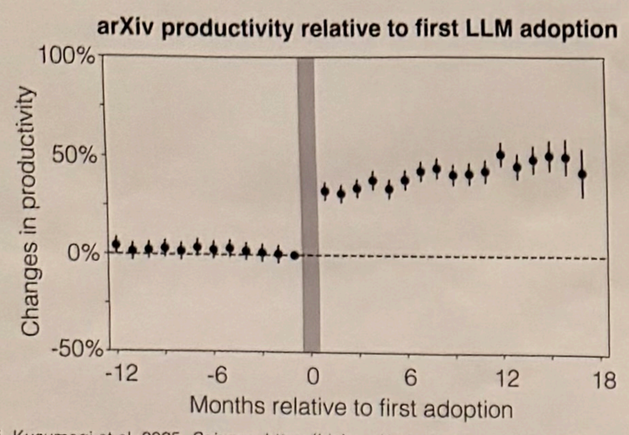
Agentic AI Challenges Science of Science to Strengthen Its Foundations

Vincent Holst[1], Andres Algaba[1], Sylvia Wenmackers[2], Vincent Ginis[1,3]

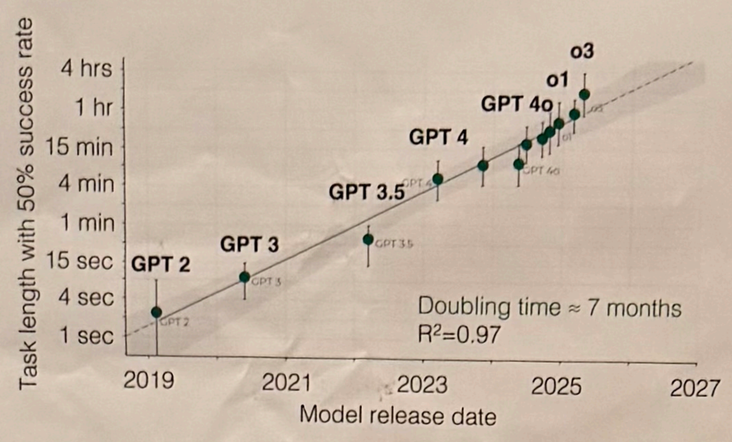
[1] Vrije Universiteit Brussel; [2] KU Leuven; [3] Harvard University

Large Language Models and Agentic AI Are Changing Science

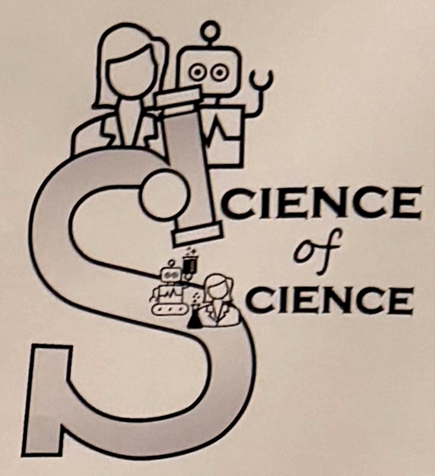
LLMs are reshaping how papers are written[1]



Agentic systems can increasingly solve long-horizon tasks[2]



Science of Science (*SciSci*) is in a peculiar position

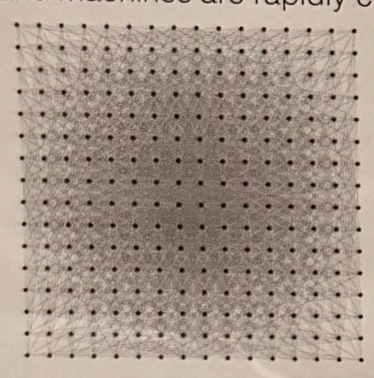


1. Kusumegi et al. 2025, *Science*, https://doi.org/10.1126/science.adw3000
2. Kwa et al. 2025, *NeurIPS 2025*, https://doi.org/10.48550/arXiv.2503.14499

