

Text analytics using large language models




Large Language Models Outperform Humans in Text Annotation

1. Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30), e2305016120.
2. Le Mens, G., Kovács, B., Hannan, M. T., & Pros, G. (2023). Uncovering the semantics of concepts using GPT-4. *Proceedings of the National Academy of Sciences*, 120(49), e2309350120.
3. Peiyao Li, Noah Castelo, Zsolt Katona, Miklos Sarvary (2024) Frontiers: Determining the Validity of Large Language Models for Automated Perceptual Analysis. *Marketing Science*, 43(2):254-266.
4. Krugmann, J. O., & Hartmann, J. (2024). Sentiment Analysis in the Age of Generative AI. *Customer Needs and Solutions*, 11(1), 1-19.



Uncovering the semantics of concepts using GPT-4

Gaël Le Mens ^{a,1}, Balázs Kovács ^b, Michael T. Hannan ^c, and Guillem Pros ^a

Edited by Kenneth Wachter, University of California, Berkeley, CA; received June 3, 2023; accepted October 13, 2023

November 30, 2023 | 120 (49) e2309350120 | <https://doi.org/10.1073/pnas.2309350120>







Significance

We use GPT-4 to create “typicality measures” that quantitatively assess how closely text documents align with a specific concept or category. Unlike previous methods that required extensive training on large text datasets, the GPT-4-based measures achieve state-of-the-art correlation with human judgments without such training. Because training data is not needed, this dramatically reduces the data requirements for obtaining high performing model-based typicality measures. Our analysis spans two domains: judging the typicality of books in literary genres and the typicality of tweets in the Democratic and Republican parties. Our results demonstrate that modern Large Language Models (LLMs) can be used for text analysis in the social sciences beyond simple classification or labelling.

The ability of recent Large Language Models (LLMs) such as GPT-3.5 and GPT-4 to generate human-like texts suggests that social scientists could use these LLMs to construct measures of semantic similarity that match human judgment. In this article, we provide an empirical test of this intuition. We use GPT-4 to construct a measure of typicality—the similarity of a text document to a concept. We evaluate its performance against other model-based typicality measures in terms of the correlation with human typicality ratings. We conduct this comparative analysis in two domains: the typicality of books in literary genres (using an existing dataset of book descriptions) and the typicality of tweets authored by US Congress members in the Democratic and Republican parties (using a novel dataset). The typicality measure produced with GPT-4 meets or exceeds the performance of the previous state-of-the art typicality measure we introduced in a recent paper [G. Le Mens, B. Kovács, M. T. Hannan, G. Pros Rius, *Sociol. Sci.* 2023, 82–117 (2023)]. It accomplishes this without any training with the research data (it is zero-shot learning). This is a breakthrough because the previous state-of-the-art measure required fine-tuning an LLM on hundreds of thousands of text documents to achieve its performance.

categories | chatGPT | deep learning | typicality | LLM

GPT is an effective tool for multilingual psychological text analysis

Steve Rathje ^{a,2,1}, Dan-Mircea Mirea ^{b,2,1}, Ilia Sucholutsky ^c, Raja Marjeh ^b, Claire E. Robertson ^a, and Jay J. Van Bavel ^{a,d,e}


Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received May 30, 2023; accepted June 18, 2024

August 12, 2024 | 121 (34) e2308950121

Significance

Many fields—including psychology, sociology, communications, political science, and computer science—use computational methods to analyze text data. However, existing text analysis methods have a number of shortcomings. Dictionary methods, while easy to use, are often not very accurate when compared to recent methods. Machine learning models, while more accurate, can be difficult to train and use. We demonstrate that the large-language model GPT is capable of accurately detecting various psychological constructs (as judged by manual annotators) in text across 12 languages, using simple prompts and no additional training data. GPT thus overcomes the limitations present in existing methods. GPT is also effective in several lesser-spoken languages, which could facilitate text analysis research from understudied contexts.

