Large Language Models Amplify the Matthew Effect in Scientific Research

Andres Algaba[1], **Vincent Holst[1]**, Floriano Tori[1],
Melika Mobini[1], Brecht Verbeken[1],
Sylvia Wenmackers[2] and Vincent Ginis[1,3]

Contact:
vincent.thorge.holst@vub.be

1 Data Analytics Laboratory, Vrije Universiteit Brussel, Belgium
2 Centre for Logic and Philosophy of Science (CLPS), KU Leuven, Belgium
3 School of Engineering and Applied Sciences, Harvard University, USA

Motivation

- Human citation behavior (~training data) has some well-documented biases (Letchford et al., 2015; Price, 1976; Wang, 2014; Wuchtly et al., 2007)

Letchford, A., Moat, H. S., & Preis, T. (2015). The advantage of short paper titles. Royal Society open science, 2(8), 150266.

Ng, J. Y., Maduranayagam, S. G., Suthakar, N., Li, A., Lokker, C., Iorio, A., Haynes, R. B., & Moher, D. (2025). Attitudes and perceptions of medical researchers towards the use of artificial intelligence chatbots in the scientific process: An international cross-sectional survey. The Lancet Digital Health, 7(1), e94–e102.

Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American society for Information science, 27(5), 292-306.

Wang, J. (2014). Unpacking the Matthew effect in citations. Journal of Informetrics, 8(2), 329-339.

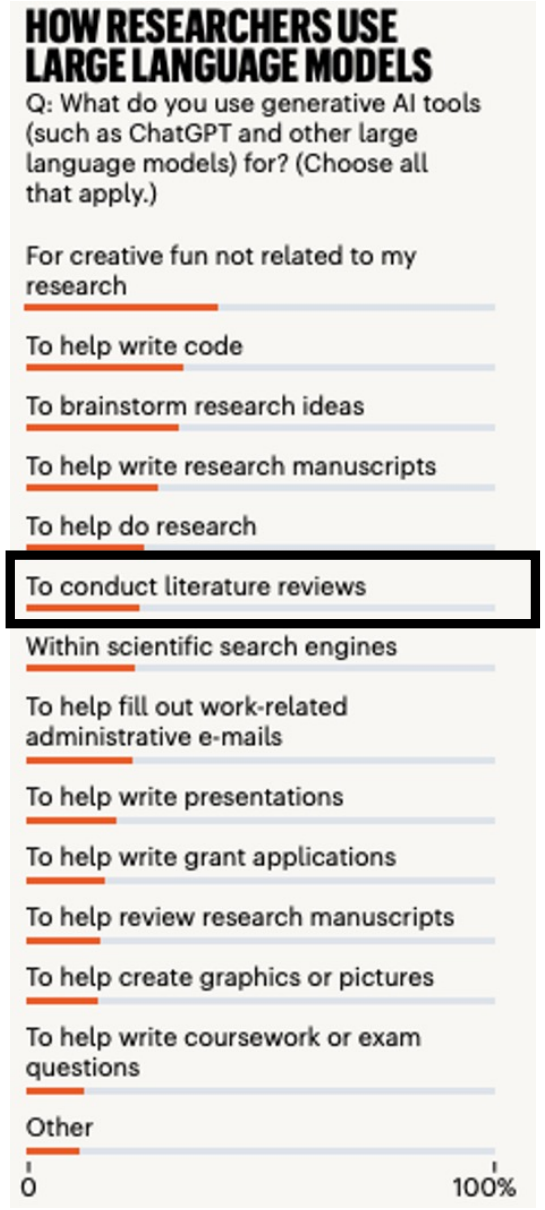
Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. Science, 316(5827), 1036-1039.

https://www.nature.com/articles/d41586-023-02980-0

Motivation

- Human citation behavior (~training data) has some well-documented biases (Letchford et al., 2015; Price, 1976; Wang, 2014; Wuchtly et al., 2007)
- Researchers are using LLMs in literature reviews (Ng et al., 2025), although a lot is still unknown about "real-world" usage

A Nature survey in September 2023



Letchford, A., Moat, H. S., & Preis, T. (2015). The advantage of short paper titles. Royal Society open science, 2(8), 150266.

Ng, J. Y., Maduranayagam, S. G., Suthakar, N., Li, A., Lokker, C., Iorio, A., Haynes, R. B., & Moher, D. (2025). Attitudes and perceptions of medical researchers towards the use of artificial intelligence chatbots in the scientific process: An international cross-sectional survey. The Lancet Digital Health, 7(1), e94–e102.

Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American society for Information science, 27(5), 292-306.

Wang, J. (2014). Unpacking the Matthew effect in citations. Journal of Informetrics, 8(2), 329-339.

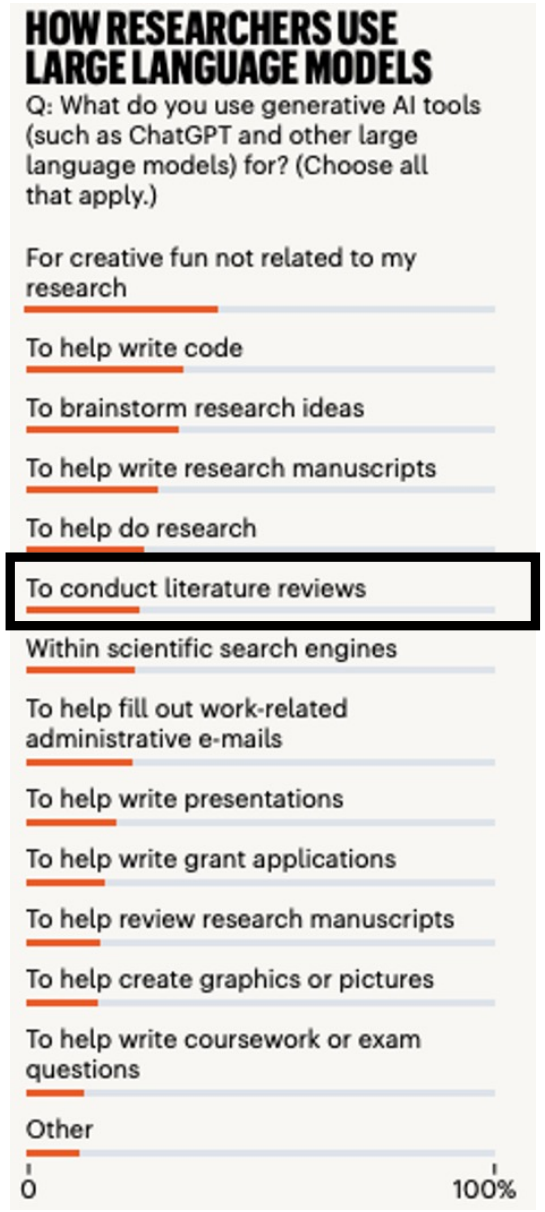
Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. Science, 316(5827), 1036-1039.

https://www.nature.com/articles/d41586-023-02980-0

Motivation

- Human citation behavior (~training data) has some well-documented biases (Letchford et al., 2015; Price, 1976; Wang, 2014; Wuchtly et al., 2007)
- Researchers are using LLMs in literature reviews (Ng et al., 2025), although a lot is still unknown about "real-world" usage
- Shortage of experiments that focus on a "controlled laboratory" setting (~parametric knowledge)

A Nature survey in September 2023



Letchford, A., Moat, H. S., & Preis, T. (2015). The advantage of short paper titles. Royal Society open science, 2(8), 150266.

Ng, J. Y., Maduranayagam, S. G., Suthakar, N., Li, A., Lokker, C., Iorio, A., Haynes, R. B., & Moher, D. (2025). Attitudes and perceptions of medical researchers towards the use of artificial intelligence chatbots in the scientific process: An international cross-sectional survey. The Lancet Digital Health, 7(1), e94–e102.

Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American society for Information science, 27(5), 292-306.

Wang, J. (2014). Unpacking the Matthew effect in citations. Journal of Informetrics, 8(2), 329-339.