The Developments Raise New Research Questions for Science of Science

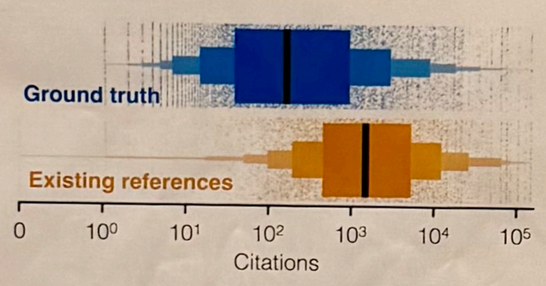
Organization of science

Collaboration and division of labor between humans and machines are rapidly changing[3]

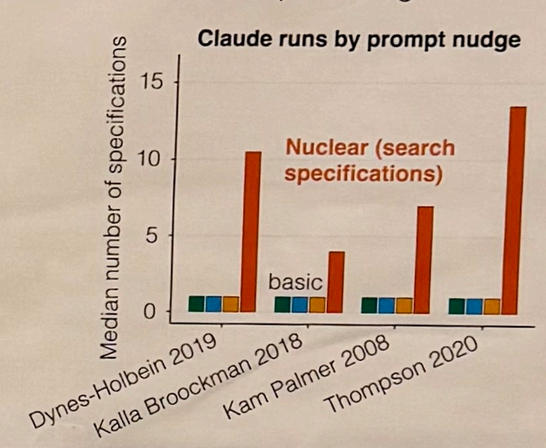


Methods used to produce scientific knowledge

LLMs may amplify heigthened citation bias[4]



Sycophantic tendencies in agentic systems may increase the risk of p-hacking[5]



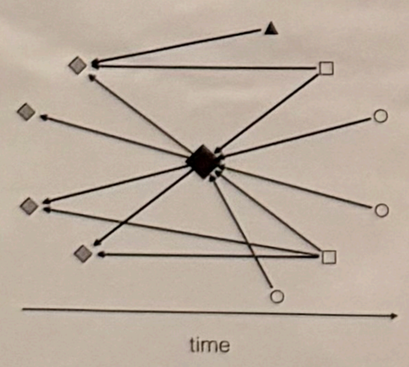
3. OpenAI 2026, "An OpenAI model has disproved a central conjecture in discrete geometry."
4. Algaba et al. 2025, *Findings of ACL: NAACL*, https://doi.org/10.18653/v1/2025.findings-naacl.381
5. Asher et al. 2026, "Do Claude Code and Codex P-Hack?", https://www.andrewbenjaminhall.com/pdfs/asher_et_al_LLM_sycophancy.pdf

Science of Science Is Vulnerable to the Risks Because It Is Under-Theorized

One branch of *SciSci* depends on large-scale analysis of citations

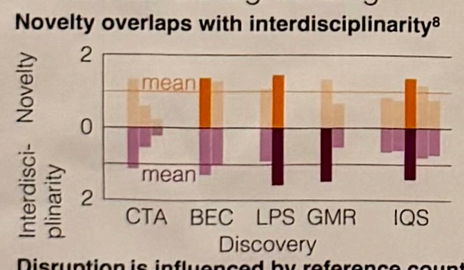
Novelty[6] — Atypical combinations of journal pairs in ◆

Disruption[7] — $\frac{\#\circ - \#\square}{\#\circ + \#\square + \#\blacktriangle}$

