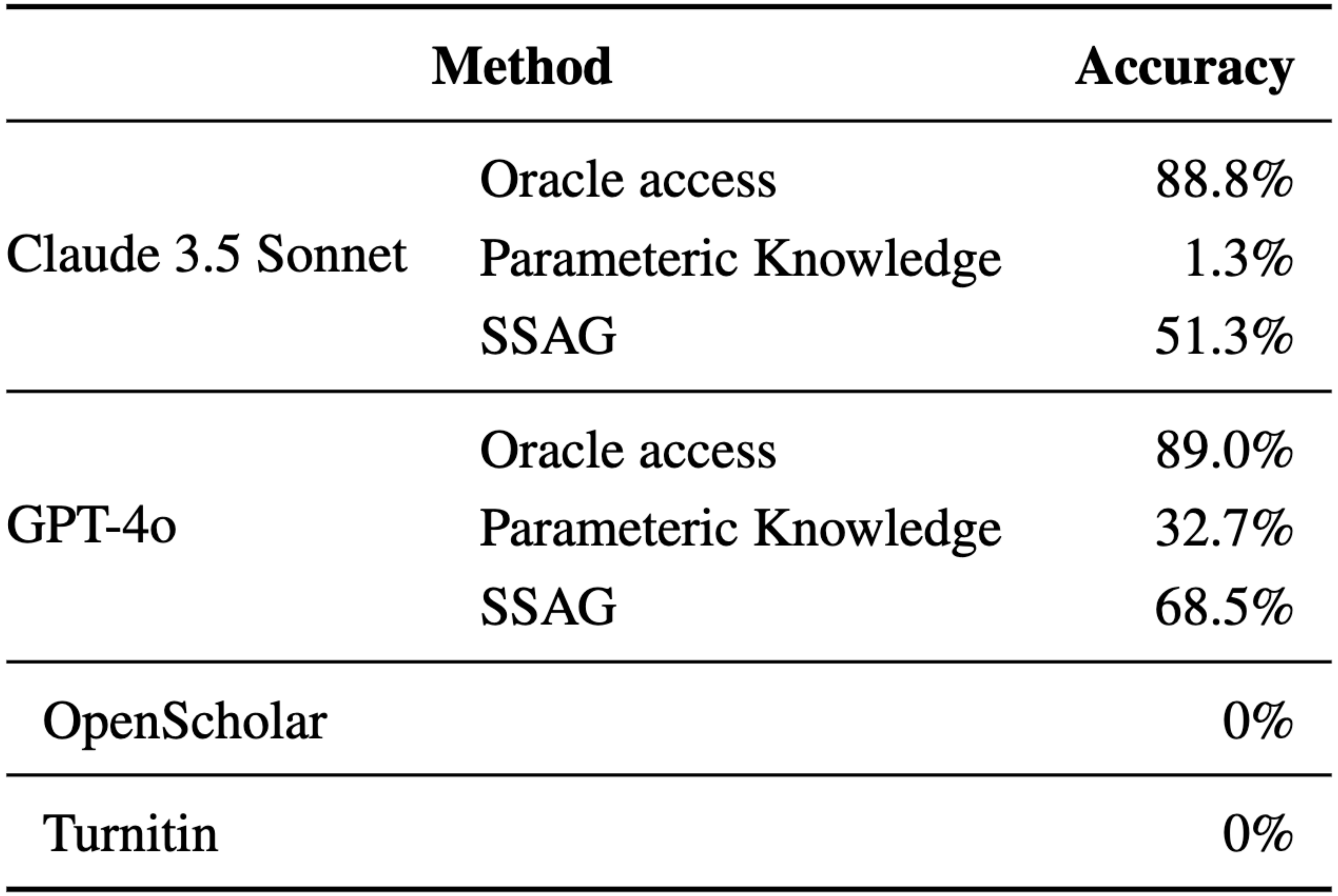
Plagiarized Research Article

Can plagiarism be automatically detected?

- Can detectors identify deliberately plagiarized articles?

Can plagiarism be automatically detected?

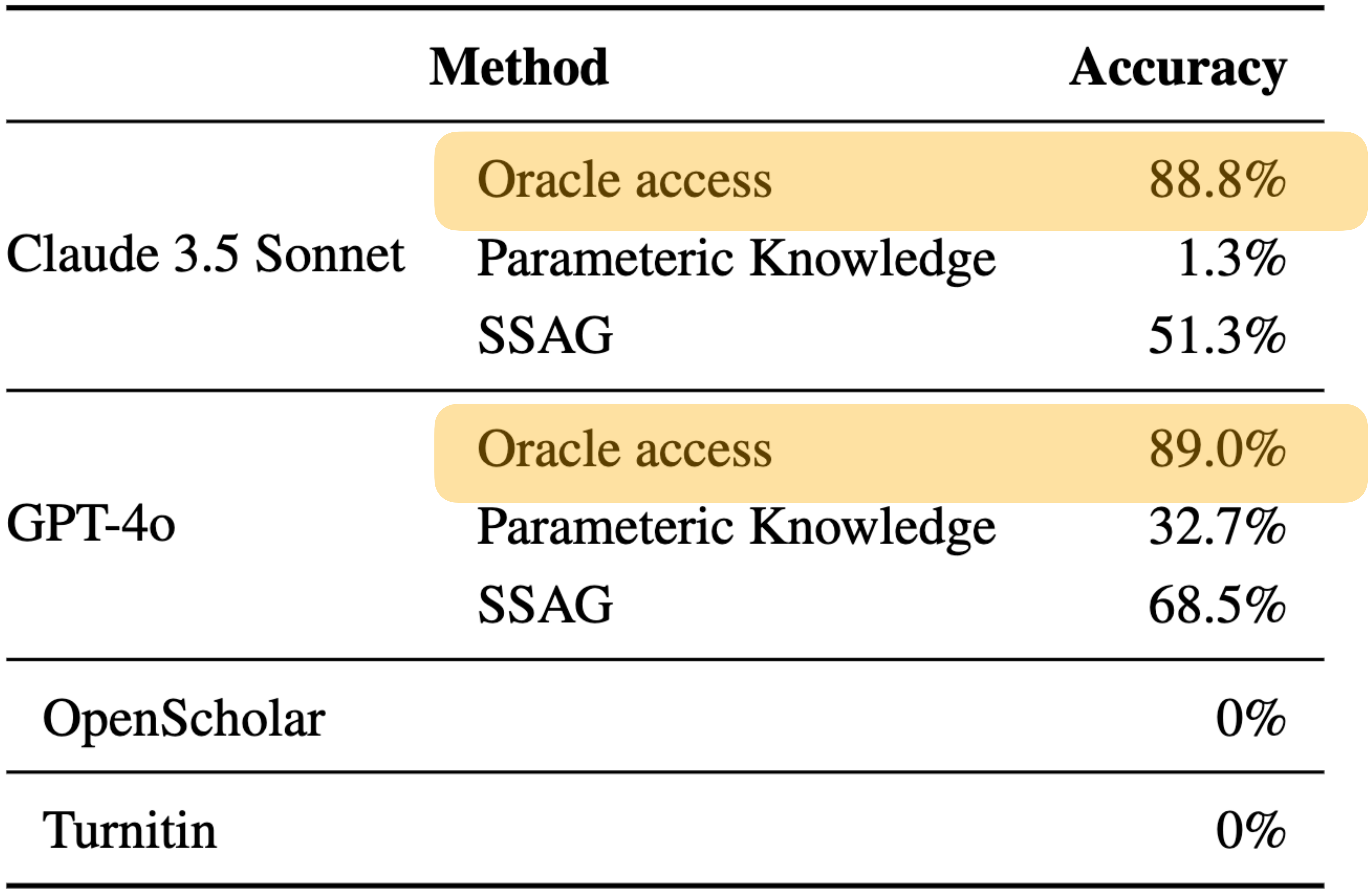
- **Can detectors identify deliberately plagiarized articles?**



Method		Accuracy
Claude 3.5 Sonnet	Oracle access	88.8%
	Parameteric Knowledge	1.3%
	SSAG	51.3%
GPT-4o	Oracle access	89.0%
	Parameteric Knowledge	32.7%
	SSAG	68.5%
OpenScholar		0%
Turnitin		0%

Can plagiarism be automatically detected?

