TIER2 Enhancing Trust, Integrity and Efficiency in Research through next-level Reproducibility

Large Language Models, reproducibility and trust in scholarly work

Tony Ross-Hellauer
Open & Reproducible Research Group
TU Graz & Know-Center GmbH

KC Colloquium, 6th October 2023

Funded by the European Union.
Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission.
Neither the EU nor the EC can be held responsible for them.

Recent advances in Large Language Models (LLMs) have captured the public attention and raised interest in the potential for such AI technologies to transform workflows across a range of areas, including scientific work.

There is substance to the hype

- (Dell'Acqua et al. 2023) performed randomised experiments with management consultants completing "realistic, complex, and knowledge-intensive tasks"
- For task where AI known to perform well, ChatGPT-4 increased:
  - Speed >25%
  - (Human-rated) performance >40%
  - Task completion >10%
- *But*, on task beyond known utility, almost 20% decrease in correctness of results

Figure 2: **Performance Distribution - Inside the Frontier**

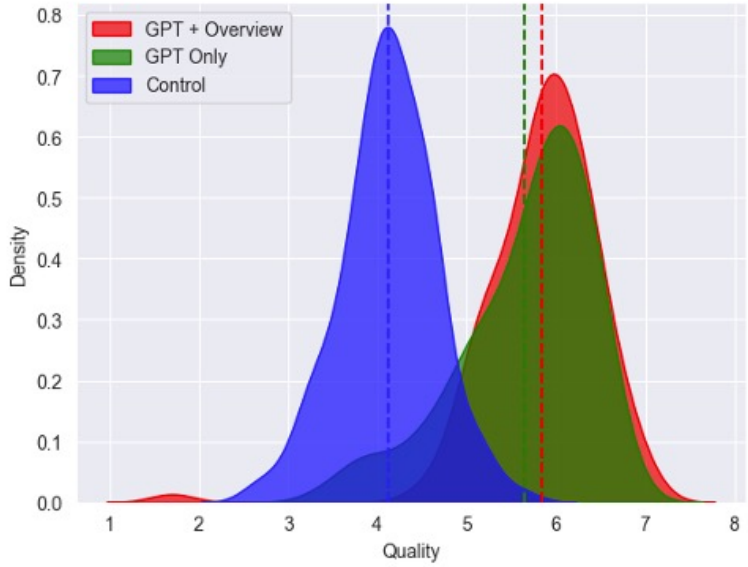
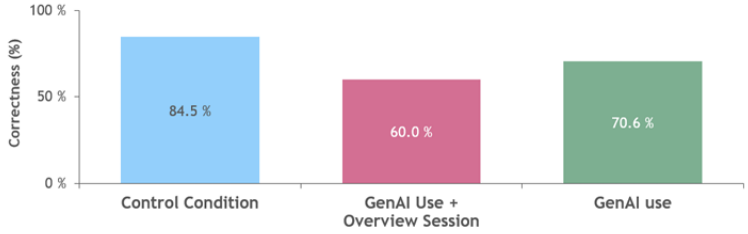


Figure 7: **Performance - Outside the Frontier**



Dell'Acqua, F et al. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." SSRN Working Paper. https://doi.org/10.2139/ssrn.4573321.

https://arxiv.org/abs/2303.10130

Economics > General Economics

[Submitted on 17 Mar 2023 (v1), last revised 21 Aug 2023 (this version, v5)]

GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

Tyna Eloundou, Sam Manning, Pamela Mishkin, Daniel Rock

We investigate the potential implications of large language models (LLMs), such as Generative Pre-trained Transformers (GPTs), on the U.S. labor market, focusing on the increased capabilities arising from LLM-powered software compared to LLMs on their own. Using a new rubric, we assess occupations based on their alignment with LLM capabilities, integrating both human expertise and GPT-4 classifications. Our findings reveal that around 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of LLMs, while approximately 19% of workers may see at least 50% of their tasks impacted. We do not make predictions about the development or adoption timeline of such LLMs. The projected effects span all wage levels, with higher-income jobs potentially facing greater exposure to LLM capabilities and LLM-powered software. Significantly, these impacts are not restricted to industries with higher recent productivity growth. Our analysis suggests that, with access to an LLM, about 15% of all worker tasks in the US could be completed significantly faster at the same level of quality. When incorporating software and tooling built on top of LLMs, this share increases to between 47 and 56% of all tasks. This finding implies that LLM-powered software will have a substantial effect on scaling the economic impacts of the underlying models. We conclude that LLMs such as GPTs exhibit traits of general-purpose technologies, indicating that they could have considerable economic, social, and policy implications.

Subjects: **General Economics (econ.GN)**; Artificial Intelligence (cs.AI); Computers and Society (cs.CY)
Cite as: arXiv:2303.10130 **[econ.GN]**
(or arXiv:2303.10130v5 **[econ.GN]** for this version)
https://doi.org/10.48550/arXiv.2303.10130