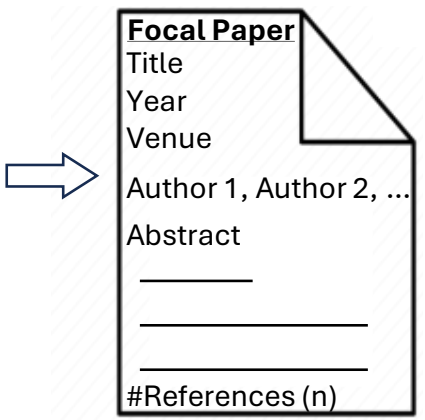
Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. Science, 316(5827), 1036-1039.

https://www.nature.com/articles/d41586-023-02980-0

Prompt GPT-4o to Generate Reference Suggestions

SciSciNet

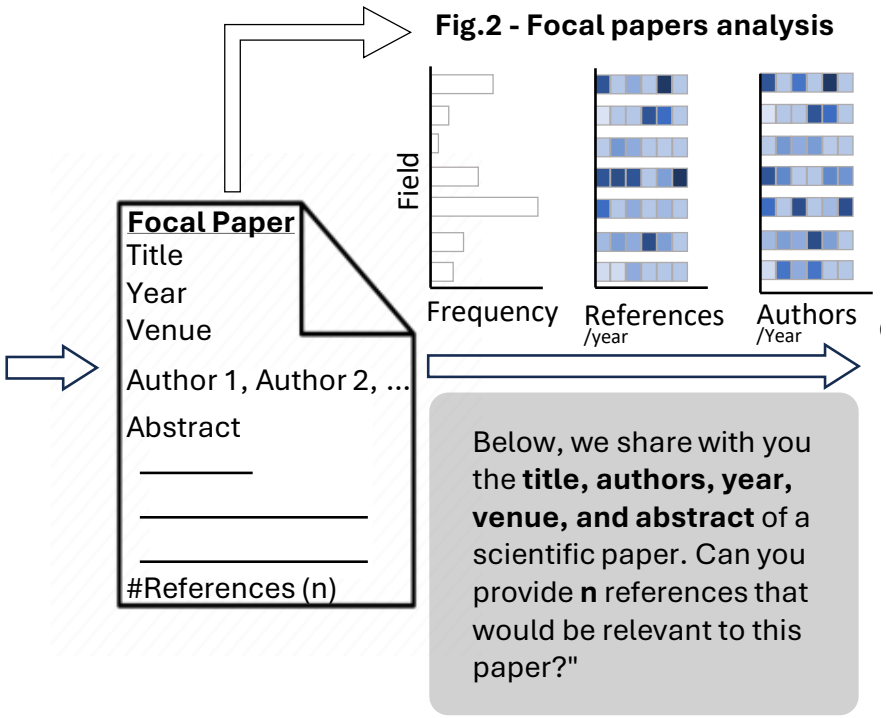
- Q1 Journal
- 1999 -2021
- 3 < #references < 54
- #citations > 1



Prompt GPT-4o to Generate Reference Suggestions

SciSciNet

- Q1 Journal
- 1999 -2021
- 3 < #references < 54
- #citations > 1



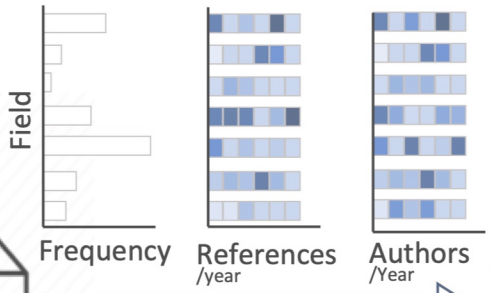
Prompt GPT-4o to Generate Reference Suggestions

SciSciNet

- Q1 Journal
- 1999 -2021
- 3 < #references < 54
- #citations > 1

Focal Paper
Title
Year
Venue
Author 1, Author 2, ...
Abstract
#References (n)

Fig.2 - Focal papers analysis



Below, we share with you the **title, authors, year, venue, and abstract** of a scientific paper. Can you provide **n** references that would be relevant to this paper?"

Generated Reference List

Fig.3 – Reinforcing Matthew

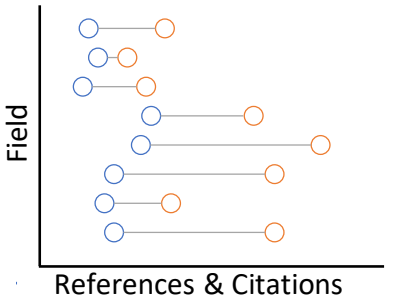


Fig.4 – Time, Author & Title Bias

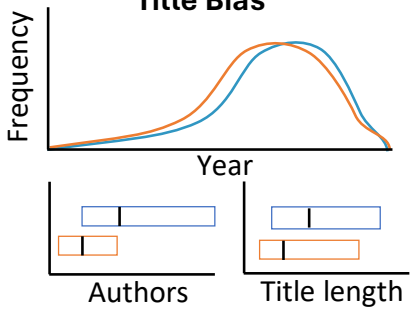


Fig.5 – Textual embedding

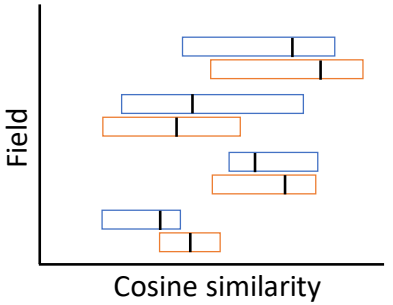
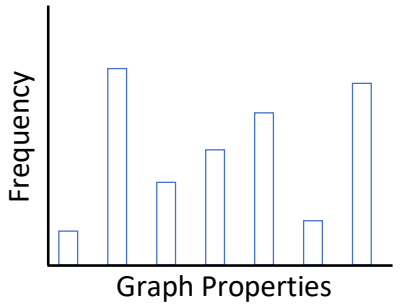
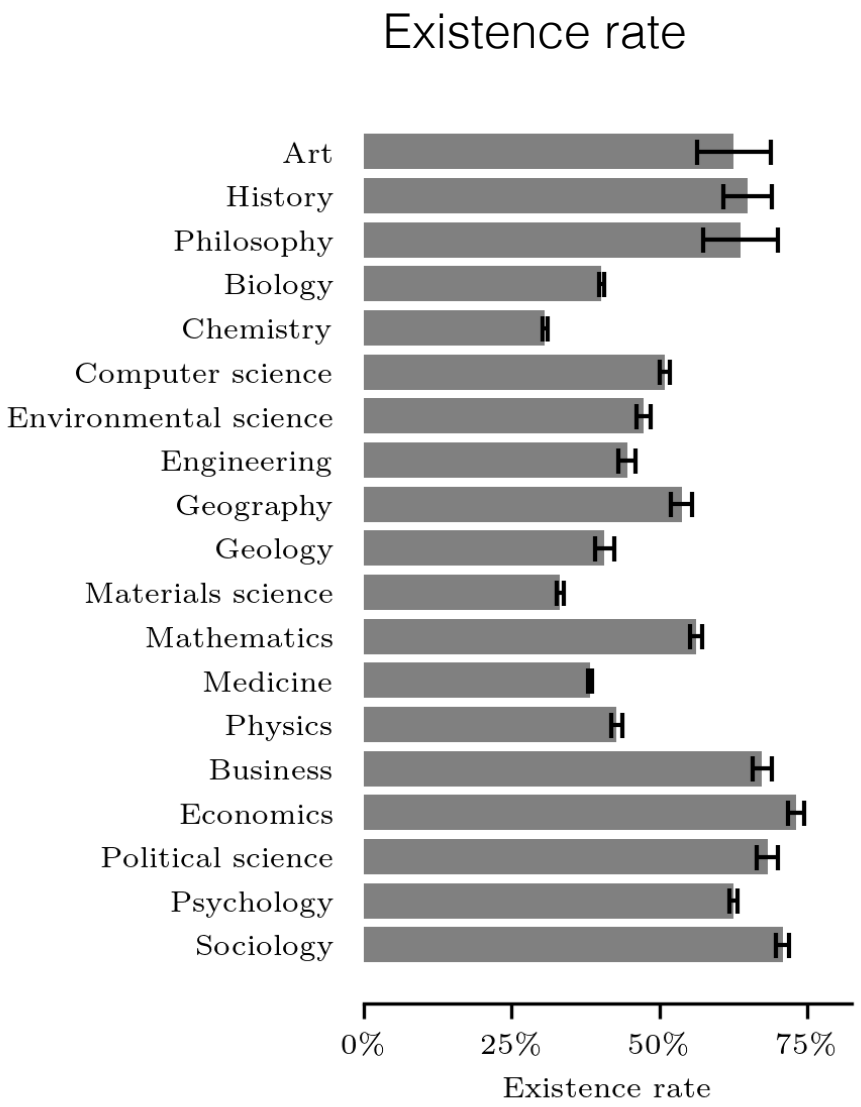