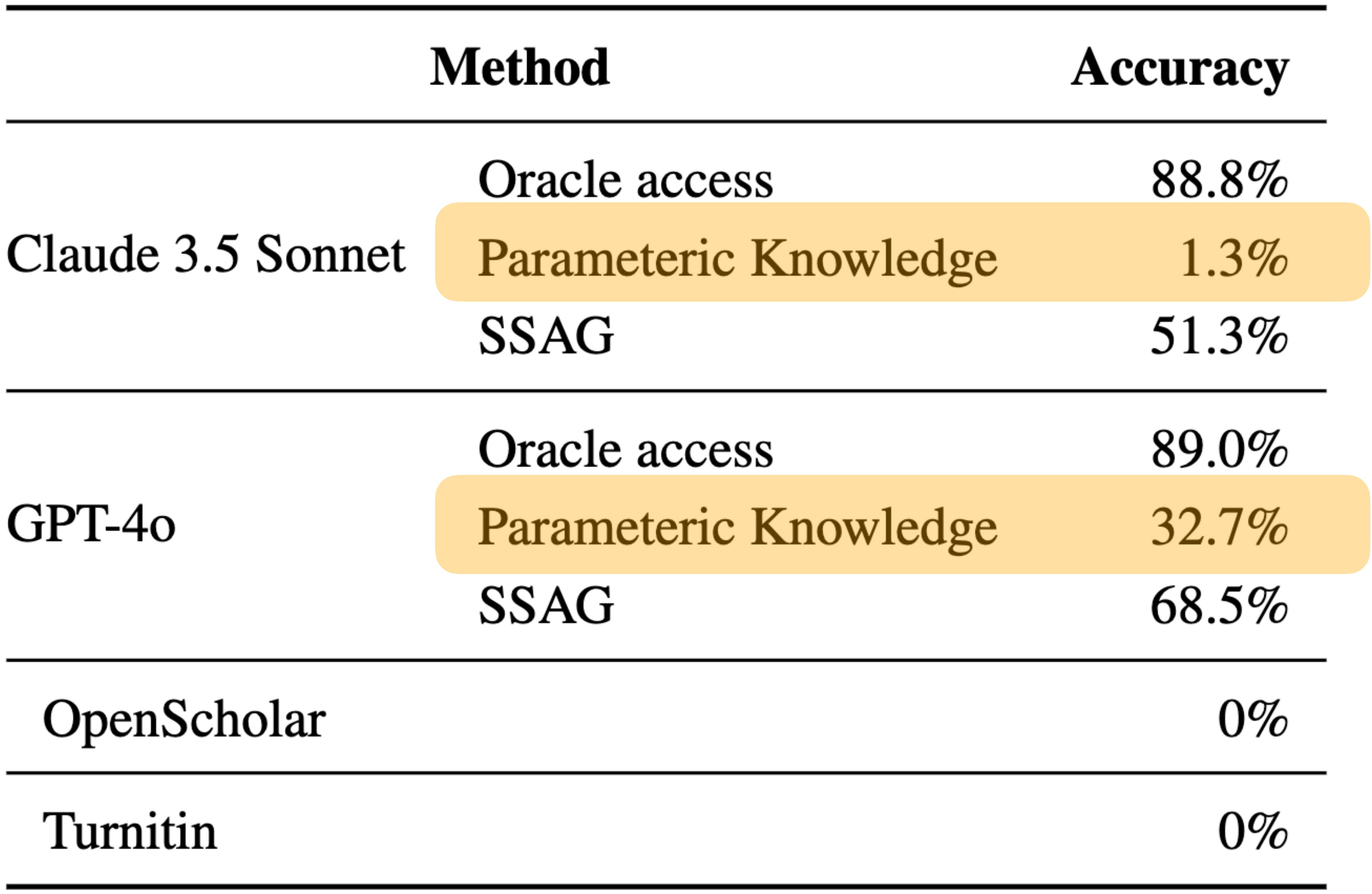
- Can detectors identify deliberately plagiarized articles?



	Method	Accuracy
Claude 3.5 Sonnet	Oracle access	88.8%
	Parameteric Knowledge	1.3%
	SSAG	51.3%
GPT-4o	Oracle access	89.0%
	Parameteric Knowledge	32.7%
	SSAG	68.5%
OpenScholar		0%
Turnitin		0%

Can plagiarism be automatically detected?

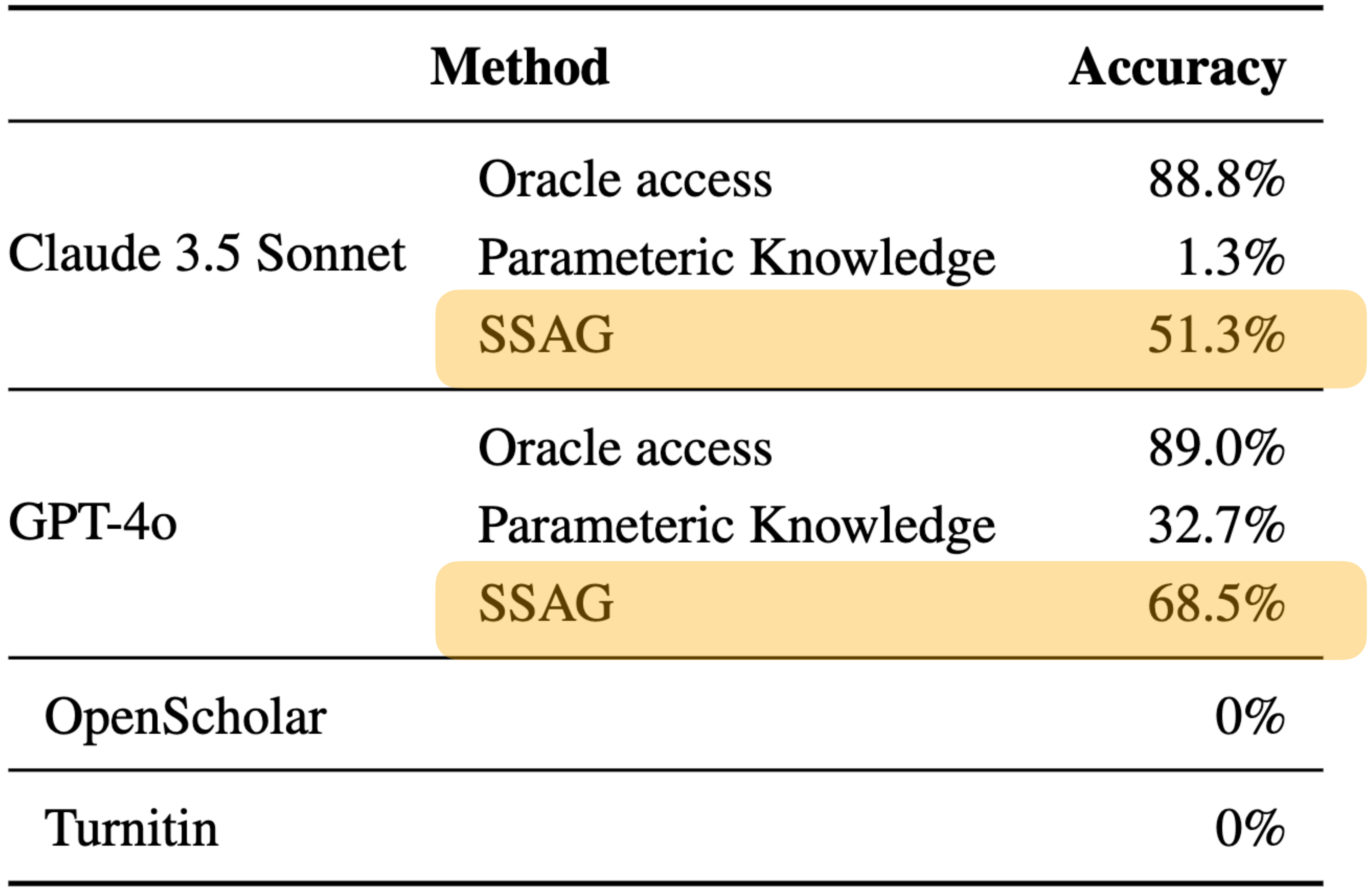
- Can detectors identify deliberately plagiarized articles?



Method		Accuracy
Claude 3.5 Sonnet	Oracle access	88.8%
	Parameteric Knowledge	1.3%
	SSAG	51.3%
GPT-4o	Oracle access	89.0%
	Parameteric Knowledge	32.7%
	SSAG	68.5%
OpenScholar		0%
Turnitin		0%

Can plagiarism be automatically detected?

- Can detectors identify deliberately plagiarized articles?



	Method	Accuracy
Claude 3.5 Sonnet	Oracle access	88.8%
	Parameteric Knowledge	1.3%
	SSAG	51.3%
GPT-4o	Oracle access	89.0%
	Parameteric Knowledge	32.7%
	SSAG	68.5%
OpenScholar		0%
Turnitin		0%

Plagiarism in human-written papers?

Plagiarism in human-written papers?