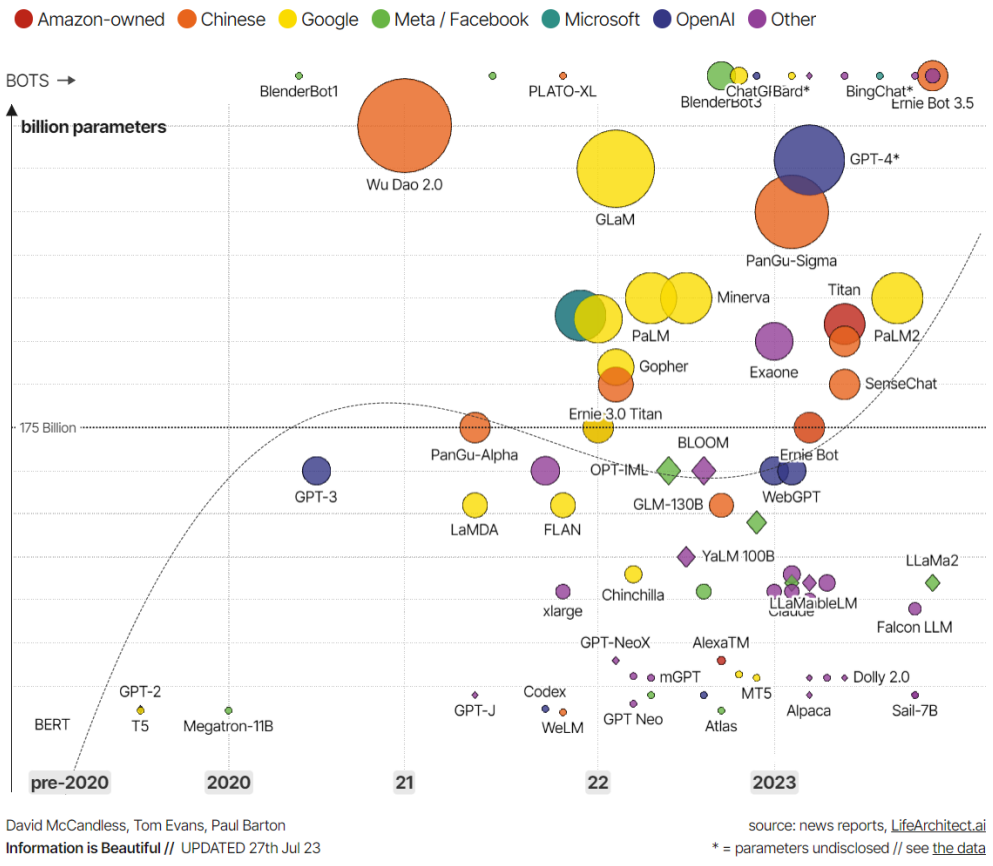
Submission history

From: Daniel Rock [view email]

TIER 2

Rapid expansion

- Growth: parameters, capabilities
- Generalised models exhibit emergent capabilities beyond simple textual comprehension/prediction: domain specialism, ideation, text summarisation, chain-of-thought reasoning
- Capabilities often identified through user interaction:
  - E.g., “Let’s think step by step” prompting to identify GPT-3 as capable of zero-shot reasoning (Kojima et al. 2022)



Source: https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/

Kojima, T et al. 2022. “Large Language Models Are Zero-Shot Reasoners.” *Advances in Neural Information Processing Systems* 35: 22199–213.

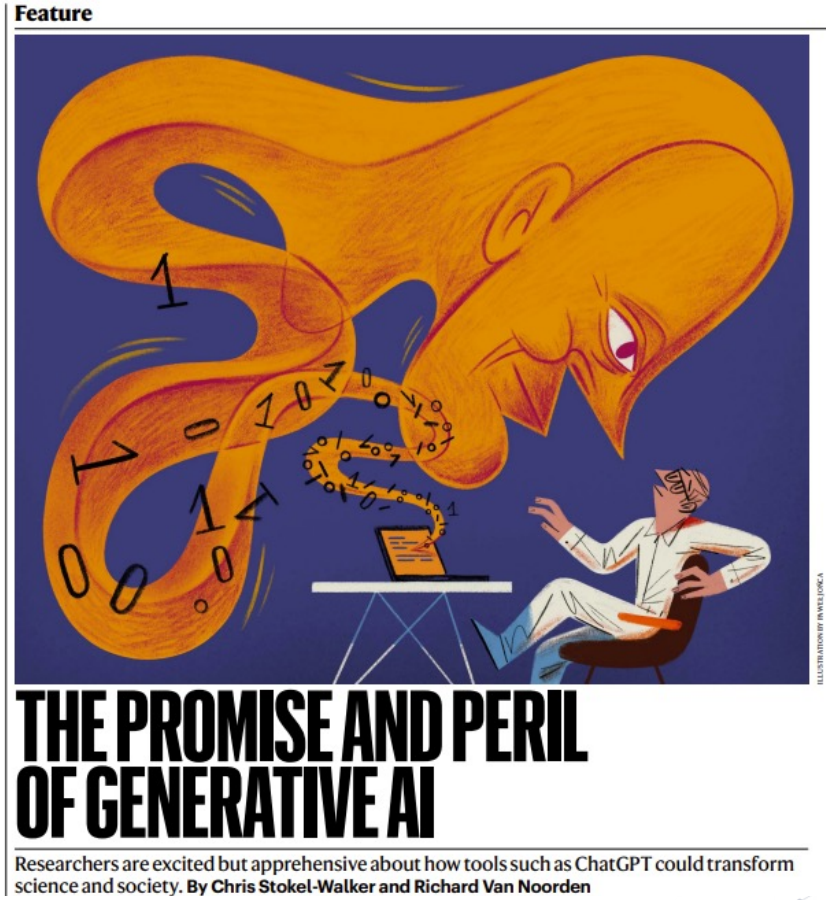
The jagged frontier of AI capabilities

- Notion of a ***jagged frontier*** where AI excels in some tasks, less adequate in others

- “Tasks with the same perceived difficulty may be on one side or the other of the frontier” (Dell’Acqua et al. 2023)

Potential impacts on research

- Recent Delphi study highlighted “transformative potential of LLMs in science, particularly in administrative, creative, and analytical tasks” (Fecher et al. 2023)
- Rich experimentation examining potential for scientific tasks like:
  - research design/ideation, information retrieval, evidence synthesis, assessment, textual analysis, code/analysis scripts, dissemination (translation, summarisation, etc.)
- Potential downsides too of course …


Feature

THE PROMISE AND PERIL OF GENERATIVE AI

Researchers are excited but apprehensive about how tools such as ChatGPT could transform science and society. By Chris Stokel-Walker and Richard Van Noorden

Fecher B et al. 2023. “Friend or Foe? Exploring the Implications of Large Language Models on the Science System.” arXiv. https://doi.org/10.48550/arXiv.2306.09928.

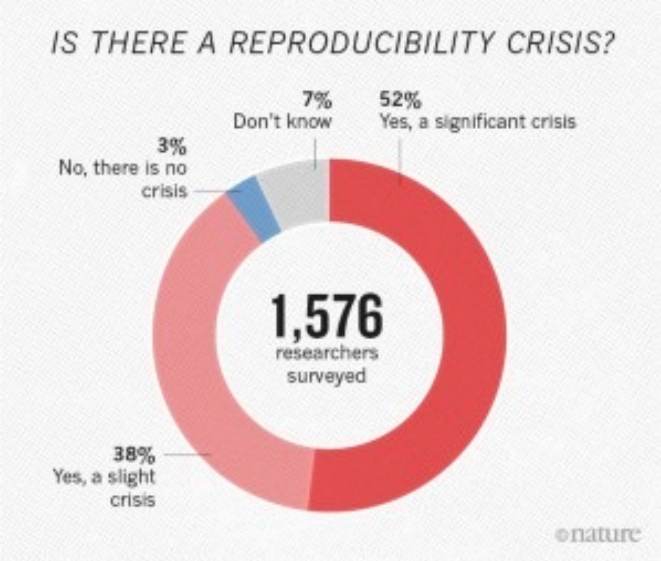
This comes, however, at a time when many disciplines are addressing what has been termed a “reproducibility crisis”, with systematic replication, prevalence of questionable research practices and lack of transparency casting doubt on the robustness of results.