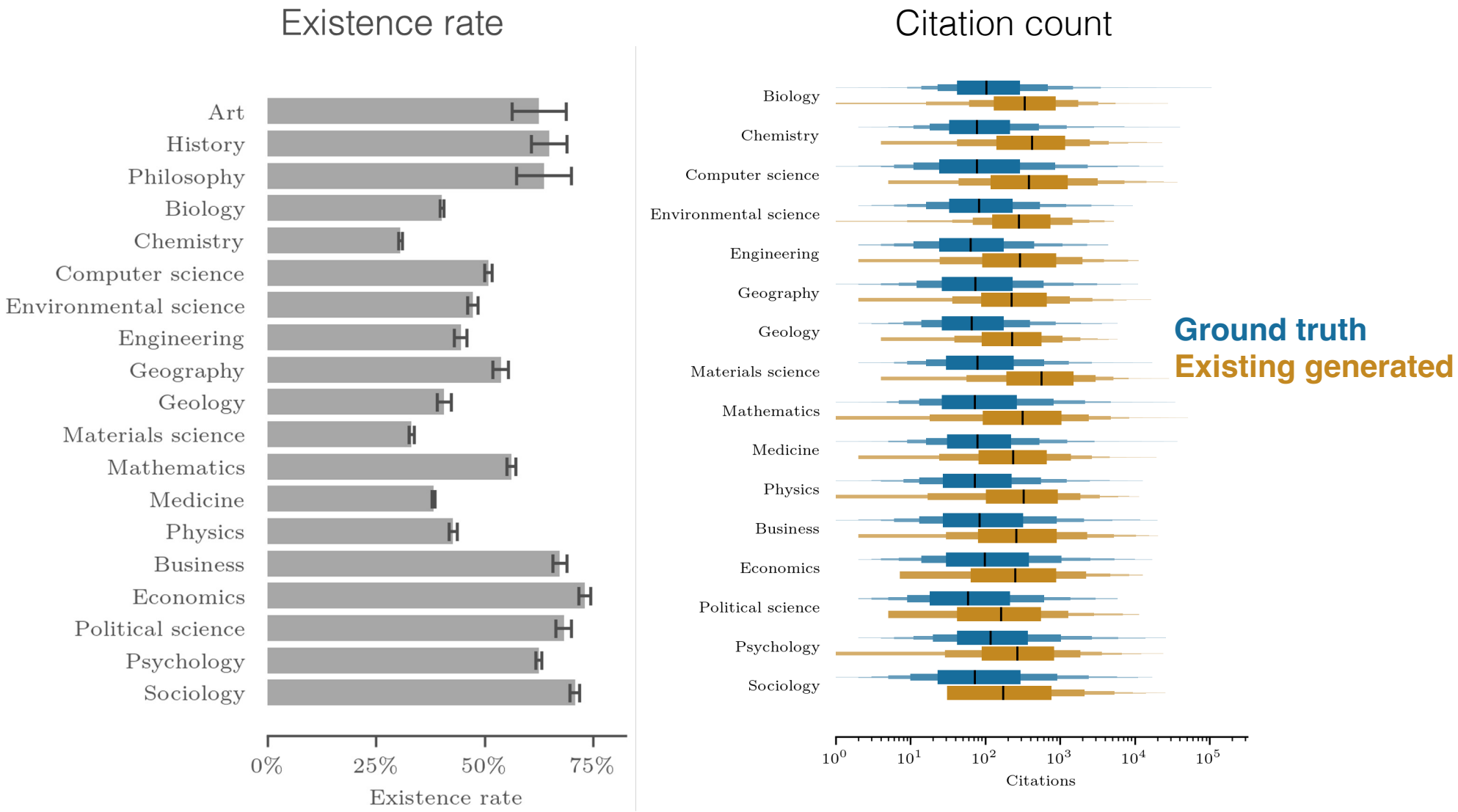
Fig.6 – Network Properties



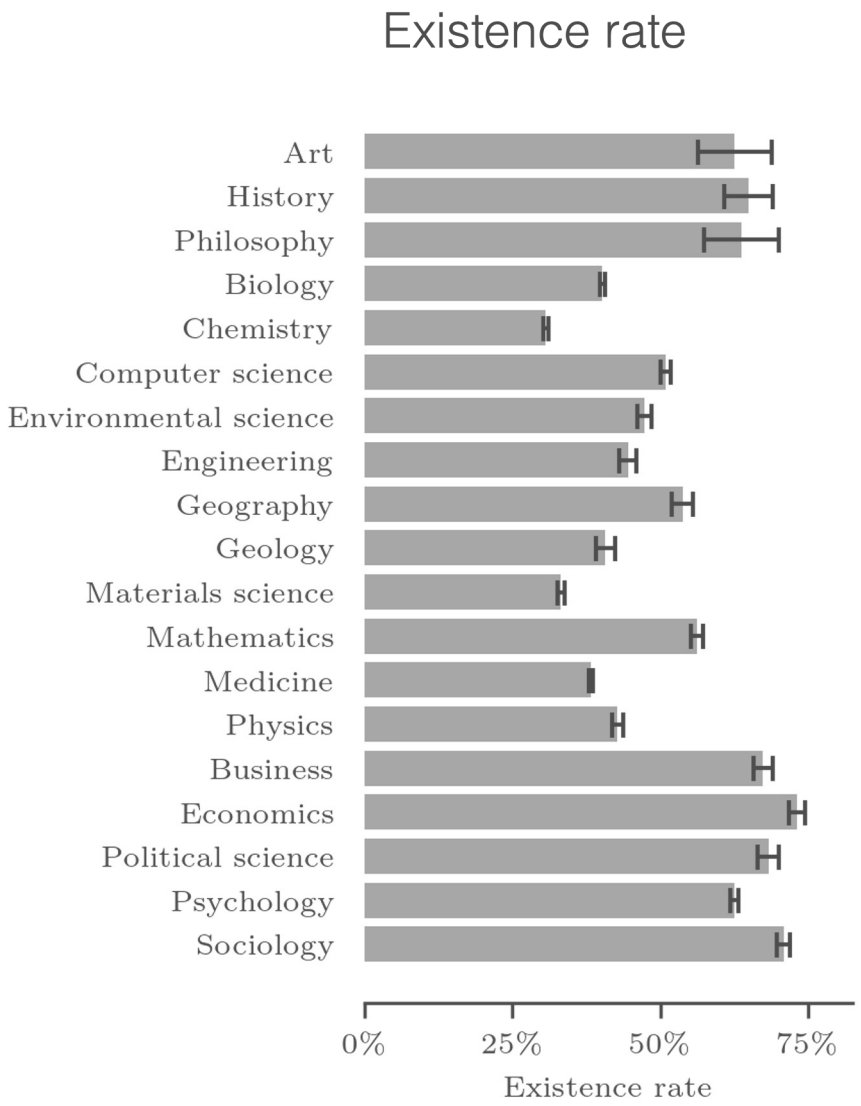
Generated References Exhibit a Heightened Citation Bias