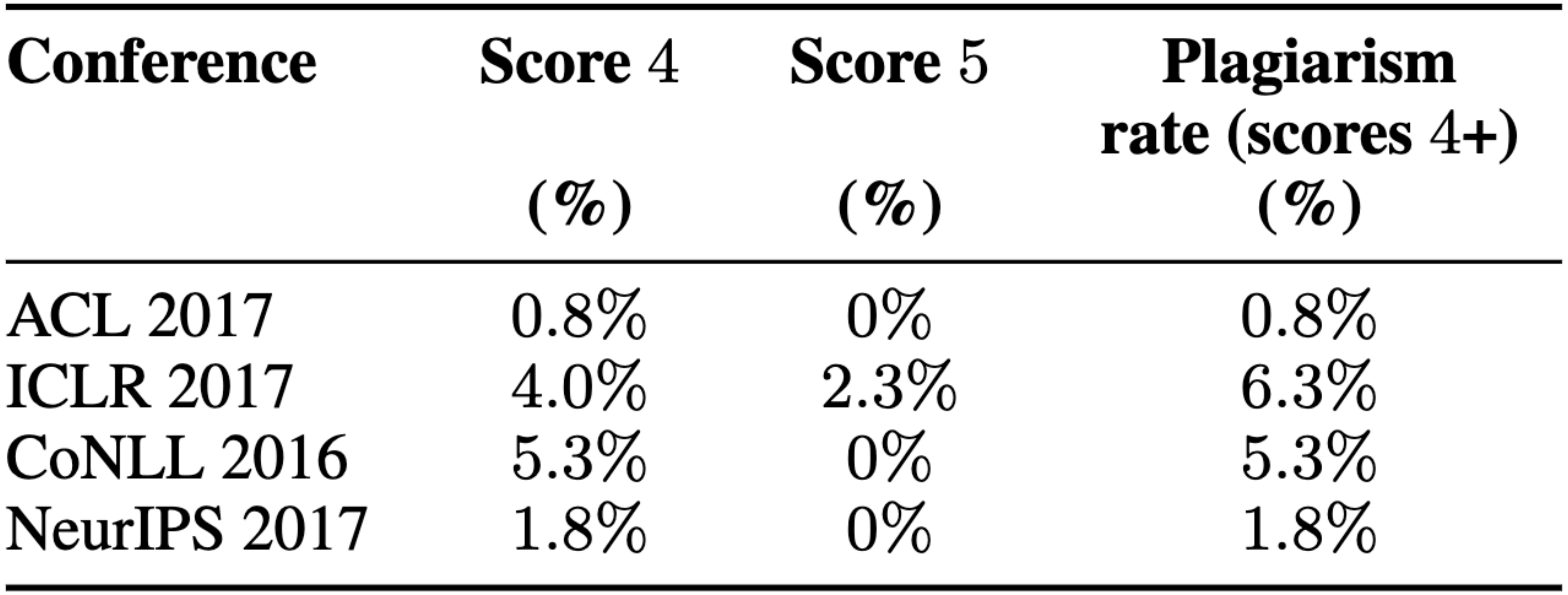
- **Experts only looked through AI-generated research;**

Plagiarism in human-written papers?

- **Experts only looked through AI-generated research;**
- **But experts regularly peer-review papers…**
    - So we extract signs of plagiarism in peer-reviews

Plagiarism in human-written papers?

- **Experts only looked through AI-generated research;**
- **But experts regularly peer-review papers…**
  - So we extract signs of plagiarism in peer-reviews



Conference	**Score** 4 (%)	**Score** 5 (%)	**Plagiarism rate (scores** 4**+)** (%)
ACL 2017	0.8%	0%	0.8%
ICLR 2017	4.0%	2.3%	6.3%
CoNLL 2016	5.3%	0%	5.3%
NeurIPS 2017	1.8%	0%	1.8%

Broader Implications of AI Scientists

- Regurgitate old ideas, without any attribution
- Overwhelm conferences
- Fracture scientific discourse
- Sow distrust about other legitimate AI capabilities

Broader Implications of AI Scientists

- Regurgitate old ideas, without any attribution
- Overwhelm conferences
- Fracture scientific discourse
- Sow distrust about other legitimate AI capabilities

All That Glitters is Not Novel: Plagiarism in AI Generated Research

By Tarun Gupta, Danish Pruthi

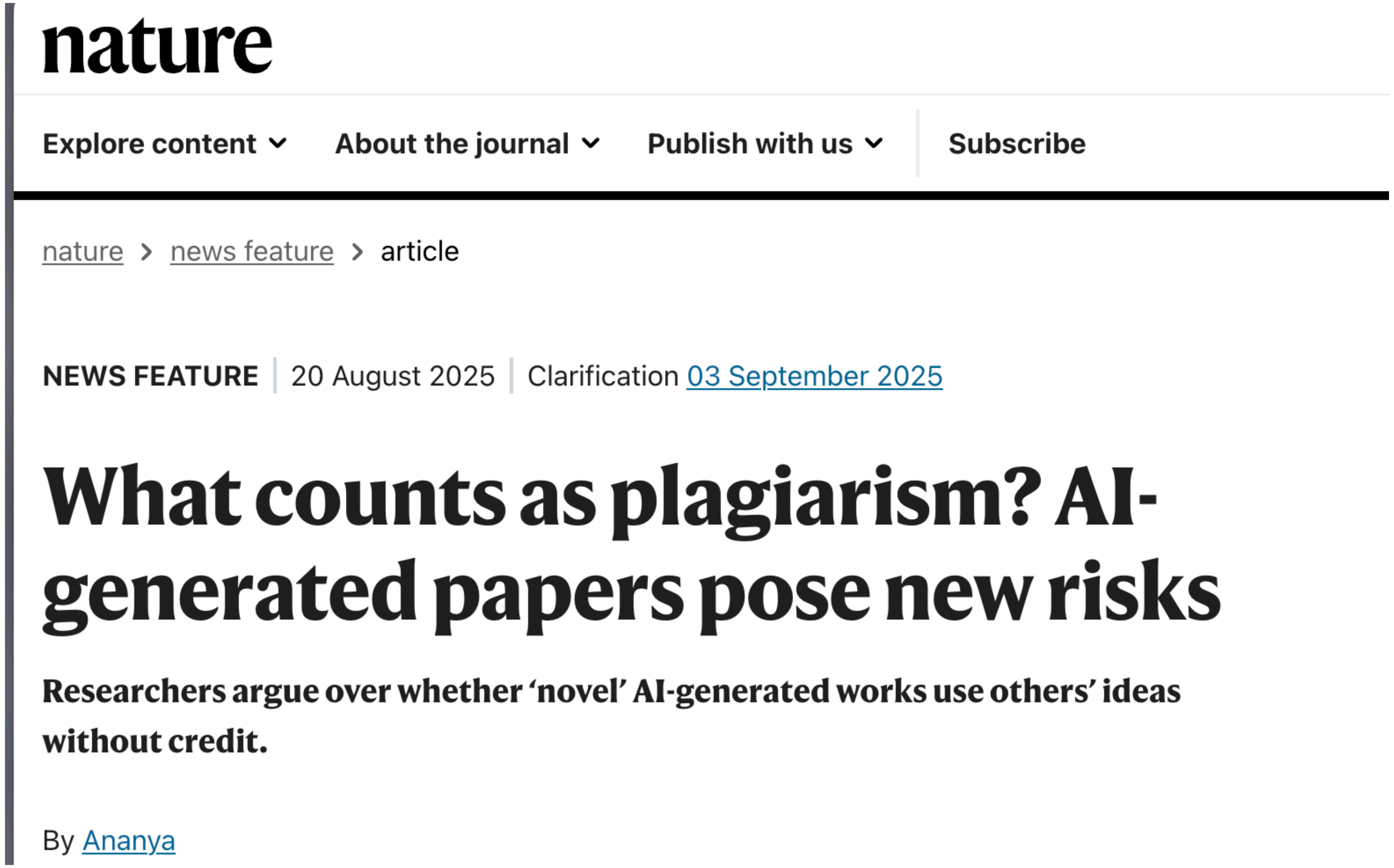
ACL 2025

Impact of the work

- **Paper received the outstanding paper award at ACL**

Impact of the work

- Paper received the outstanding paper award at ACL



NEWS FEATURE | 20 August 2025 | Clarification 03 September 2025

What counts as plagiarism? AI-generated papers pose new risks

Researchers argue over whether 'novel' AI-generated works use others' ideas without credit.

By Ananya

Broadly: Detecting AI Slop

AI Slop: ***low-quality, often nonsensical or misleading, content generated by artificial intelligence.***

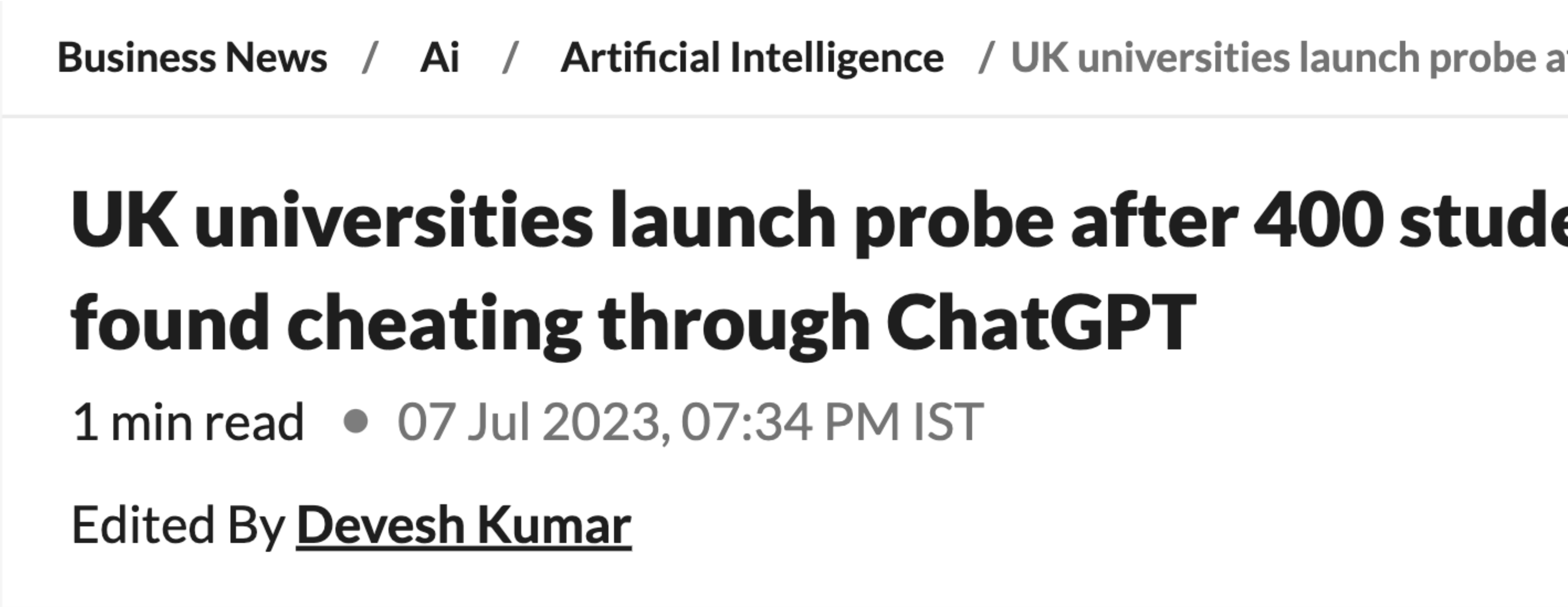
With 'AI slop' distorting our reality, the world is sleepwalking into disaster