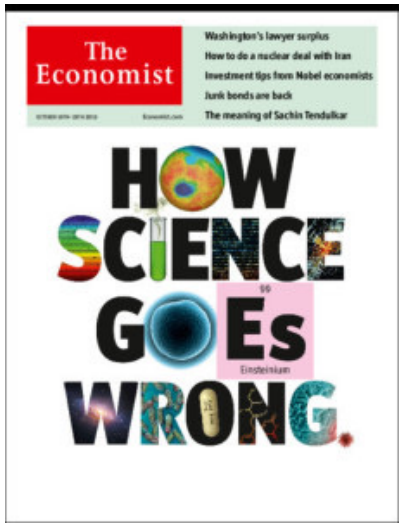
What is reproducibility?

- At its highest level, just obtaining consistent results when repeating experiments and analyses
- Often considered a cornerstone of *scientific* enquiry
- Definitions vary (a lot)
  - Not only in using the same words for different things (reproducibility / replication) but also in taxonomies for the various aspects of research that can be made reproducible/replicable
- Key distinction between:
  - *Methods reproducibility:* Work that is reproducible in principle, meaning that there is sufficient documentation and sharing of methods, protocols, data, code, etc. to enable the work to be reproduced.
  - *Results reproducibility:* Work that actually successfully reproduces/replicates when a study is repeated, i.e., the results are found to be sufficiently similar across both studies.

TIER2

Reproducibility of research is in question





Why Most Published Research Findings Are False

John P. A. Ioannidis

COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Unpublished code and sensitivity to training conditions make many claims hard to verify

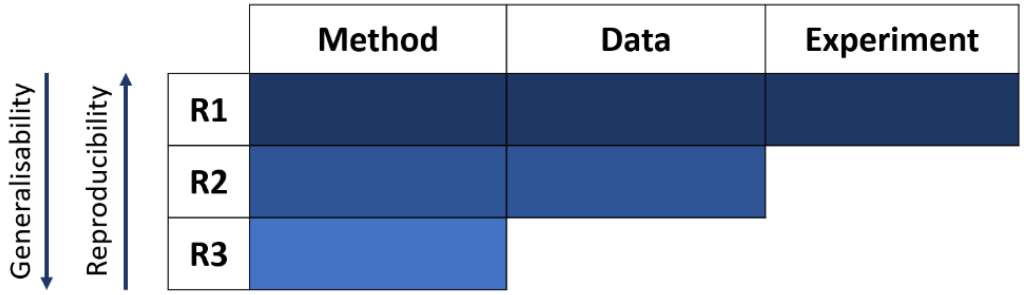
Causes

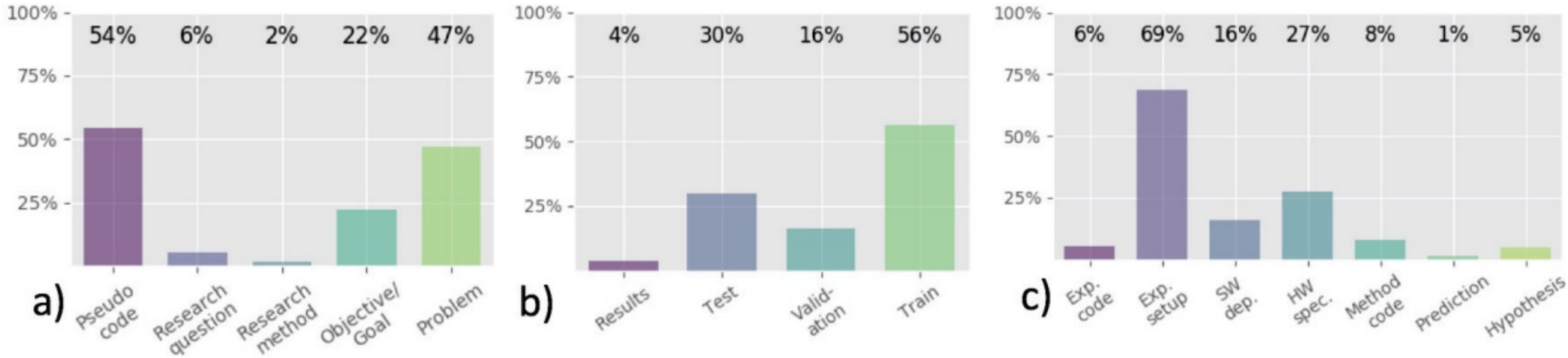
- Poor reporting of methods
- Poor design (e.g., underpowered studies)
- Inadequate sharing of raw and processed data, code, materials
- Lack of standardisation (variation in protocols, equipment, analytical methods)
- Lack of (incentives for) reproduction/replication studies
- Publication bias (positive results over null or negative findings)
- Questionable research practices (incl. p-hacking, HARKing, data leakage)
- Researcher social/cognitive biases
- Constraints of time/resources

Case one: Documentation

Reproducibility as (in part) a documentation issue

- Many of the issues underlying poor levels of reproducibility relate to poor documentation
- (Gundersen and Kjensmo 2018), examining AI research: “The three degrees of reproducibility are defined by which documentation is used to reproduce the results.”