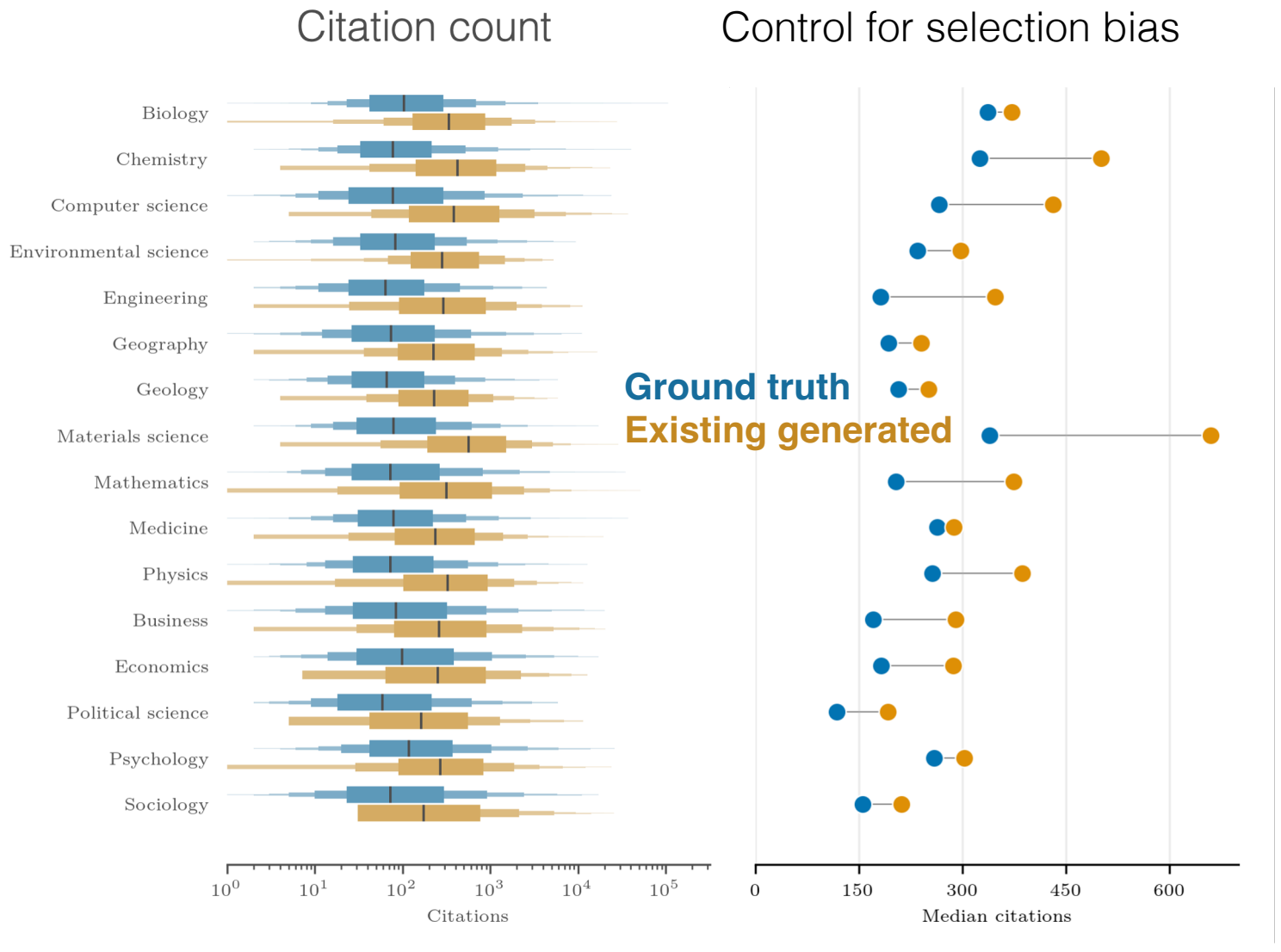


Generated References Exhibit a Heightened Citation Bias

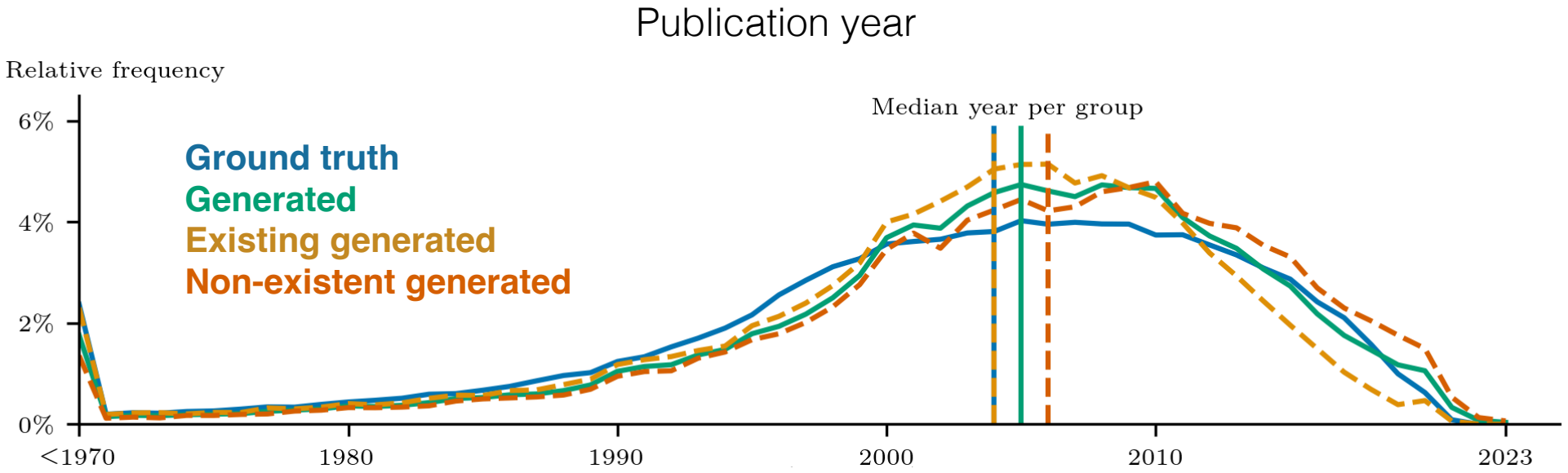


Generated References Exhibit a Heightened Citation Bias

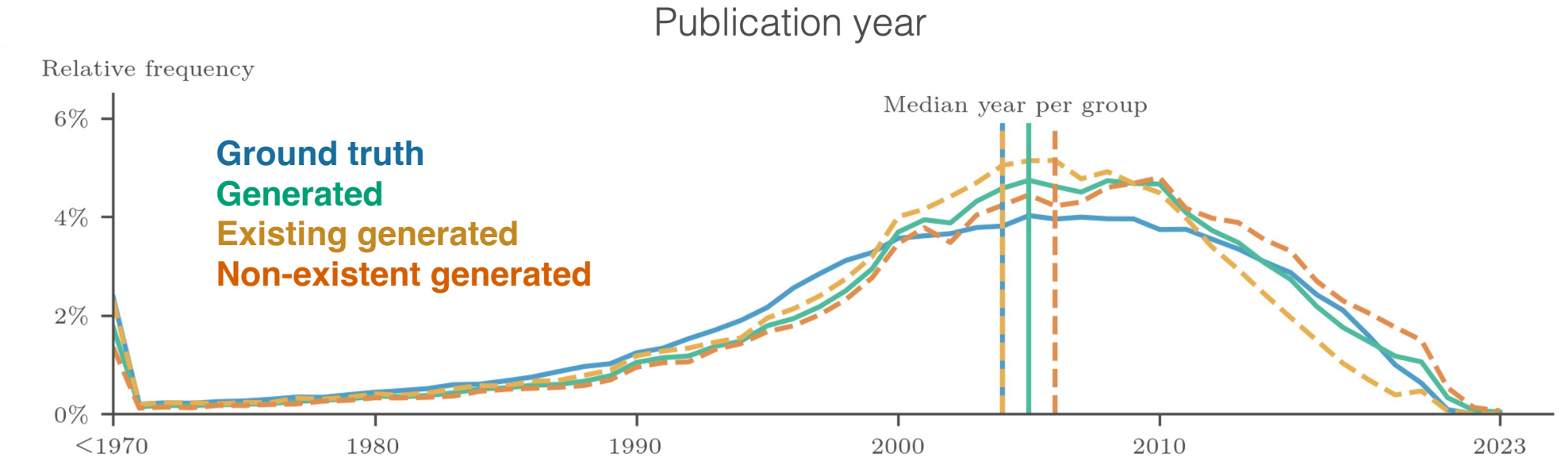


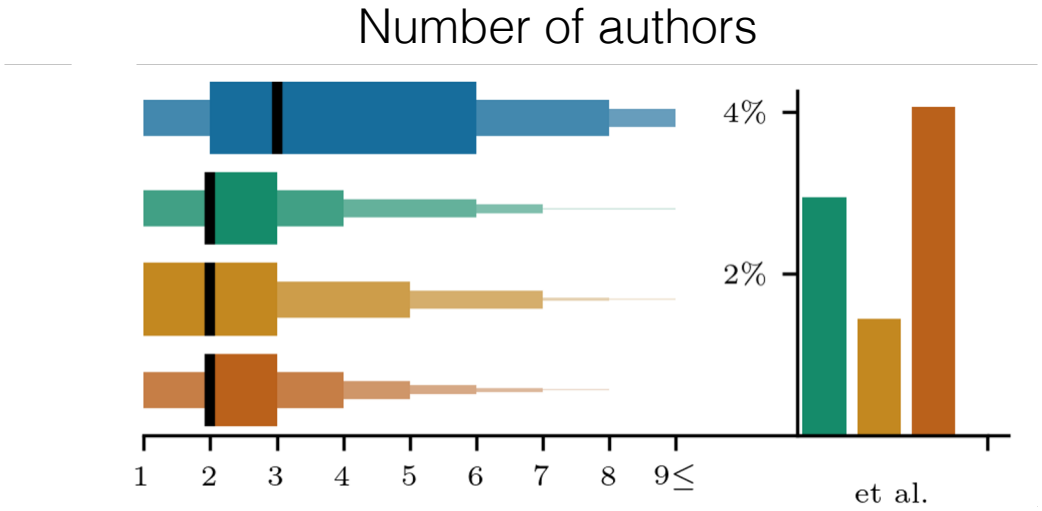


Generated References Reveal a Preference for Recent Publications, Fewer Authors, and Shorter Titles