By Nesrine Malik in the Guardian

AI-generated 'slop' is slowly killing the internet

By Arwa Mahdawi in the Guardian

Increasing Reports of Plagiarism

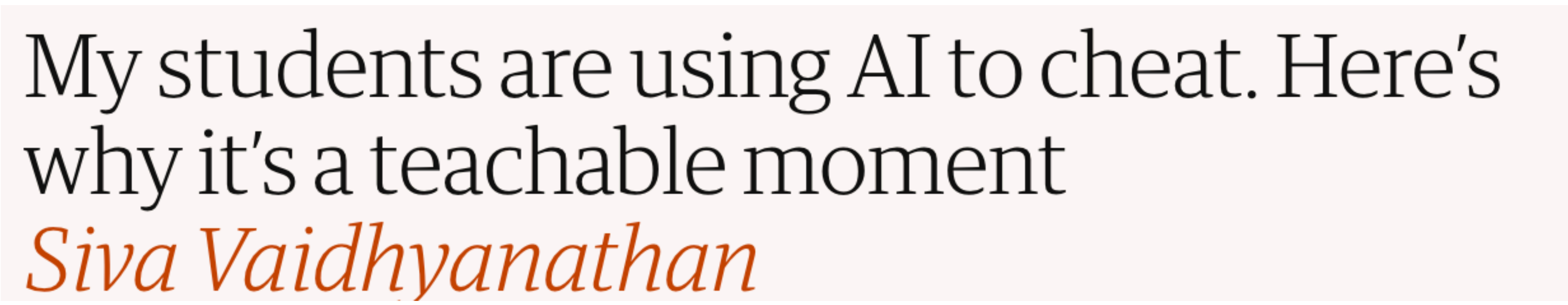


Business News / Ai / Artificial Intelligence / UK universities launch probe after 400 stud...

UK universities launch probe after 400 students found cheating through ChatGPT

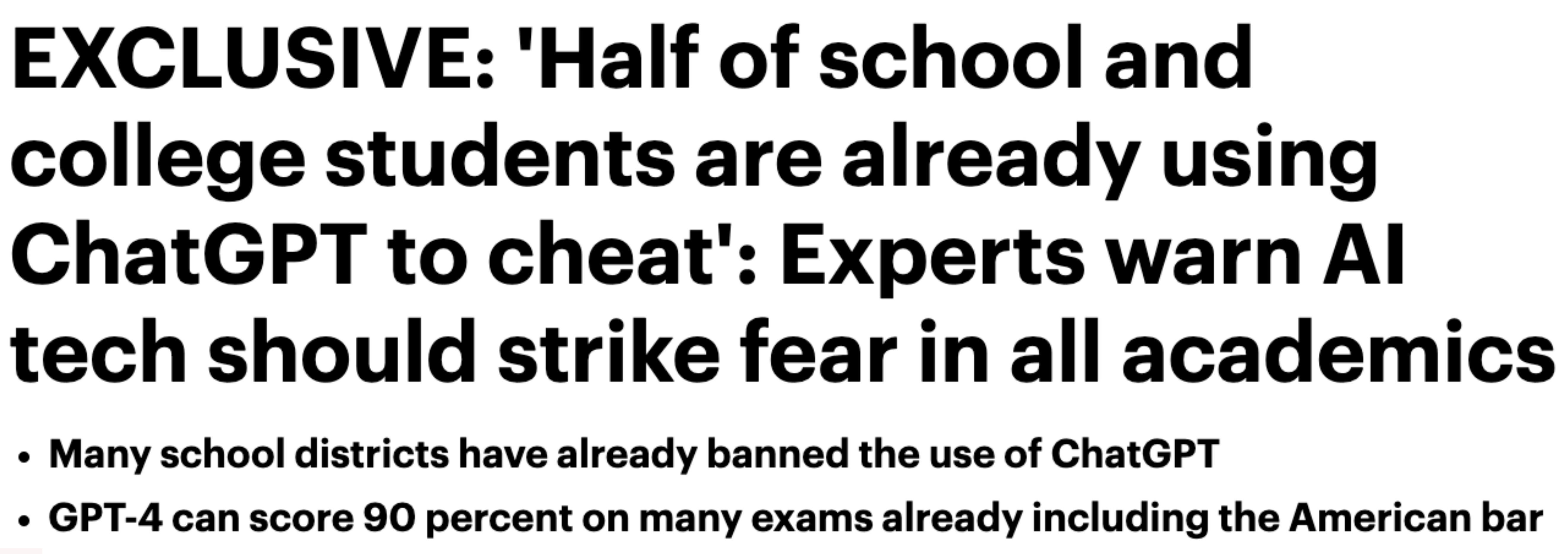
1 min read • 07 Jul 2023, 07:34 PM IST

Edited By Devesh Kumar



My students are using AI to cheat. Here's why it's a teachable moment

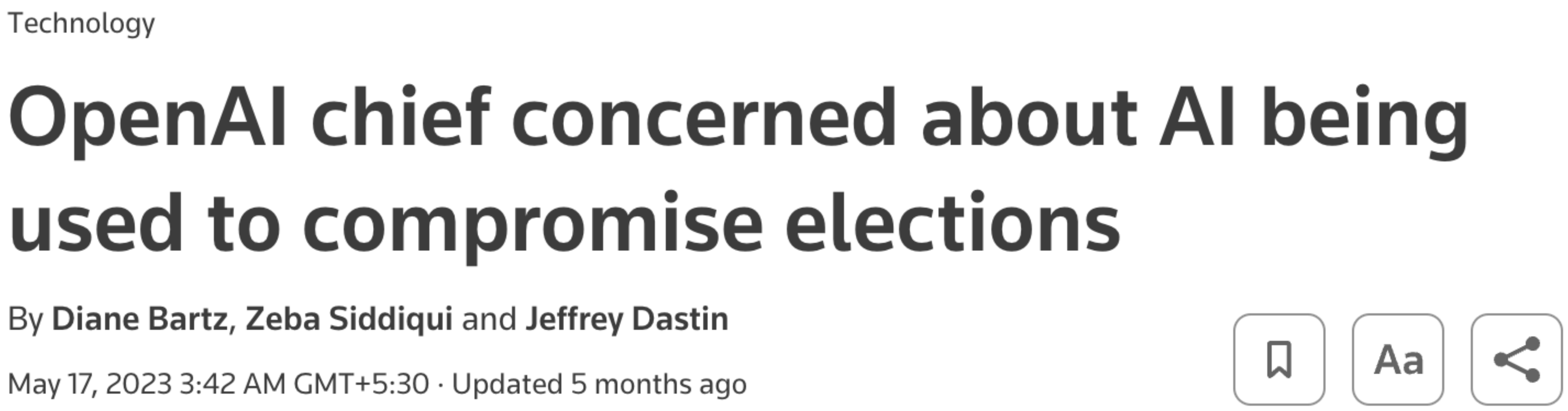
Siva Vaidhyanathan



EXCLUSIVE: 'Half of school and college students are already using ChatGPT to cheat': Experts warn AI tech should strike fear in all academics

- **Many school districts have already banned the use of ChatGPT**
- **GPT-4 can score 90 percent on many exams already including the American bar**