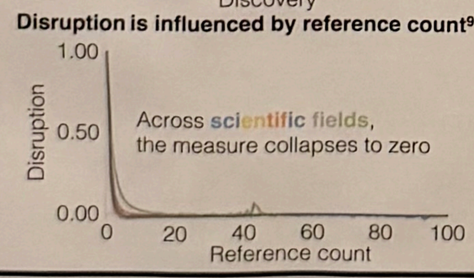


Yet key metrics were popularized without sufficient grounding

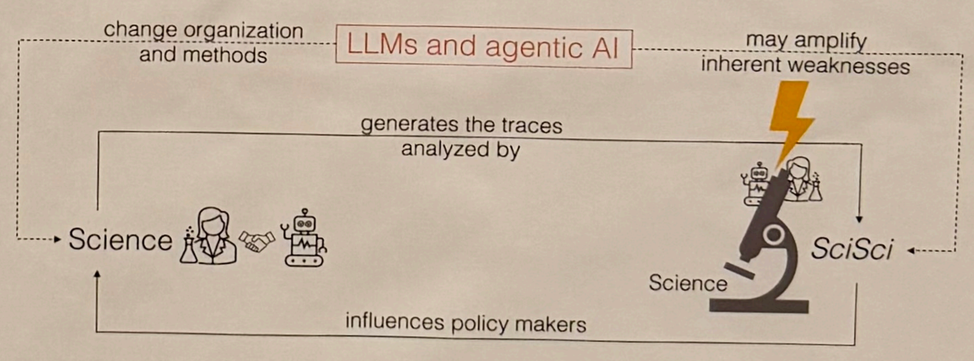
Novelty overlaps with interdisciplinarity[8]



Disruption is influenced by reference count[9]



LLMs and agentic AI make it easy to generate hypotheses, run many specifications, and shape post-hoc narratives



6. Uzzi et al. 2013, *Science*, https://doi.org/10.1126/science.1240474
7. Funk & Owen-Smith 2017, *Management Science*, https://doi.org/10.1287/mnsc.2015.2366
8. Fontana et al. 2020, *Research Policy*, https://doi.org/10.1016/j.respol.2020.104063,
9. Ruan et al. 2021, *Technological Forecasting & Social Change*, https://doi.org/10.1016/j.techfore.2021.121071

How Can We Strengthen Science of Science?

Theoretical foundations

Develop new metrics from first principles to ensure
- Conceptual grounding
- Stability of interpretation
- Robustness

Methodological safeguards

CERN for *SciSci*
- Blinded analysis
- Strict significance thresholds
- Independent replication

Caveat: controlled data access

Study of automation in science

Validity standards for AI-mediated scientific traces
- Productivity signals
- References
- Reviews

Be clear abt the EXPERIMENTAL SETTING
remove studied variables from experiment tools

human oversight & evaluate LLM & Agent's outcomes

Quantify Vulnerability in a small case study

Derive new metrics for scientific progress with aligned incentives (robust to shenanigans)

Be a domain expert - understand causal and non relationships in your data and account for them. Also use statistically principled & appropriate instruments

Learn from history, more responsibility

~~For~~ In the age of AI, we should think of building a scientific AI system that can actually "challenge" your idea.

Benchmark AI "Skills"

Concern: Homogenization of LLMs, losing human creativity
Idea: Similar to day 1 talk on risk, maybe reward high-risk, unique ideas?

Quality of Research
- checks for plagiarism.
- how core users engaged with articles themselves?
- reproducibility of data & study

Be careful with theory - it's misled medical research

Add your thoughts here!

DATA ANALYTICS LABORATORY RESEARCH GROUP

Check out our research in *SciSci* here. The poster with all post-its will be uploaded on this page.

Contact
vincent.thorge.holst@vub.be

Derive new metrics
for scientific
progress with
aligned incentives
(robust to shakegizing)

Ba

Be careful
with theory
- it's misled
medical
research

Be a domain expert -
understand causal
and noise
relationships in your
data and account for them.

Also, use statistically
principled & appropriate
instruments.

Be clear about the EXPERIMENTAL SETTING

---

remove studied variables from experiment tools

Learn from
history,
more
responsibility

In the era of AI,
we should think of building
a scientific AI system
that can actually "challenge"
your idea.

Benchmark

Ai "Skills"

& evaluation

human oversight on

LLM & Agent's outcomes

Concern:

Homogenization of LLMs, losing human creativity

Idea:

Similar to day 1 talk on risk, maybe reward high-risk, unique ideas?

Quality of Research

- checks for plagirism.
- how are users engaged with articles themselves?
- reproducibility of data & study

Quantify
vulnerability
in a small
case study