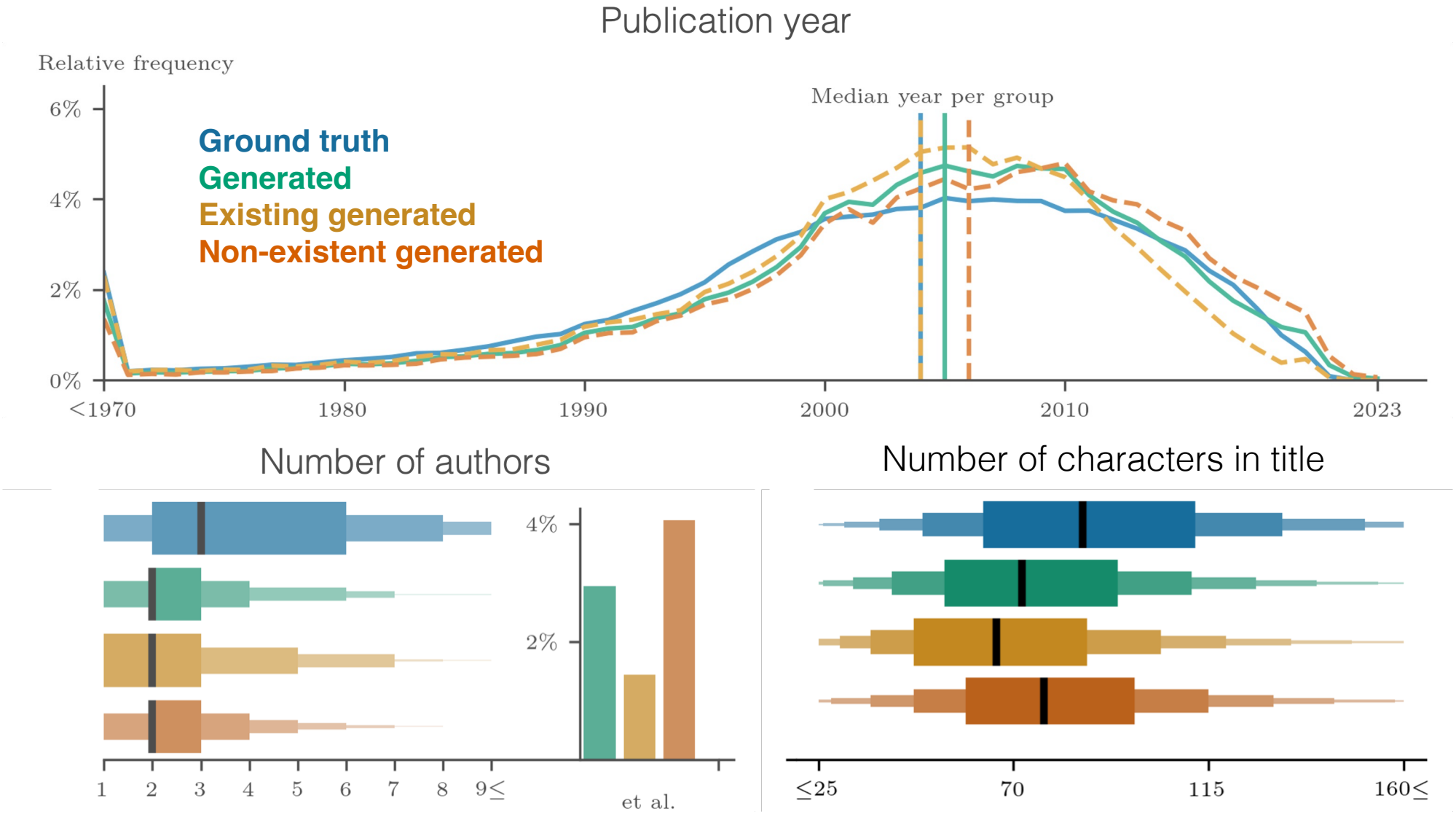


Generated References Reveal a Preference for Recent Publications, Fewer Authors, and Shorter Titles





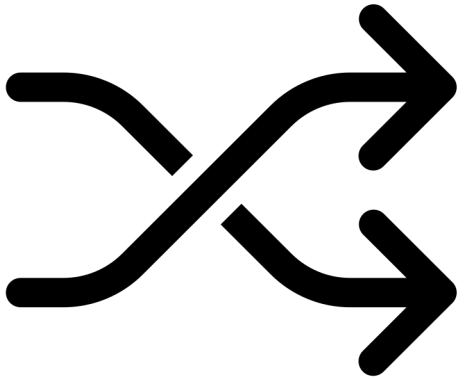
Generated References Reveal a Preference for Recent Publications, Fewer Authors, and Shorter Titles



Generated References Reflect Human Citation Patterns in the OpenAI Embedding Space

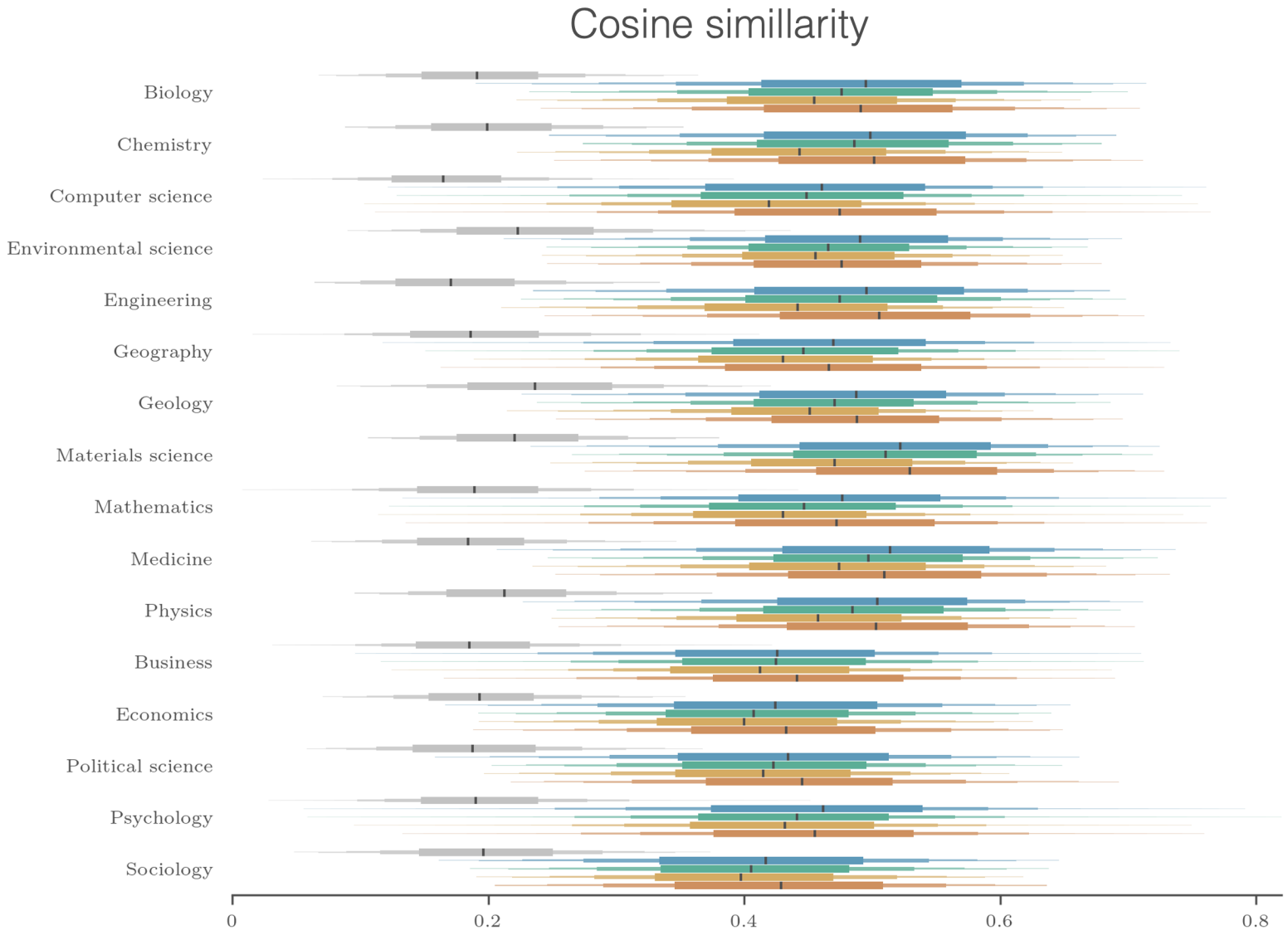
Random baseline

Ground truth references



(while fixing the field of study)