Increasing Concerns of Targeted Misinformation

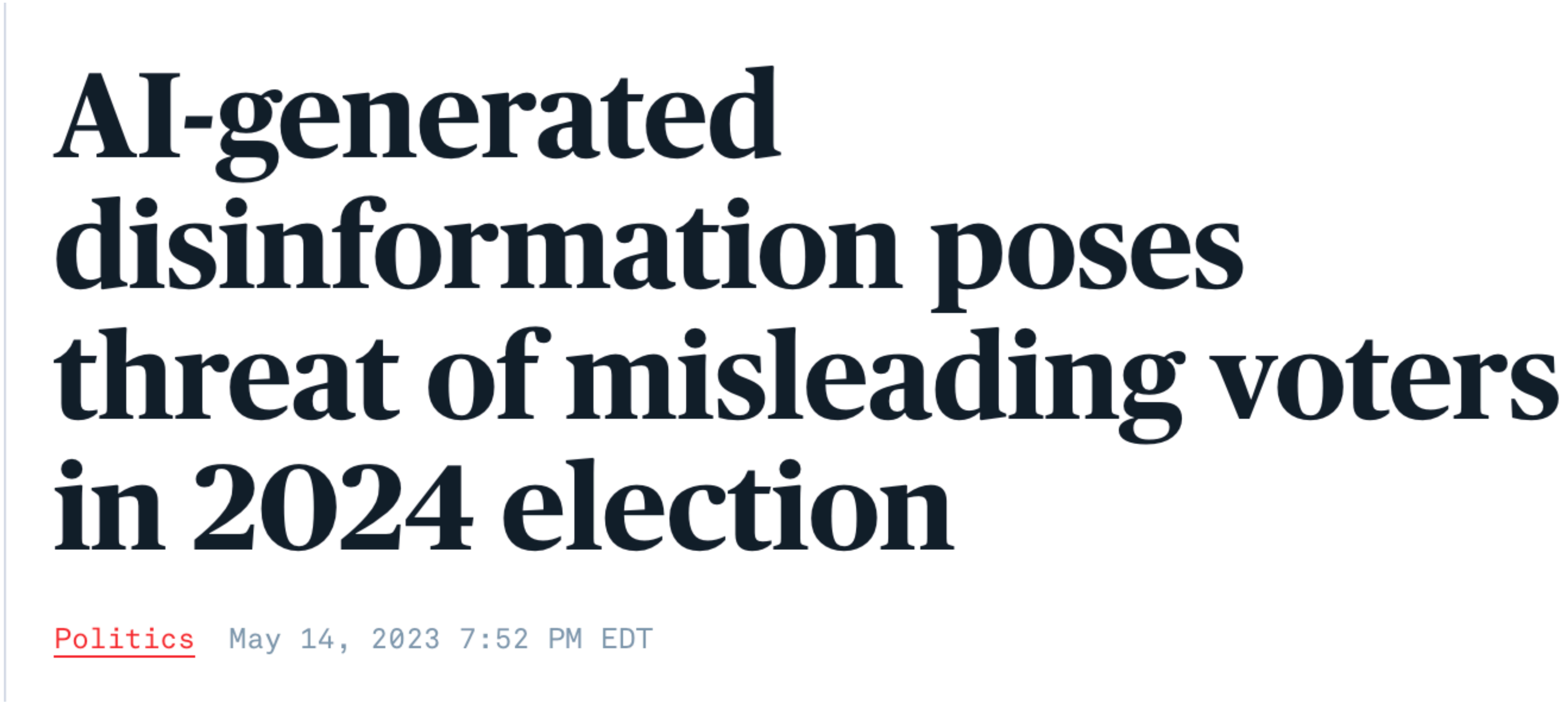


Technology

OpenAI chief concerned about AI being used to compromise elections

By **Diane Bartz**, **Zeba Siddiqui** and **Jeffrey Dastin**

May 17, 2023 3:42 AM GMT+5:30 · Updated 5 months ago



AI-generated disinformation poses threat of misleading voters in 2024 election

Politics May 14, 2023 7:52 PM EDT

Need to Distinguish LLM Outputs from Human Text

- **Detect plagiarism**
- **Combat targeted large-scale misinformation/spam/abuse**
- **Avoid re-training on LLM outputs**

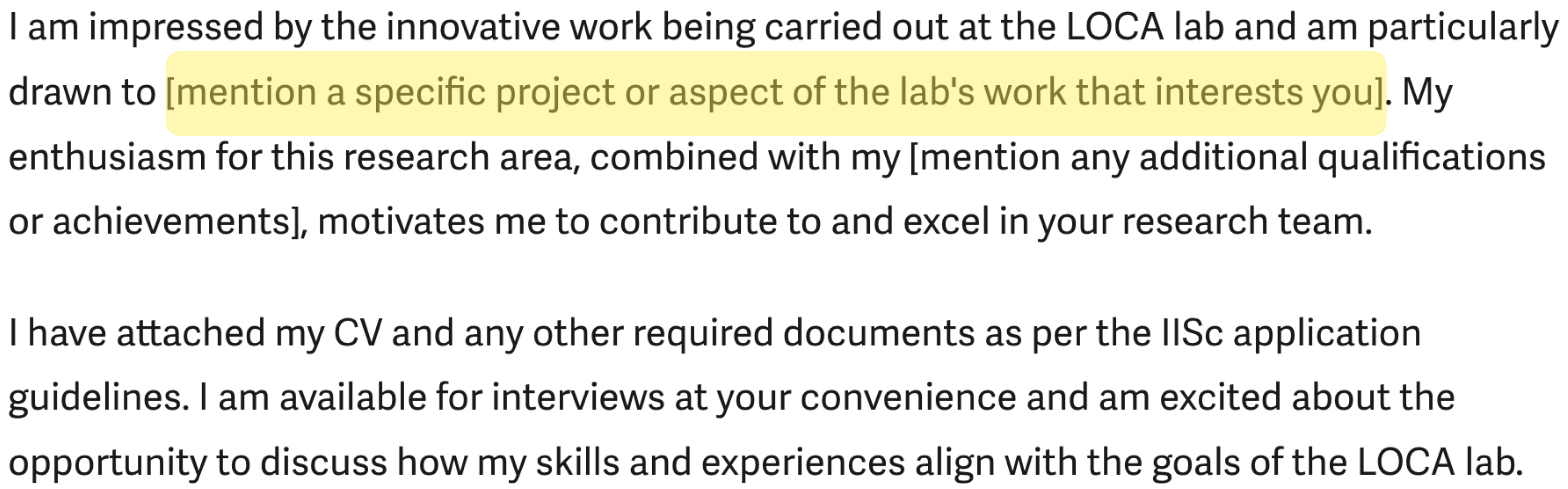
Possible Approaches

Possible Approaches

- **You might get lucky:**

Possible Approaches

- **You might get lucky:**

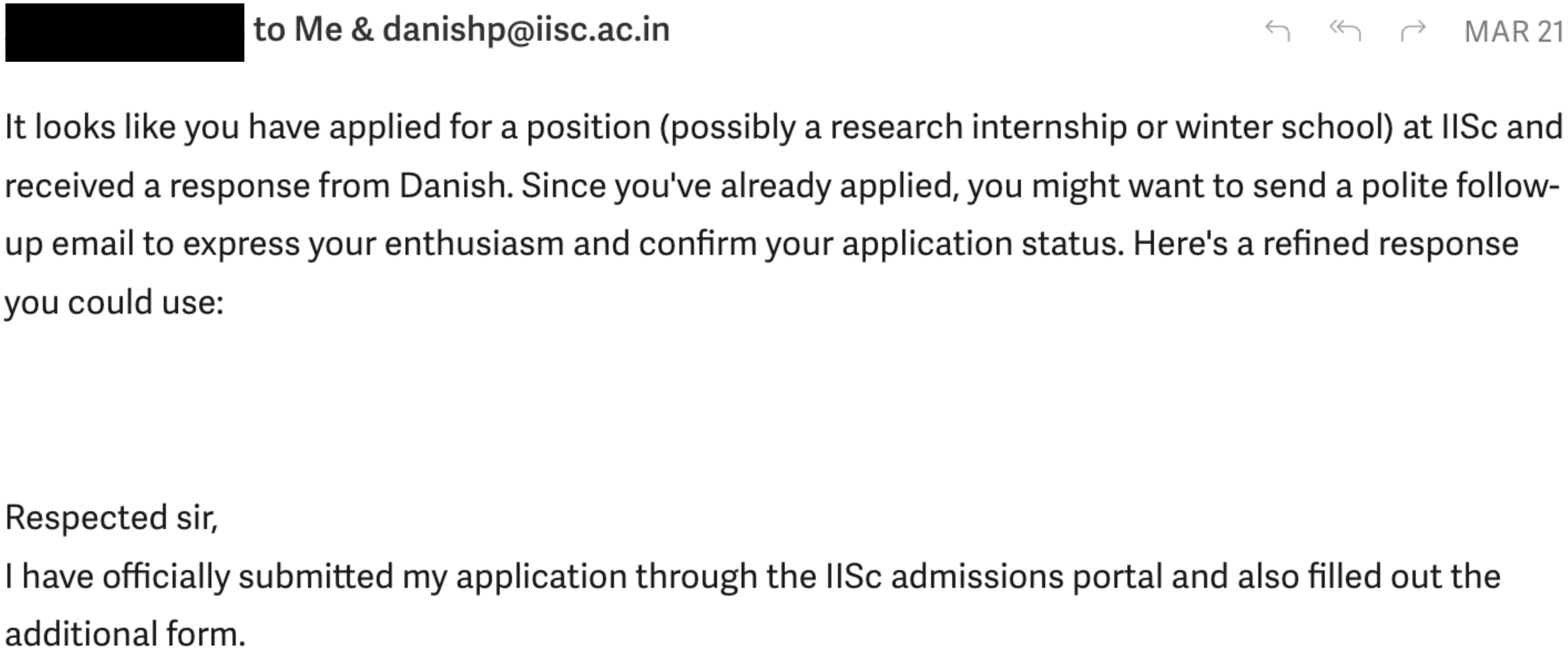


I am impressed by the innovative work being carried out at the LOCA lab and am particularly drawn to [mention a specific project or aspect of the lab's work that interests you]. My enthusiasm for this research area, combined with my [mention any additional qualifications or achievements], motivates me to contribute to and excel in your research team.