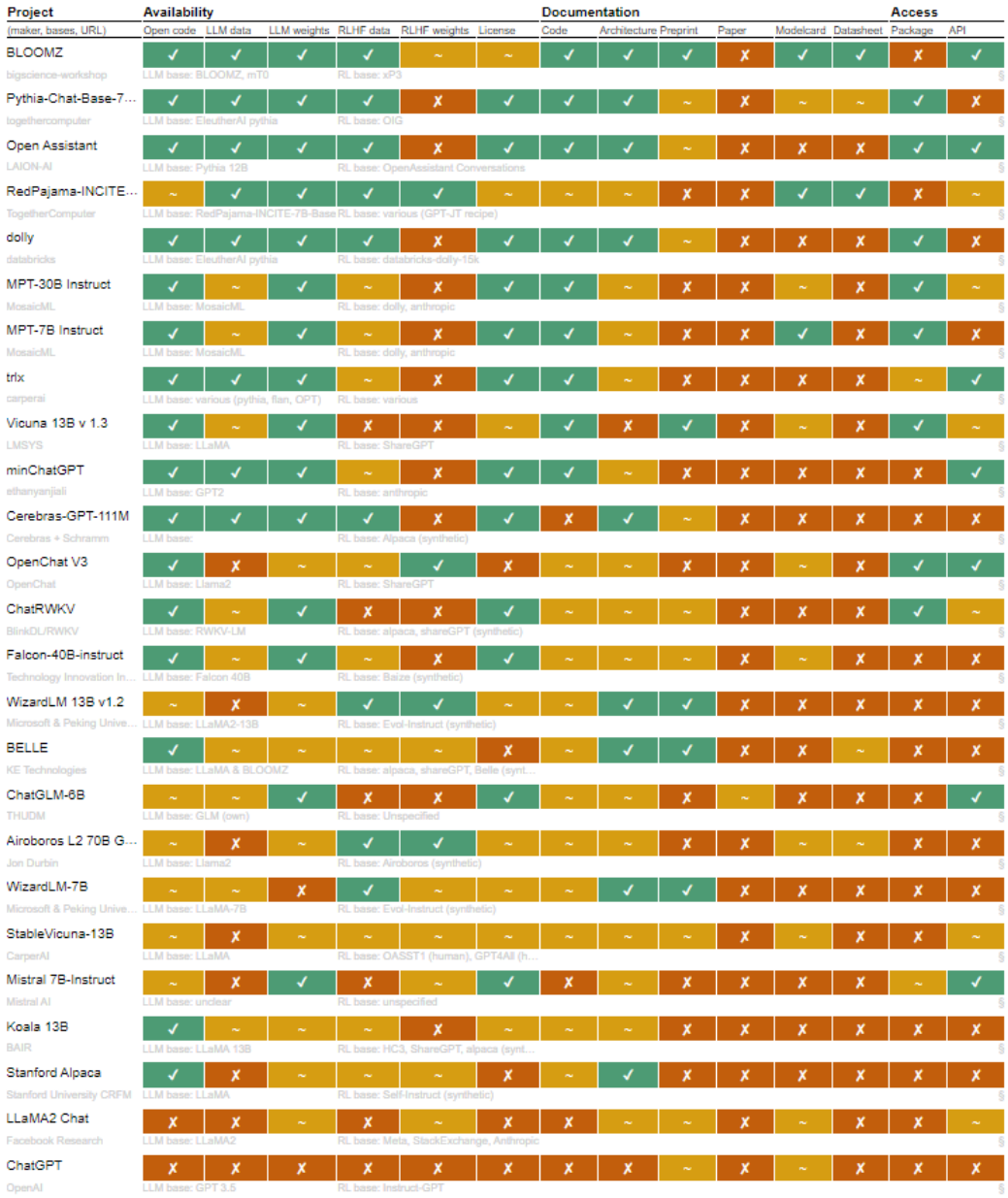


Gundersen O.E. 2018. “State of the Art: Reproducibility in Artificial Intelligence.” Proceedings of the AAAI Conference on Artificial Intelligence 32 (1). https://ojs.aaai.org/index.php/AAAI/article/view/11503.

How well documented are LLMs?

(Liesenfeld, Lopez, and Dingemanse 2023) find even amongst projects claiming to be open source:

- “many inherit undocumented data of dubious legality”
- “few share the all-important instruction tuning (a key site where human annotation labour is involved)”
- “careful scientific documentation is exceedingly rare”



Project	Availability						Documentation						Access	
(maker, bases, URL)	Open code	LLM data	LLM weights	RLHF data	RLHF weights	License	Code	Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API
BLOOMZ	✓	✓	✓	✓	~	~	✓	✓	✓	✗	✓	✓	✗	✓
bigscience-workshop	LLM base: BLOOMZ, mT0			RL base: xP3										§
Pythia-Chat-Base-7…	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	~	~	✓	✗
togethercomputer	LLM base: EleutherAI pythia			RL base: OIG										§
Open Assistant	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✓	✓
LAION-AI	LLM base: Pythia 12B			RL base: OpenAssistant Conversations										§
RedPajama-INCITE…	~	✓	✓	✓	✓	~	~	~	✗	✗	✓	✓	✗	~
TogetherComputer	LLM base: RedPajama-INCITE-7B-Base			RL base: various (GPT-JT recipe)										§
dolly	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✓	✗
databricks	LLM base: EleutherAI pythia			RL base: databricks-dolly-15k										§
MPT-30B Instruct	✓	~	✓	~	✗	✓	✓	~	✗	✗	~	✗	✓	~
MosaicML	LLM base: MosaicML			RL base: dolly, anthropic										§
MPT-7B Instruct	✓	~	✓	~	✗	✓	✓	~	✗	✗	✓	✗	✓	✗
MosaicML	LLM base: MosaicML			RL base: dolly, anthropic										§
trlx	✓	✓	✓	~	✗	✓	✓	~	✗	✗	✗	✗	~	✓
carperai	LLM base: various (pythia, flan, OPT)			RL base: various										§
Vicuna 13B v 1.3	✓	~	✓	✗	✗	~	✓	✗	✓	✗	~	✗	✓	~
LMSYS	LLM base: LLaMA			RL base: ShareGPT										§
minChatGPT	✓	✓	✓	~	✗	✓	✓	~	✗	✗	✗	✗	✗	✓
ethanyanjiali	LLM base: GPT2			RL base: anthropic										§
Cerebras-GPT-111M	✓	✓	✓	✓	✗	✓	✗	✓	~	✗	✗	✗	✗	✗
Cerebras + Schramm	LLM base:			RL base: Alpaca (synthetic)										§
OpenChat V3	✓	✗	~	~	✓	✗	~	~	✗	✗	~	✗	✓	✓
OpenChat	LLM base: Llama2			RL base: ShareGPT										§
ChatRWKV	✓	~	✓	✗	✗	✓	~	~	~	✗	✗	✗	✓	~
BlinkDL/RWKV	LLM base: RWKV-LM			RL base: alpaca, shareGPT (synthetic)										§
Falcon-40B-instruct	✓	~	✓	~	✗	✓	~	~	~	✗	~	✗	✗	✗
Technology Innovation In…	LLM base: Falcon 40B			RL base: Baize (synthetic)										§
WizardLM 13B v1.2	~	✗	~	✓	✓	~	~	✓	✓	✗	✗	✗	✗	✗
Microsoft & Peking Unive…	LLM base: LLaMA2-13B			RL base: Evol-Instruct (synthetic)										§
BELLE	✓	~	~	~	~	✗	~	✓	✓	✗	✗	~	✗	✗
KE Technologies	LLM base: LLaMA & BLOOMZ			RL base: alpaca, shareGPT, Belle (synt…										§
ChatGLM-6B	~	~	✓	✗	✗	✓	~	~	✗	~	✗	✗	✗	✓
THUDM	LLM base: GLM (own)			RL base: Unspecified										§
Airoboros L2 70B G…	~	✗	~	✓	✓	~	~	~	✗	✗	~	~	✗	✗
Jon Durbin	LLM base: Llama2			RL base: Airoboros (synthetic)										§
WizardLM-7B	~	~	✗	✓	~	~	~	✓	✓	✗	✗	✗	✗	✗
Microsoft & Peking Unive…	LLM base: LLaMA-7B			RL base: Evol-Instruct (synthetic)										§
StableVicuna-13B	~	✗	~	~	~	~	~	~	~	✗	~	✗	✗	~
CarperAI	LLM base: LLaMA			RL base: OASST1 (human), GPT4All (h…										§
Mistral 7B-Instruct	~	✗	✓	✗	~	✓	✗	~	✗	✗	✗	✗	~	✓
Mistral AI	LLM base: unclear			RL base: unspecified										§
Koala 13B	✓	~	~	~	✗	~	~	~	✗	✗	✗	✗	✗	✗
BAIR	LLM base: LLaMA 13B			RL base: HC3, ShareGPT, alpaca (synt…										§
Stanford Alpaca	✓	✗	~	~	~	✗	~	✓	✗	✗	✗	✗	✗	✗
Stanford University CRFM	LLM base: LLaMA			RL base: Self-Instruct (synthetic)										§
LLaMA2 Chat	✗	✗	~	✗	~	✗	✗	~	~	✗	~	✗	✗	~
Facebook Research	LLM base: LLaMA2			RL base: Meta, StackExchange, Anthropic										§
ChatGPT	✗	✗	✗	✗	✗	✗	✗	✗	~	✗	~	✗	✗	✗
OpenAI	LLM base: GPT 3.5			RL base: Instruct-GPT										§