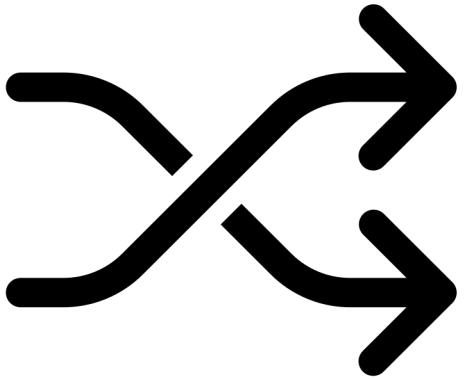
Random
Ground truth
Generated
Existing generated
Non-existent generated



Generated References Reflect Human Citation Patterns in the OpenAI Embedding Space

Random baseline

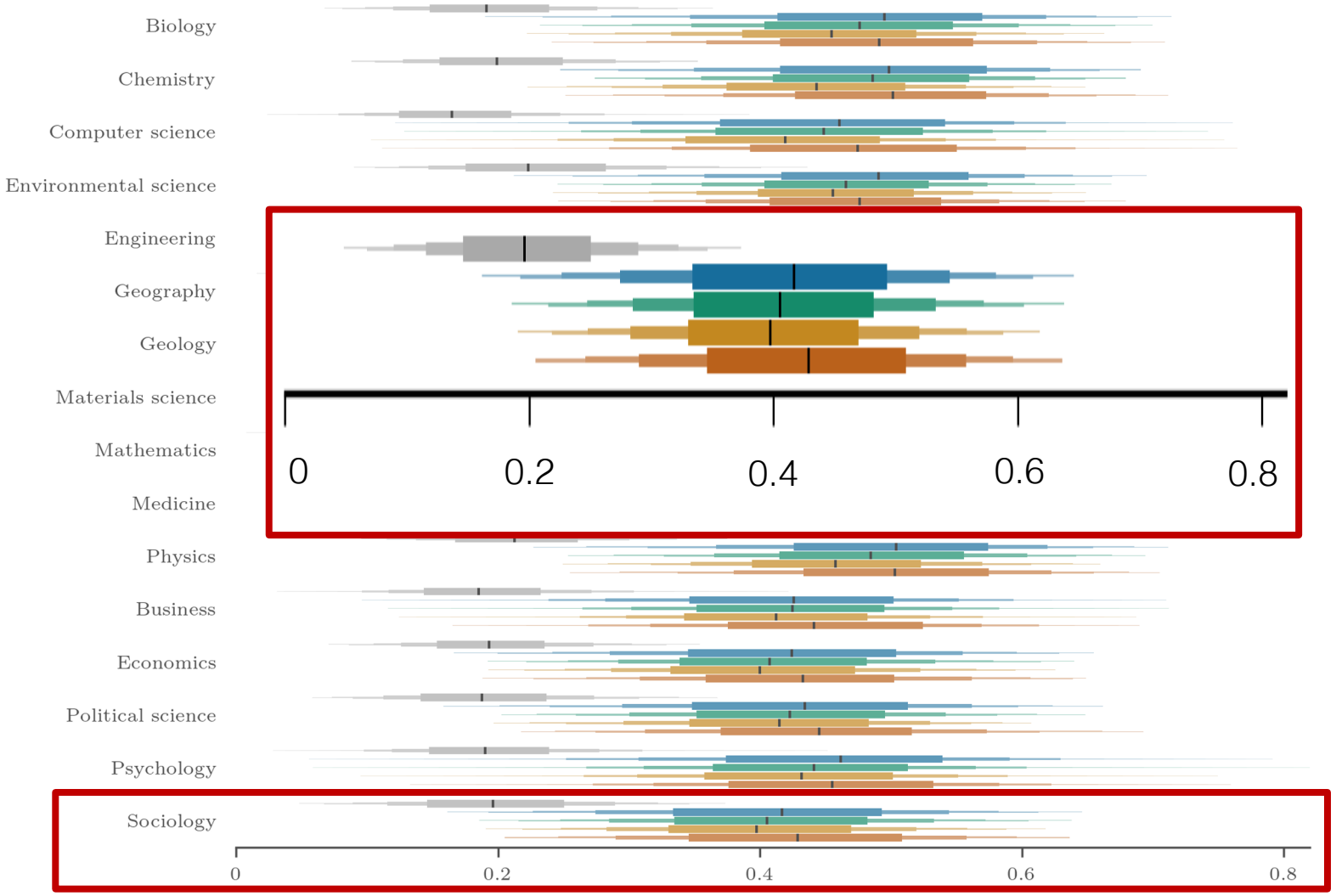
Ground truth references



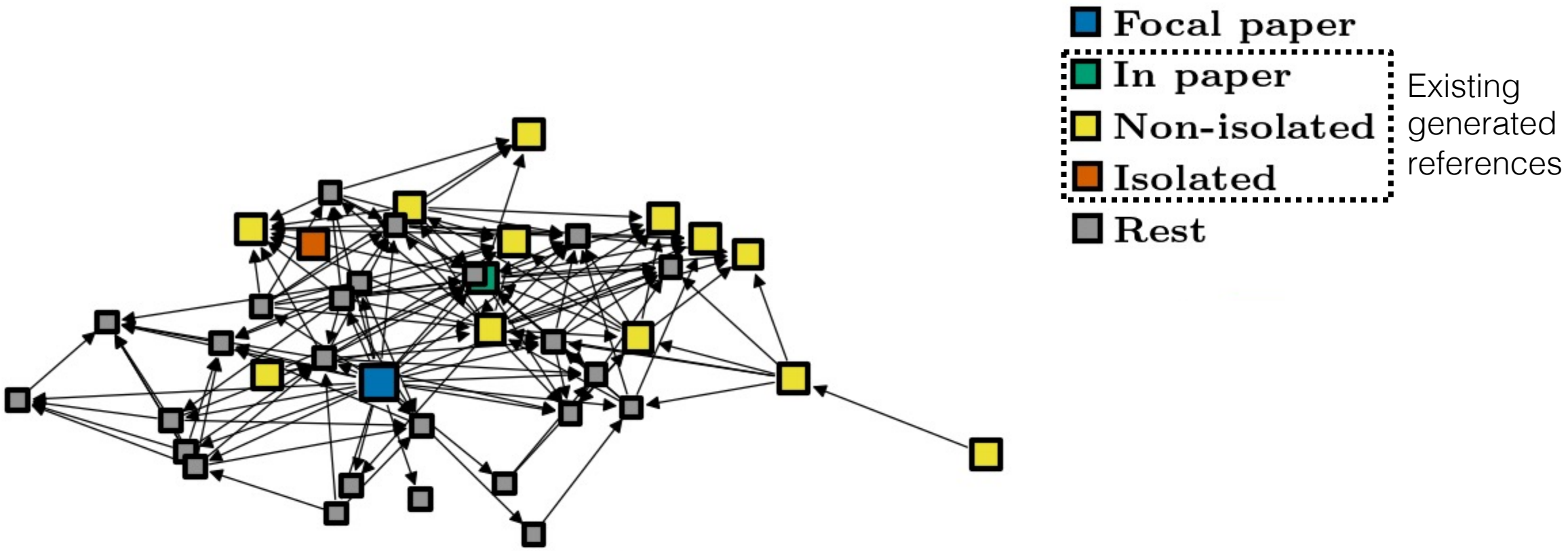
(while fixing the field of study)