I have attached my CV and any other required documents as per the IISc application guidelines. I am available for interviews at your convenience and am excited about the opportunity to discuss how my skills and experiences align with the goals of the LOCA lab.

Possible Approaches

- **You might get lucky:**



[redacted] to Me & danishp@iisc.ac.in MAR 21

It looks like you have applied for a position (possibly a research internship or winter school) at IISc and received a response from Danish. Since you've already applied, you might want to send a polite follow-up email to express your enthusiasm and confirm your application status. Here's a refined response you could use:

Respected sir,
I have officially submitted my application through the IISc admissions portal and also filled out the additional form.

Possible Approaches

- Model developers could store all the responses generated

Possible Approaches

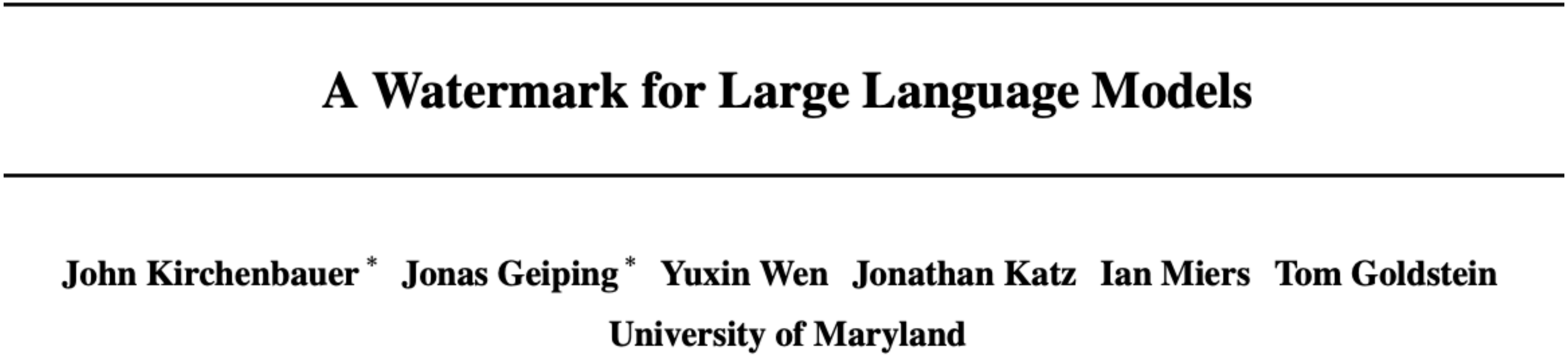
- Model developers could store all the responses generated

Possible Approaches

- Model developers could store all the responses generated

- Wouldn't work for open-source models
- Privacy concerns

Possible Approaches: Watermarking



A Watermark for Large Language Models

John Kirchenbauer[*] Jonas Geiping[*] Yuxin Wen Jonathan Katz Ian Miers Tom Goldstein
University of Maryland

ICML 2023 (Best Paper Award)

Watermarking Language Models

Approach from Kirchenbauer et al.

Watermarking Language Models

I work at the Indian Institute of Science. I live in ________

Approach from Kirchenbauer et al.

Watermarking Language Models

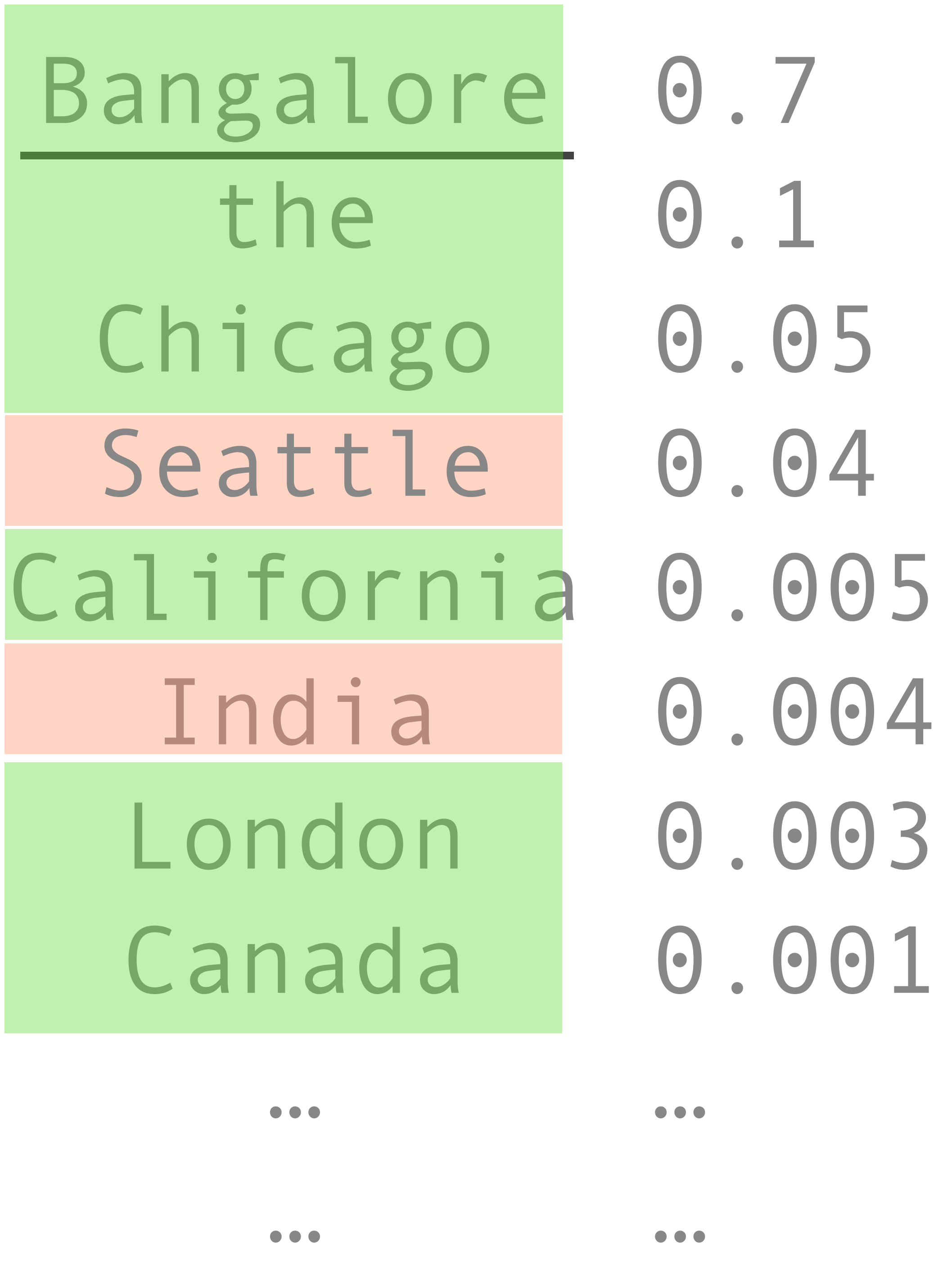
I work at the Indian Institute of Science. I live in

Bangalore	0.7
the	0.1
Chicago	0.05
Seattle	0.04
California	0.005
India	0.004
London	0.003
Canada	0.001
…	…
…	…

Approach from Kirchenbauer et al.

Watermarking Language Models

I work at the Indian Institute of Science. I live in



🟩	Bangalore	0.7
🟩	the	0.1
🟩	Chicago	0.05
🟥	Seattle	0.04
🟩	California	0.005
🟥	India	0.004
🟩	London	0.003
🟩	Canada	0.001
	…	…
	…	…

1. Randomly partition the vocabulary based on the last word

2. Don't generate a word from the red list

Approach from Kirchenbauer et al.

How does detection work

- **If all words are from green list, then we know it's from a model**

Approach from Kirchenbauer et al.