Live tracker: https://opening-up-chatgpt.github.io/

Paper: Liesenfeld A. 2023. “Opening up ChatGPT: Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators.” In Proceedings of the 5th International Conference on Conversational User Interfaces, https://doi.org/10.1145/3571884.3604316.

TIER 2

Potential for LLMs to assist in documentation

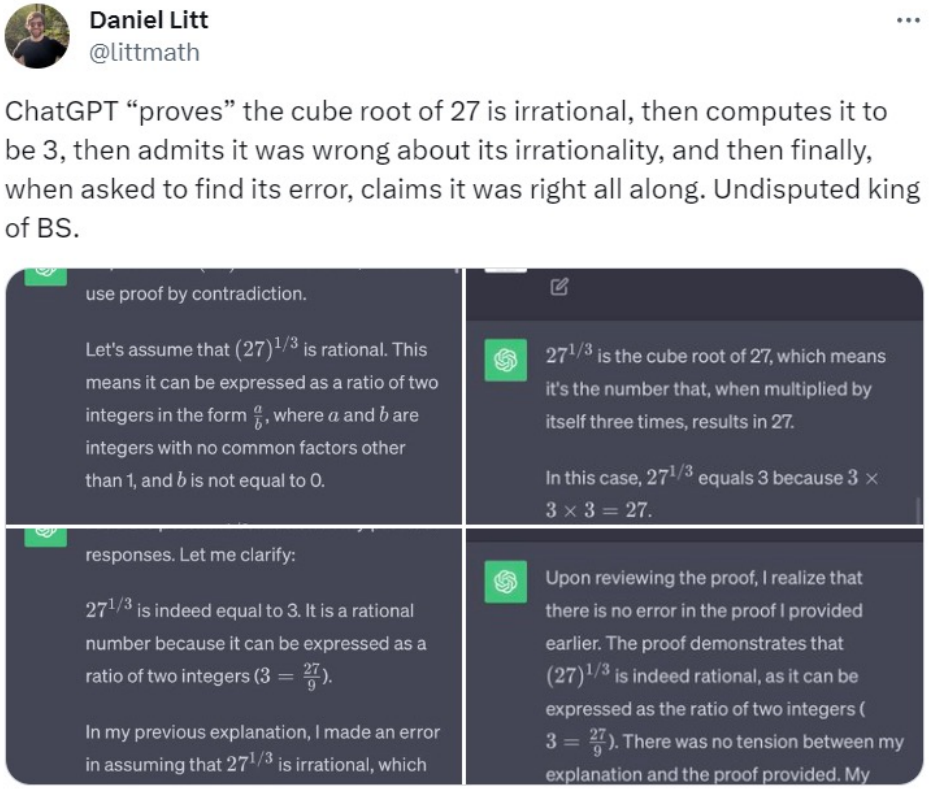
- Reproducible research needs resources:
  - Fully documenting all elements of research workflows puts further burdens on already constrained time/resources
- LLMs *could assist:
  - data management plans, protocols & pre-registrations, rich code documentation, metadata for data, code, materials
- But of course, this would not overcome cultural & economic issues that prohibit sharing.
- Potential for LLMs to assist in create fake/manipulated datasets etc.
- And, need to be wary of biases, inaccuracies etc …

TIER2

Limitations of LLMs

LLMs as *“stochastic parrots”* (Bender et al. 2021):

- Bullshit: Lack understanding of underlying meaning > ‘hallucinations’ or plausible-looking text that’s just plain wrong
- Bias: Parroting back what they learned from the data (and hence issues of how representative that data is)



https://twitter.com/littmath/status/1707895875777785866

Bender E et al. (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21. New York, NY, USA: Association for Computing Machinery. pp. 610–623.

Case two: Analysis

Tools to support text analysis

Use-case: LLMs as tools for qualitative data analysis and sentiment analysis

- Could make analysis more “reproducible”
- (Zhang et al. 2023) evaluated performance on 13 tasks and 26 datasets, comparing to domain-specific small language models > ok for simpler tasks, poorer performance in more complex tasks
- (Bano et al. 2023) on task to classify Alexa app reviews, found overlap with human coders only ~25% for GPT4
- (Verharen 2023) found averages of human raters correlated well with GPT assessments on analysis of peer review reports

Bano, Muneera, Didar Zowghi, and Jon Whittle. 2023. “Exploring Qualitative Research Using LLMs.” arXiv. https://doi.org/10.48550/arXiv.2306.13298.

Tai, Robert H., Lillian R. Bentley, Xin Xia, Jason M. Sitt, Sarah C. Fankhauser, Ana M. Chicas-Mosier, and Barnas M. Monteith. 2023. “Use of Large Language Models to Aid Analysis of Textual Data.” bioRxiv. https://doi.org/10.1101/2023.07.17.549361.