Random
Ground truth
Generated
Existing generated
Non-existent generated

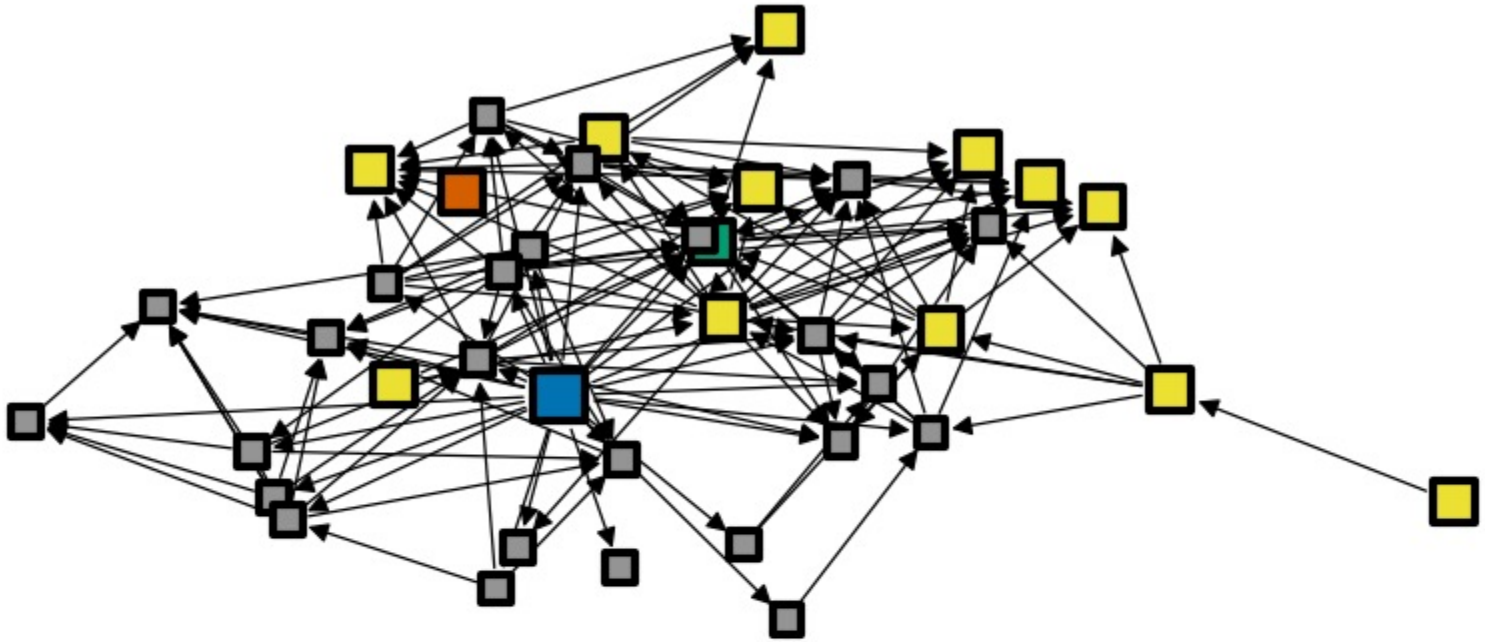
Cosine simillarity

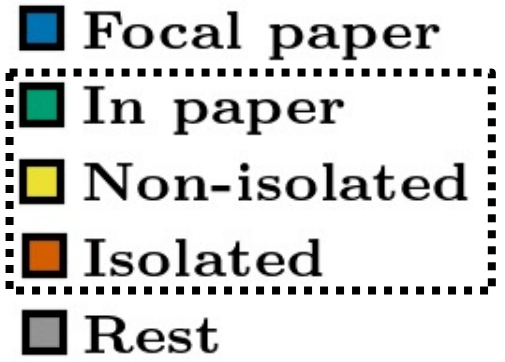


Generate References Are Well Embedded in the Local Citation Context



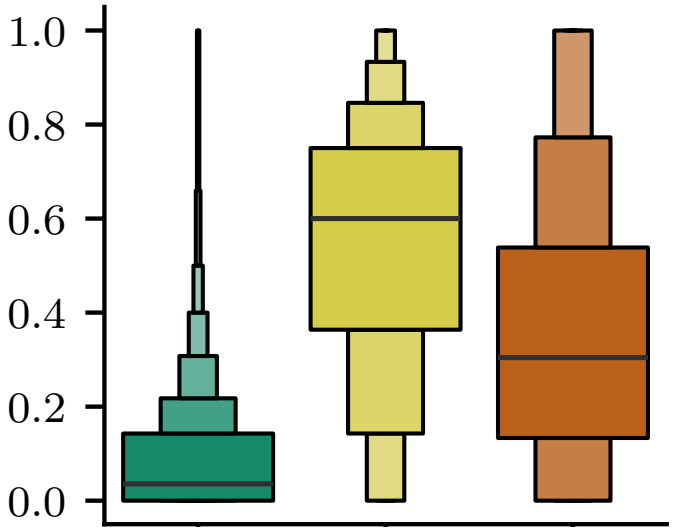
Generate References Are Well Embedded in the Local Citation Context