Watermarking Language Models

I work at the Indian Institute of Science. I live in

🟩	Bangalore	0.7
🟩	the	0.1
🟩	Chicago	0.05
🟥	Seattle	0.04
🟩	California	0.005
🟥	India	0.004
🟩	London	0.003
🟩	Canada	0.001
	…	…
	…	…

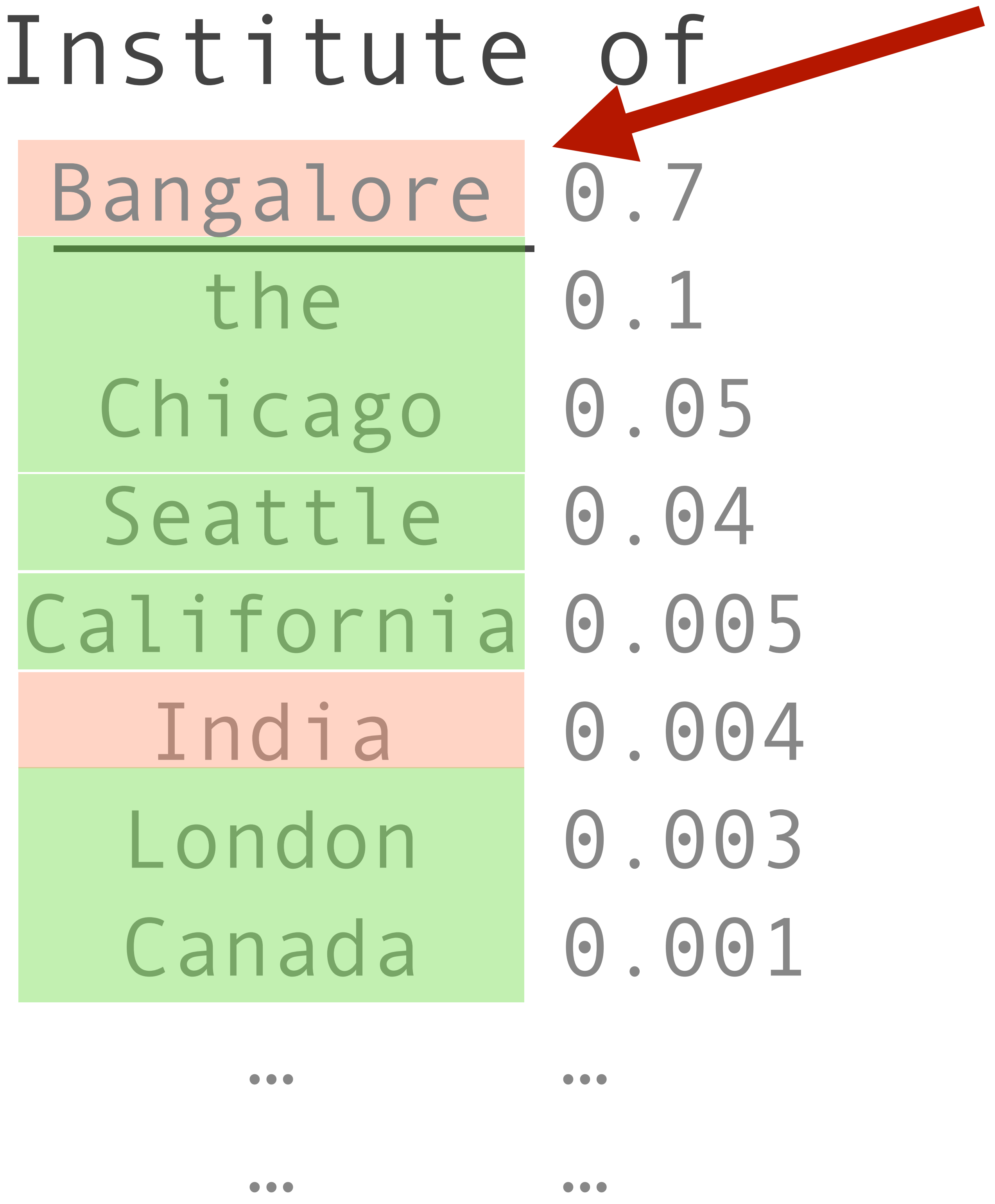
1. Randomly partition the vocabulary based on the last word

2. Don't generate a word from the red list

Approach from Kirchenbauer et al.

Watermarking Language Models

I work at the Indian Institute of Science. I live in



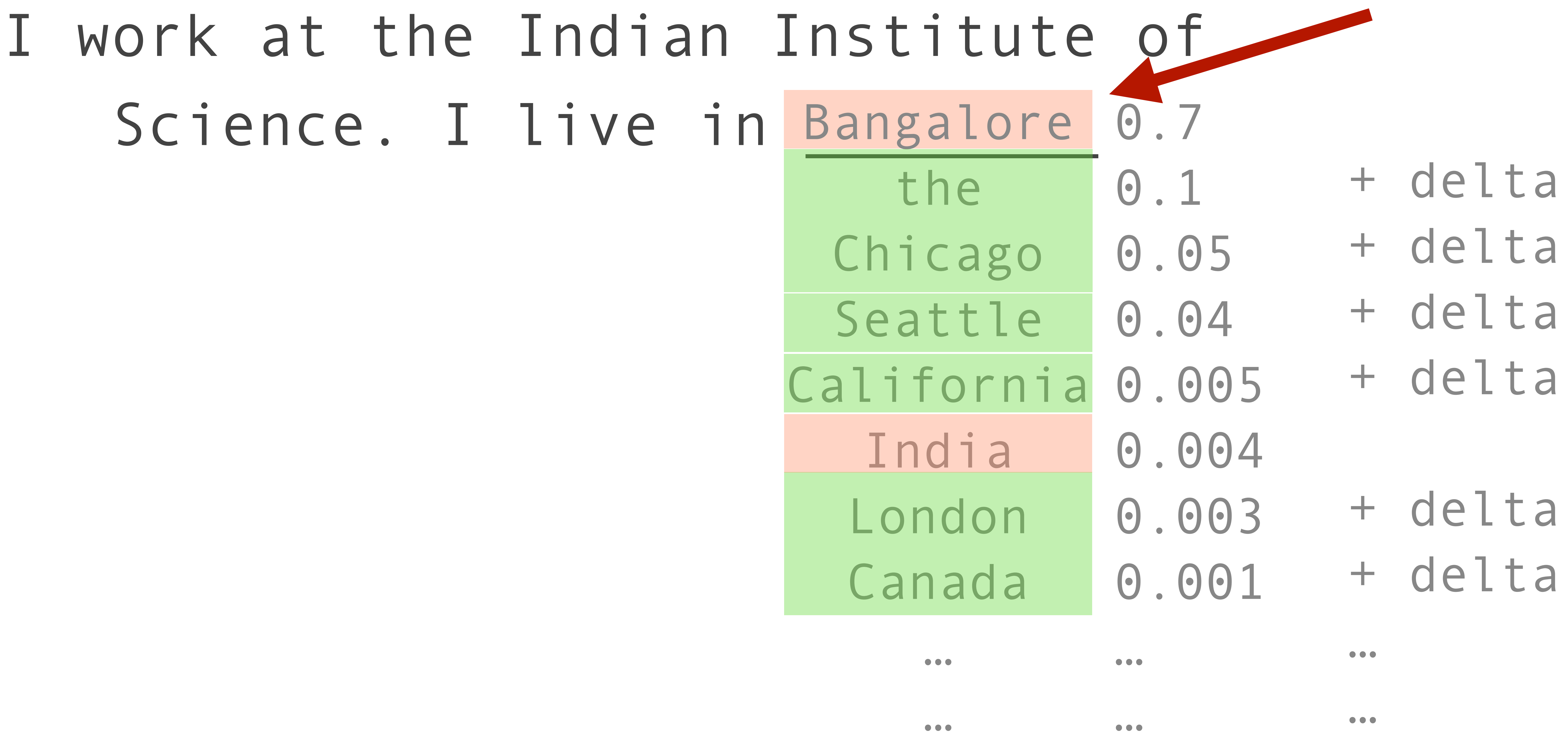
🟥	Bangalore	0.7
🟩	the	0.1
🟩	Chicago	0.05
🟩	Seattle	0.04
🟩	California	0.005
🟥	India	0.004
🟩	London	0.003
🟩	Canada	0.001
	…	…
	…	…

1. Randomly partition the vocabulary based on the last word

2. Don't generate a word from the red list

Approach from Kirchenbauer et al.

Watermarking Language Models



I work at the Indian Institute of Science. I live in

🟥 Bangalore	0.7	
🟩 the	0.1	+ delta
🟩 Chicago	0.05	+ delta
🟩 Seattle	0.04	+ delta
🟩 California	0.005	+ delta
🟥 India	0.004	
🟩 London	0.003	+ delta
🟩 Canada	0.001	+ delta
…	…	…
…	…	…

1. Randomly partition the vocabulary based on the last word

~~2. Don't generate a word from the red list~~ Boost green words

Approach from Kirchenbauer et al.

Watermarking Language Models

Soft

I work at the Indian Institute of Science. I live in

Bangalore	0.7	
the	0.1	+ delta
Chicago	0.05	+ delta
Seattle	0.04	+ delta
California	0.005	+ delta
India	0.004	
London	0.003	+ delta
Canada	0.001	+ delta
…	…	…
…	…	…

1. Randomly partition the vocabulary based on the last word

~~2. Don't generate a word from the red list~~ Boost green words

Approach from Kirchenbauer et al.

How does detection work

- If all words are from green list, then we know it's from a model

Approach from Kirchenbauer et al.

How does detection work

- If ~~all~~ words are from green list, then we know it's from a model

  large fraction

  (than what we would expect randomly)

Approach from Kirchenbauer et al.