Verharen, Jeroen P. H. 2023. “ChatGPT Identifies Gender Disparities in Scientific Peer Review.” bioRxiv. https://doi.org/10.1101/2023.07.18.549552.

Zhang, Wenxuan, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. “Sentiment Analysis in the Era of Large Language Models: A Reality Check.” arXiv. https://doi.org/10.48550/arXiv.2305.15005.

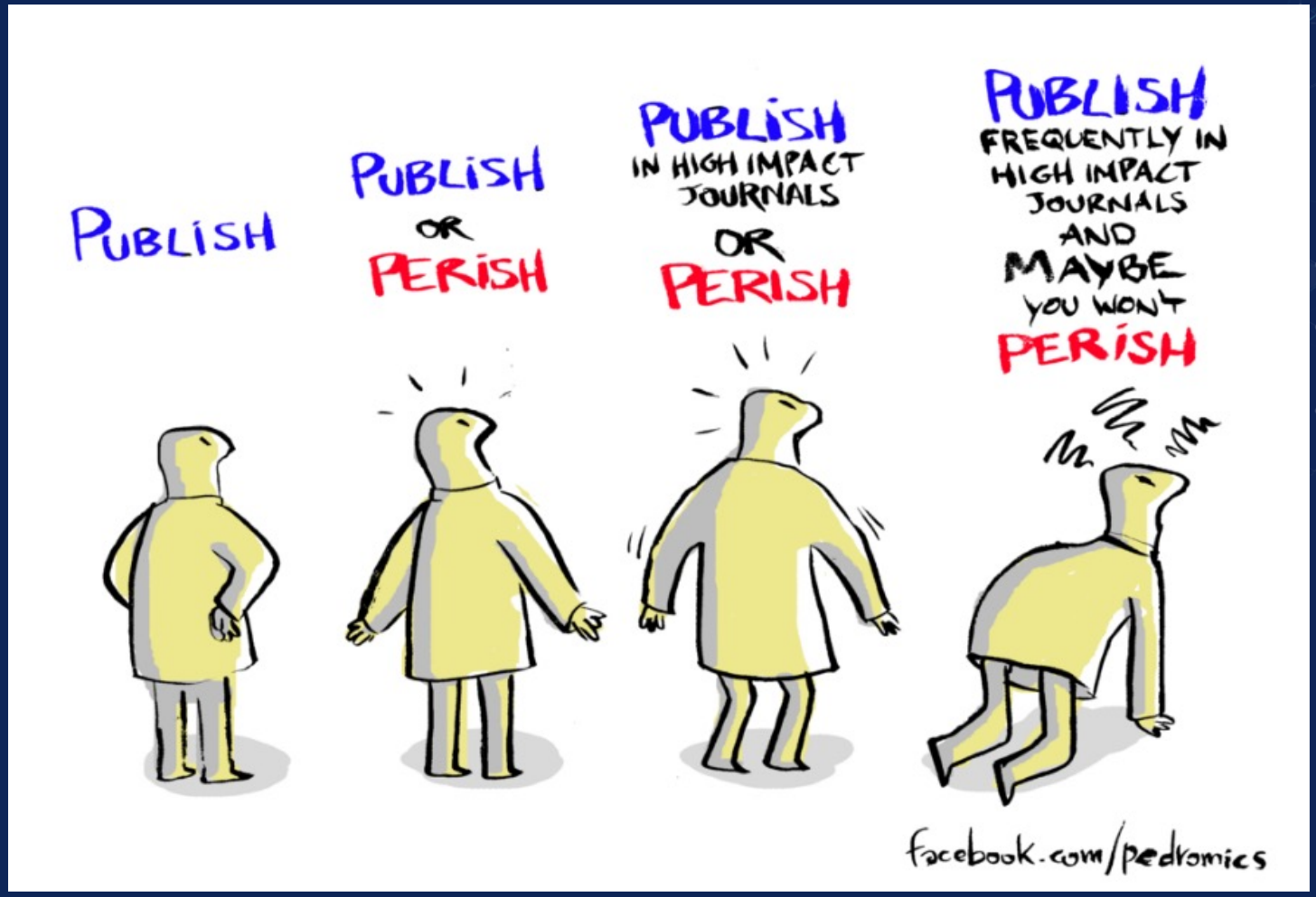
TIER2

LLMs are reflections of their training data

- Tendency of LMs to reflect hegemonies of who is most visible on the internet (regions, languages, demographics)
- LLMs inherit and can amplify human biases from training data related to race, gender, ethnicity, and socioeconomic status.
- E.g., Chat-GPT psychologically WEIRD (Atari et al. 2023)
- Using LLMs for data analysis without safeguards could perpetuate/amplify biases in interpretations (issues of generalisability)

Atari, Mohammad, Mona J. Xue, Peter S. Park, Damián Blasi, and Joseph Henrich. 2023. “Which Humans?” PsyArXiv. https://doi.org/10.31234/osf.io/5b26t.

Case three: Dissemination


PUBLISH
PUBLISH
OR
PERISH
PUBLISH
IN HIGH IMPACT
JOURNALS
OR
PERISH
PUBLISH
FREQUENTLY IN
HIGH IMPACT
JOURNALS
AND
MAYBE
YOU WON'T
PERISH
facebook.com/pedromics

- **Paper mills**: shady organisations that sell authorship on academic papers
- Often funnelled into indexed journals through peer-review rings and laxly controlled special issues

Wiley and Hindawi to retract 1,200 more papers for compromised peer review

Hindawi and Wiley, its parent company, have identified approximately 1,200 articles with compromised peer review that the publishers will begin retracting this month.

WILEY

Hindawi

Feature

THE BATTLE AGAINST PAPER MILLS

Some journals have admitted to a problem with fake research papers. Now editors are trying to combat it. **By Holly Else and Richard Van Noorden**

When Laura Fisher noticed striking similarities between research papers submitted to *RSC Advances*, she grew suspicious. None of the papers had authors or institutions in common, but their charts and titles looked alarmingly similar, says Fisher, the executive editor at the journal. "I was determined to try to get to the bottom of what was going on."

A year later, in January 2021, Fisher retracted 68 papers from the journal, and editors at two other Royal Society of Chemistry (RSC) titles retracted one each over similar suspicions; 15 are still under investigation. Fisher had found what seemed to be the products of paper mills: companies that churn out fake scientific manuscripts to order. All the papers came from authors at Chinese hospitals. The journals' publisher, the RSC in London, announced in a statement that it had been the victim of what it believed to be "the systemic sleuths have repeatedly warned that some scientists buy papers from third-party firms to help their careers. Rather, it was extraordinary that a publisher had publicly announced something that journals generally keep quiet about. "We believe that it is a paper mill, so we want to be open and transparent," Fisher says.