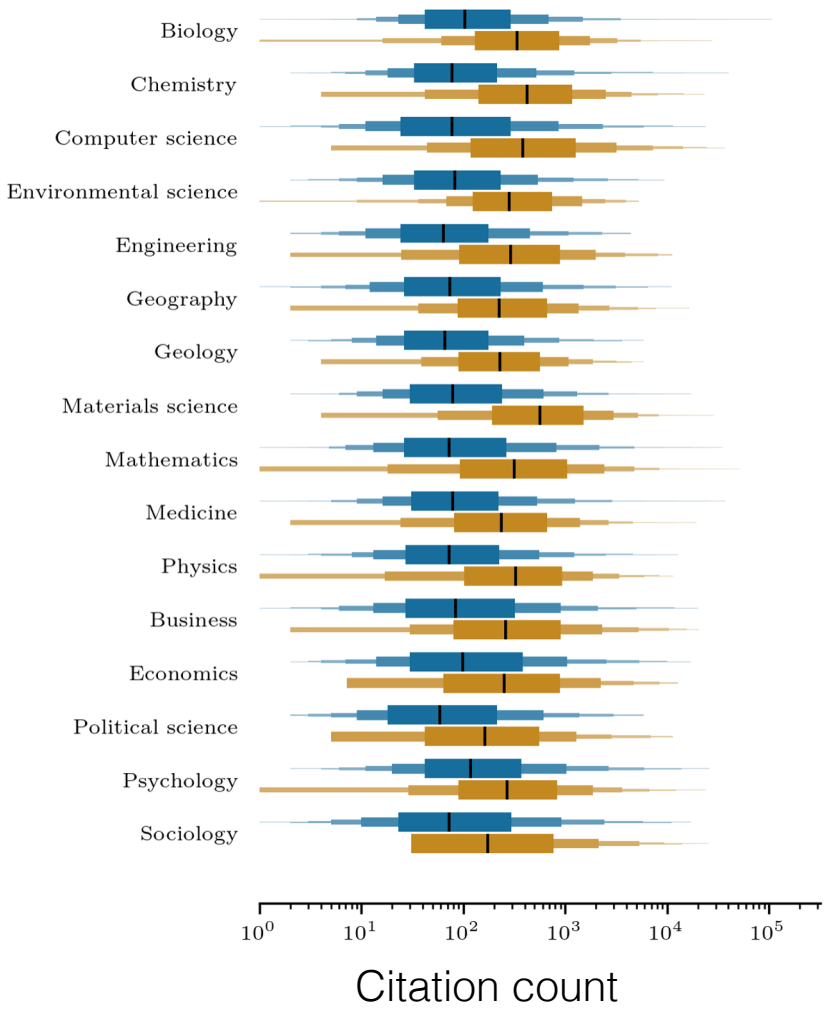
Existing generated references

Existing gen. references per focal paper

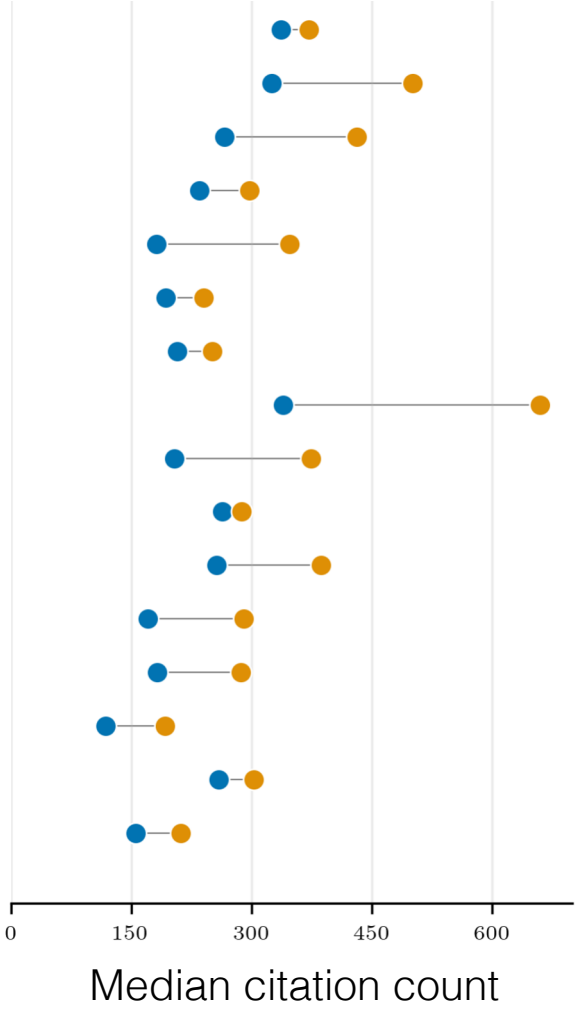


Key Takeaways

LLMs Exhibit a Heightenend Citation Bias



Even When Controlling for Selection Bias

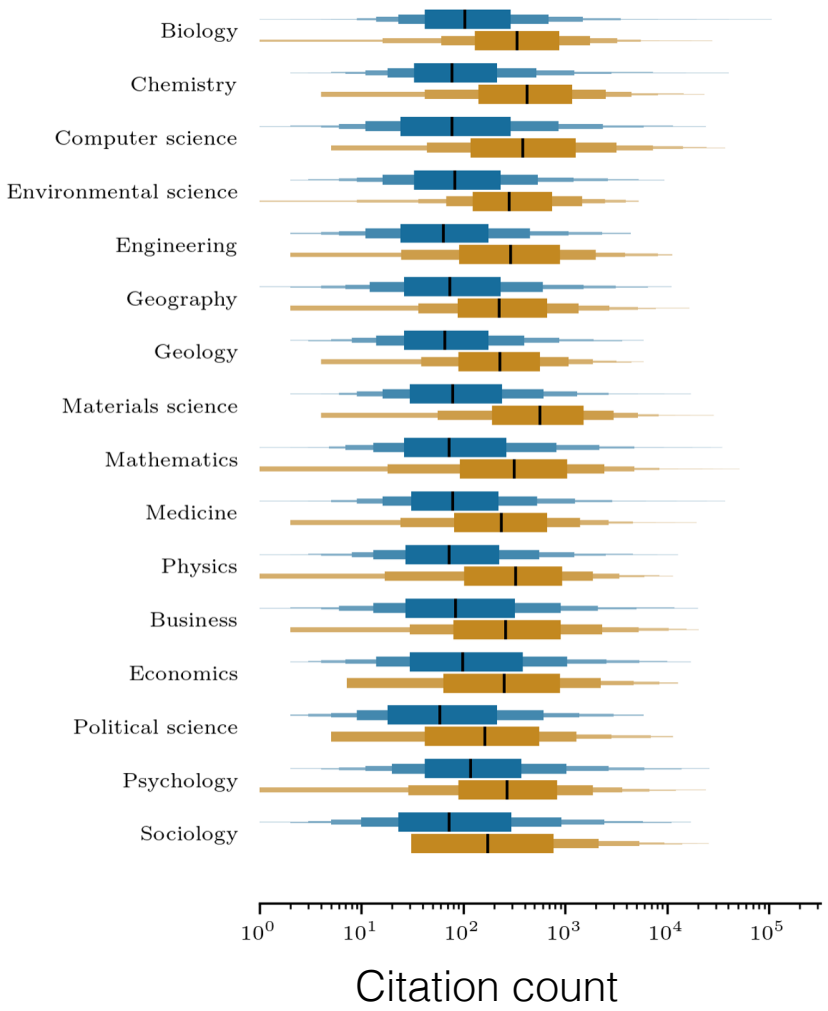


Ground truth

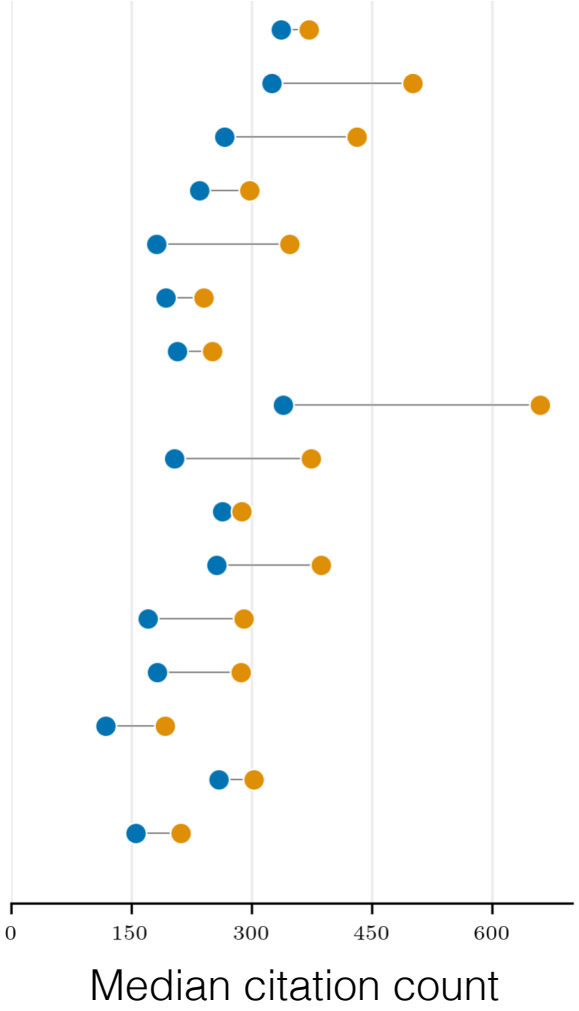
Existing generated

Key Takeaways

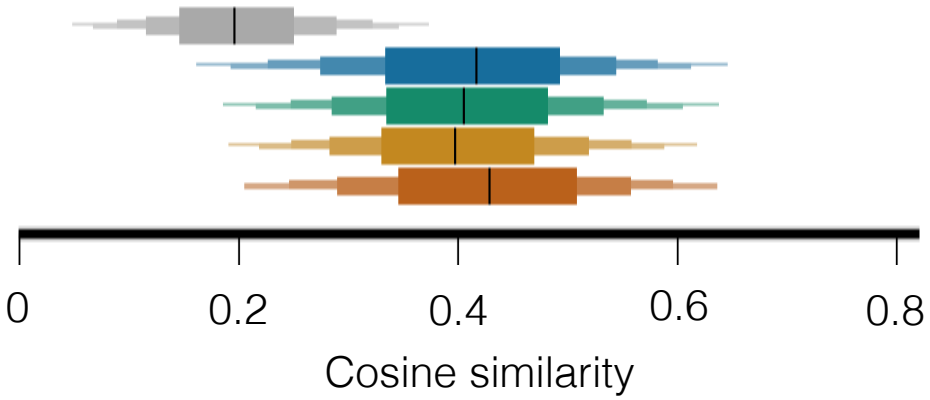
LLMs Exhibit a Heightenend Citation Bias



Even When Controlling for Selection Bias



LLMs Resemble Human Citation Patterns in the OpenAI Embedding Space



Random
Ground truth
Generated
Existing generated
Non-existent generated

Large Language Models Amplify the Matthew Effect in Scientific Research

Main paper

Related work

In-text citations for Computer Science (NAACL findings)

Graphs & embeddings (ICLR tiny paper)


Andres Algaba


Floriano Tori


Melika Mobini


Brecht Verbeken


Sylvia Wenmackers


Vincent Ginis

https://github.com/AndresAlgaba/LLMs_scientific_literature

https://zenodo.org/records/15124610

VUB DATA ANALYTICS LABORATORY RESEARCH GROUP

KU LEUVEN Centre for Logic and Philosophy of Science (CLPS)

Harvard John A. Paulson School of Engineering and Applied Sciences