How does detection work

- If ~~all~~ words are from green list, then we ~~know~~ it's from a model

  large fraction
  (than what we would expect randomly)

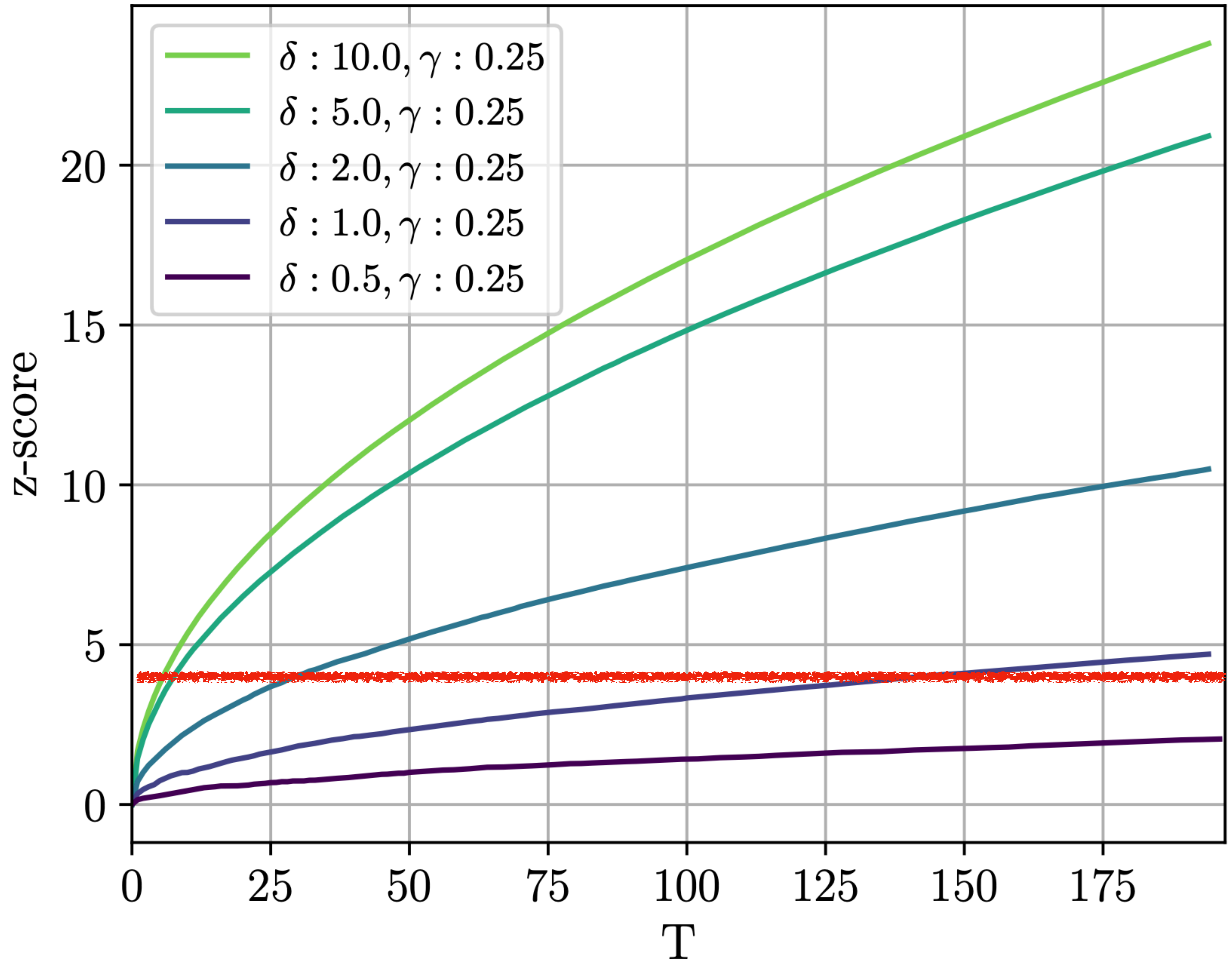
  suspect
  (we can run a statistical test)

Approach from Kirchenbauer et al.

Useful Properties

- Can be applied to any language model
- Knobs to play around with the watermarking strength
- Watermarking is conceptually simple & computationally cheap
- Detection does not depend on the model probabilities

Efficacy of watermarking



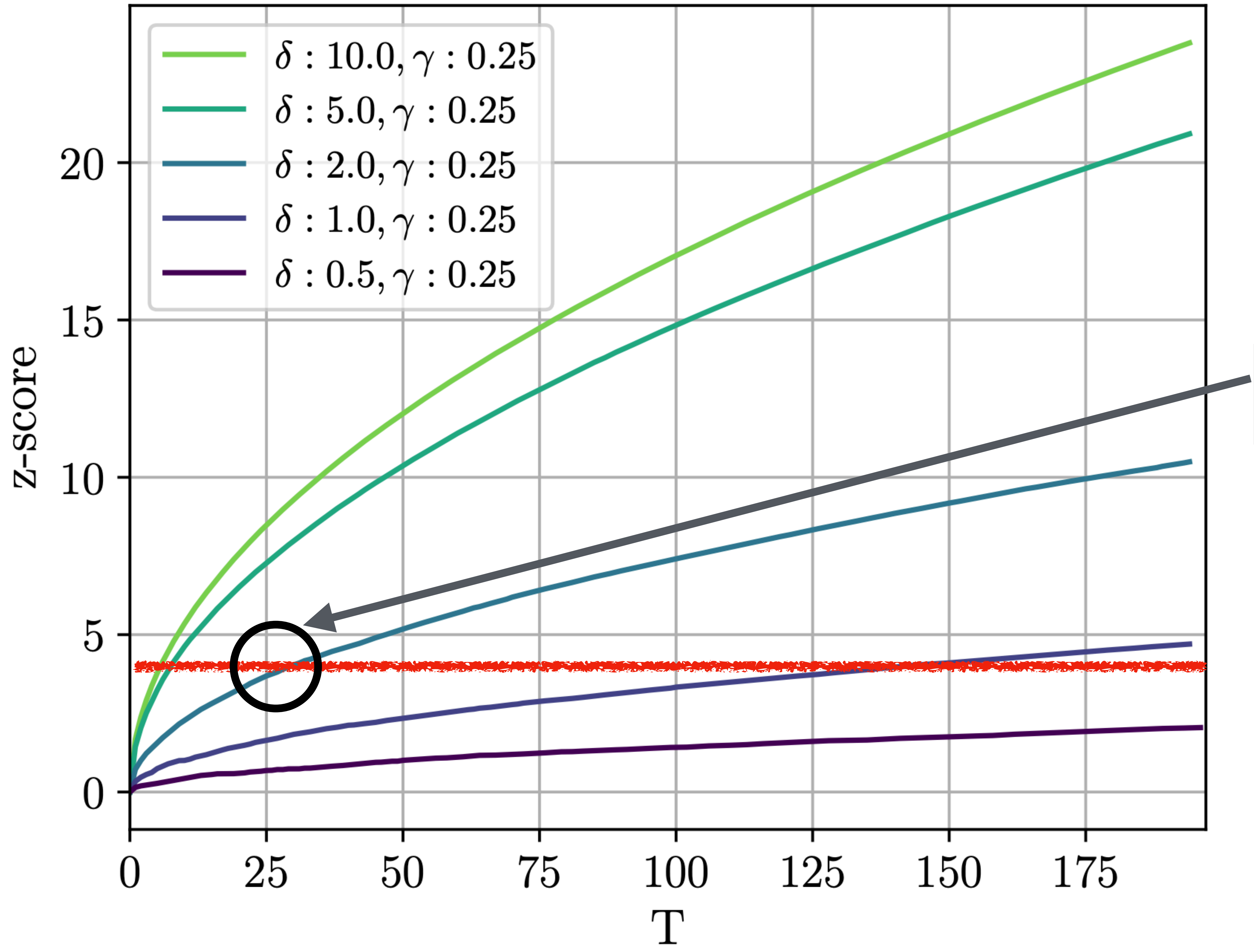
$$z = (|x_G| - \gamma T)/\sqrt{T\gamma(1-\gamma)}$$

$$z \propto r_G\sqrt{T}$$

False positive rate is 3 x 10^{-5}

From Kirchenbauer et al.

Efficacy of watermarking



$$z = (|x_G| - \gamma T)/\sqrt{T\gamma(1-\gamma)}$$

$$z \propto r_G\sqrt{T}$$

Often, about 30 tokens suffice

False positive rate is 3 x 10^{-5}

From Kirchenbauer et al.

Our Work

- **Downstream effects of watermarking (Findings of EMNLP, 2024)**
  - By Anirudh Ajith, Sameer Singh, Danish Pruthi
  - https://arxiv.org/abs/2311.09816
- **Undoing (or reverse-engineering) watermarking (EMNLP 2024)**
  - By Saksham Rastogi, Danish Pruthi
  - https://arxiv.org/abs/2411.05277
- **Watermarking your own content (ICML 2025)**
  - By Saksham Rastogi, Pratyush Maini, Danish Pruthi
  - https://arxiv.org/abs/2504.13416

In Practice: Almost No Adoption

In Practice: Almost No Adoption

- Corporations are (allegedly) worried that they'll lose customers

- Requires all developers to watermark

In Practice

- Hundreds of available detectors:
    - Originality,
    - GPTZero,
    - DetectGPT,
    - Pangram, etc.

- Which work with varying effectiveness

Thank you

Danish Pruthi

Webpage: https://danishpruthi.com/

Papers: http://bit.ly/danish037

Email: danishp@iisc.ac.in