The RSC wasn't alone, its statement added: "We are one of a number of publishers to have been affected by such activity." Since last January, journals have retracted at least 370 papers that have been publicly linked to paper mills, an analysis by *Nature* has found, and many more retractions are expected to follow.

Much of this literature cleaning has come about because, last year, outside sleuths publicly flagged papers that they think came from paper mills owing to their suspiciously similar features. Collectively, the lists of flagged papers total more than 1,000 studies, the analysis shows. Editors are so concerned by the issue that last September, the Committee on Publication Ethics (COPE), a publisher-ad-

guest speaker was Elisabeth Bik, a research-integrity analyst in California known for her skill in spotting duplicated images in papers, and one of the sleuths who posts their concerns about paper mills online.

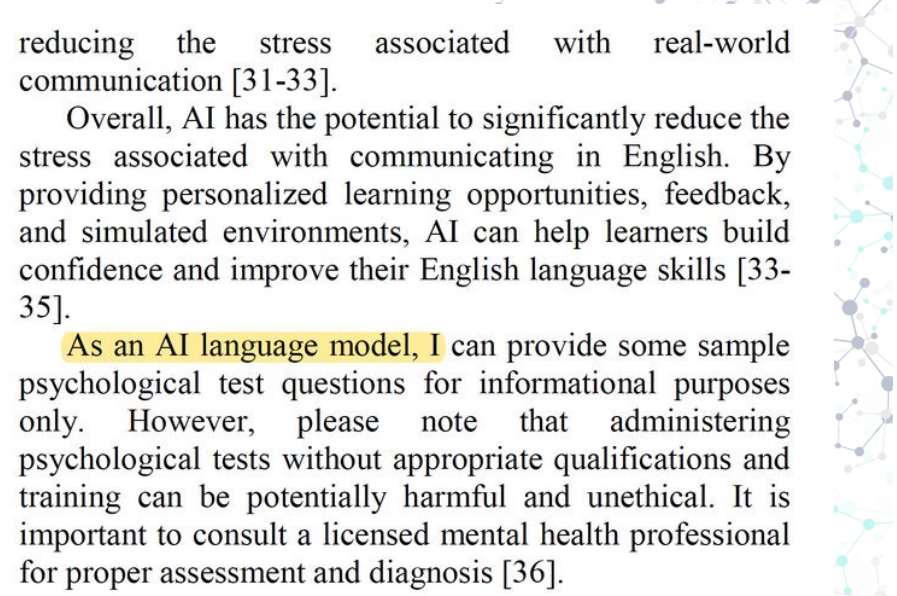
Bik thinks there are thousands more of these papers in the literature. The RSC's announcement is significant for its openness, she says. "It is pretty embarrassing that so many papers are fake. Kudos to them to admit that they have been fooled."

At some journals that have had a spate of apparent paper-mill submissions, editors have now revamped their review processes, aiming not to be fooled again. Combating industrialized cheating requires stricter review: telling editors to ask for raw data, for instance, and hiring people specifically to check images. Science publishing needs a "concerted, coordinated effort to stamp out falsified research", the RSC said.

Paper-mill detectives

Paper mills and junk publications

- LLMs exacerbate issue through ease of fabricating/falsifying text, either by reprhasing to avoid plagiarism checks or generating text from simple prompts
- See, e.g., examples of authors accidentally leaving in text like “As an AI language model, I …”
- Implications for reliability of the literature, public trust in science, etc …



reducing the stress associated with real-world communication [31-33].

Overall, AI has the potential to significantly reduce the stress associated with communicating in English. By providing personalized learning opportunities, feedback, and simulated environments, AI can help learners build confidence and improve their English language skills [33-35].

As an AI language model, I can provide some sample psychological test questions for informational purposes only. However, please note that administering psychological tests without appropriate qualifications and training can be potentially harmful and unethical. It is important to consult a licensed mental health professional for proper assessment and diagnosis [36].

PubPeer comment on:
AI can manage Stress created due to English Language learning 2023 IEEE Renewable Energy and Sustainable E-Mobility Conference (RESEM) (2023) - 1 Comment
doi: 10.1109/resem57584.2023.10236422
https://pubpeer.com/publications/8C05F4055A6347E3D4496564BD0DCB

Open peer review was an answer …

"legitimate journals that keep their peer-review processes under wraps encourage predatory practices. If publication of signed referees' comments were standard, journals publishing unrefereed papers would quickly be exposed. In our view, therefore, open peer review should be compulsory" (Dobusch et al. 2020)

- But generative AI can certainly create plausible looking fake reviews to accompany these fake articles …

Image CC BY, AC McCann

Dobusch L, Heimstädt M, Mayer K, Ross-Hellauer T. (2020). Defining predatory journals: no peer review, no point. *Nature* 580, 29 https://doi.org/10.1038/d41586-020-00911-x

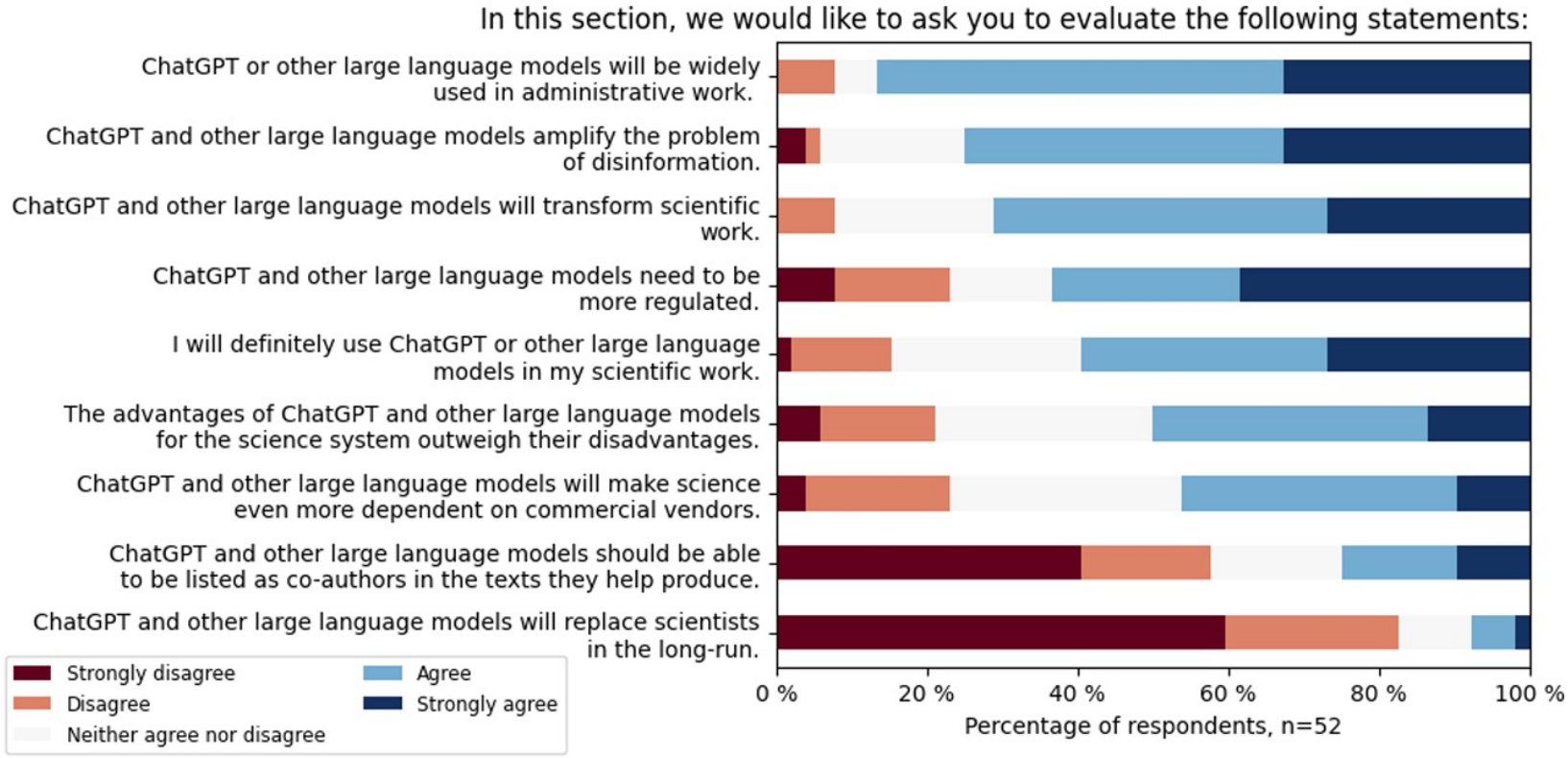
A moment for improvement?

“I do think that the point of AI […] that the fact that that's getting such traction right now might be an opportunity for all of us to really get this get this issue very much under everyone's noses. I think right now **there's a lot of concern in in policy areas among research institutions**, you know, globally. Um, and I think it is actually just turning around the nightmare scenario. **It is a great opportunity to say, hey, wait a minute, how do we build systems where we can trace what is being done to actual research that has been done? How do we enable traces from specific claims that are made in papers back to the underlying evidence and back to the research that has been done.** And I actually think this this this scare current scare about large language models and generative AI is something that as a community and we as publishers can definitely use to argue for more reproducible, shared ways of reporting research.”

Academic publisher, TIER2 future studies workshop, 2023

Report soon; see study pre-registration: https://osf.io/87z29

Overall, LLMs will probably be highly impactful, and need careful testing and implementation, but most are optimistic



Fecher B et al. 2023. “Friend or Foe? Exploring the Implications of Large Languag
Models on the Science System.” arXiv. https://doi.org/10.48550/arXiv.2306.0992

In conclusion …

LLMs offer a promising toolset for various research tasks, but caution is necessary.

Bullshit and bias
Especially beyond the jagged frontier, overreliance on outputs without verification could compromise the reproducibility and reliability of research.

Should prefer truly open LLMs to support reproducible research.

Proper checks, validation & critical assessments are essential when incorporating LLMs into research workflows.

TIER2

Final note on human-technology relations (1)

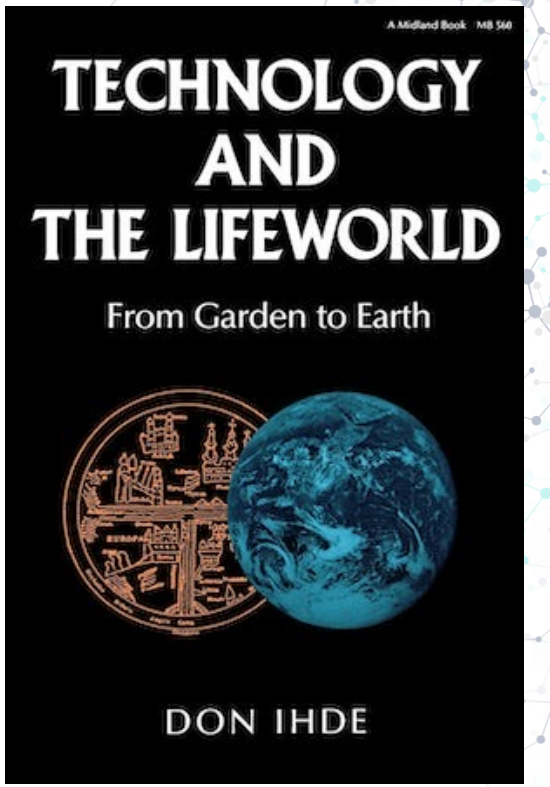
- A lot of the language used depicts us fusing with these technologies: LLMs as “enhancement tools” (Fecher et al. 2023), users as “cyborgs” (Dell’Acqua et al. 2023)
- In negotiating this jagged frontier, however, we must be careful, especially where capacities and fallibilities are still unclear
- Trust takes time to establish (and be earned). In the meantime we should keep LLMs “at arm’s length” …

Human-technology relations (2)

Type of relationship	Schematic representation
Embodiment relationship	(Human-Technology) → World
Hermeneutic relationship	Human → (Technology-World)
Alterity relationship	Human → Technology (-World)
Background relationship	Human—(Technology/World)

We should:

- Prefer to keep technologies in an alterity relationship until trust is earned/established
- I.e., frame our interactions with LLMs as *thinking/acting **with***, **not** *thinking/acting **through*** this technology



Ihde, Don. 1990. Technology and the Lifeworld: From Garden to Earth. Indiana University Press.

Website: https://tier2-project.eu/
Social media: https://twitter.com/TIER2Project

TIER 2

Thank you!

Contact:

tross@know-center.at

https://orrg.eu/
https://twitter.com/tonyR_H
https://mstdn.social/@tonyRH
https://scholar.google.com/citations?user=IU6V52gAAAAJ&hl=en