ChatGPT outperforms crowd workers for text-annotation tasks

Fabrizio Gilardi ¹, Meysam Alizadeh , and Maël Kubli 

Edited by Mary Waters, Harvard University, Cambridge, MA; received March 27, 2023; accepted June 2, 2023

July 18, 2023 | 120 (30) e2305016120 | <https://doi.org/10.1073/pnas.2305016120>

Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using four samples of tweets and news articles ($n = 6,183$), we show that ChatGPT outperforms crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection. Across the four datasets, the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25 percentage points on average, while ChatGPT’s intercoder agreement exceeds that of both crowd workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than \$0.003—about thirty times cheaper than MTurk. These results demonstrate the potential of large language models to drastically increase the efficiency of text classification.

ChatGPT | text classification | large language models | human annotations | text as data

Department of Political Science, University of Zurich, Zurich 8050, Switzerland

This article contains supporting information online at <https://doi.org/10.1073/pnas.2305016120#supplementary-materials>.

¹To whom correspondence may be addressed. Email: gilardi@ipz.uzh.ch.
This open access article is distributed under [Creative Commons Attribution License 4.0 \(CC BY\)](#).

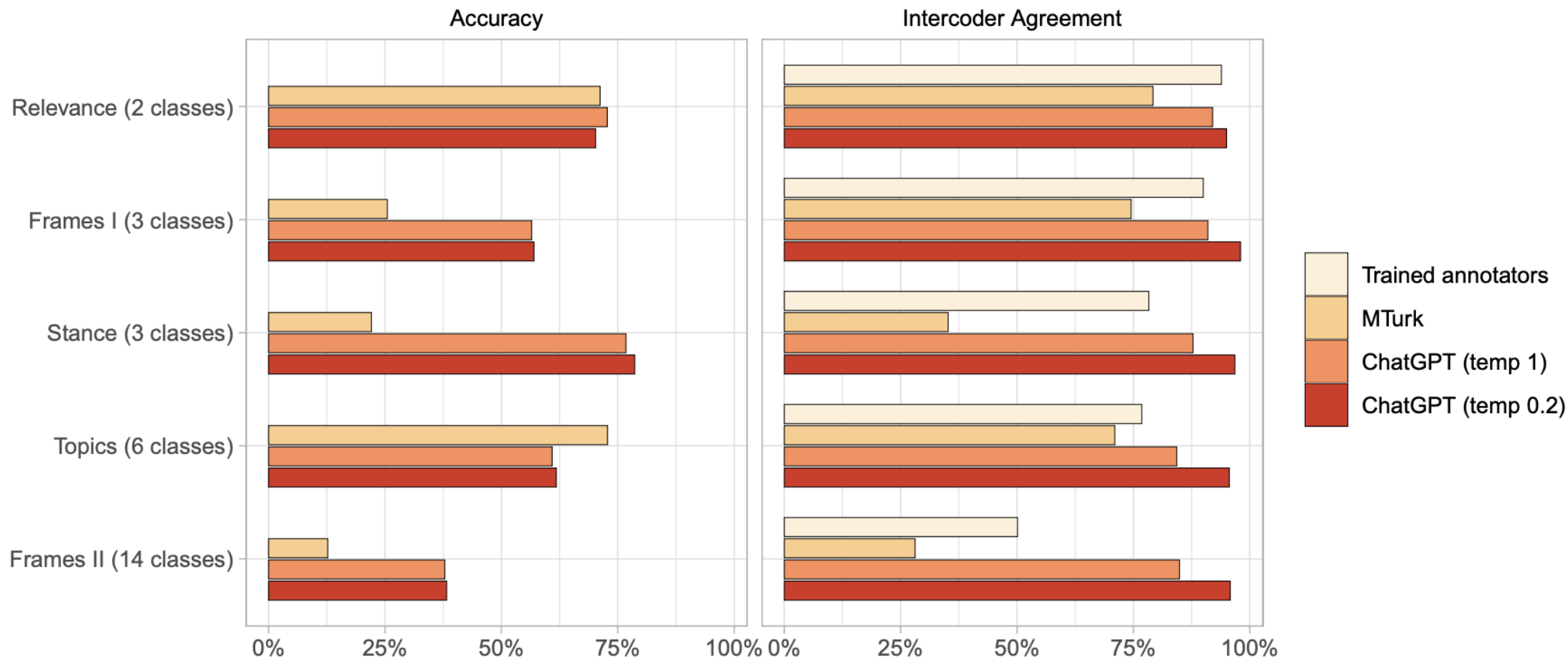


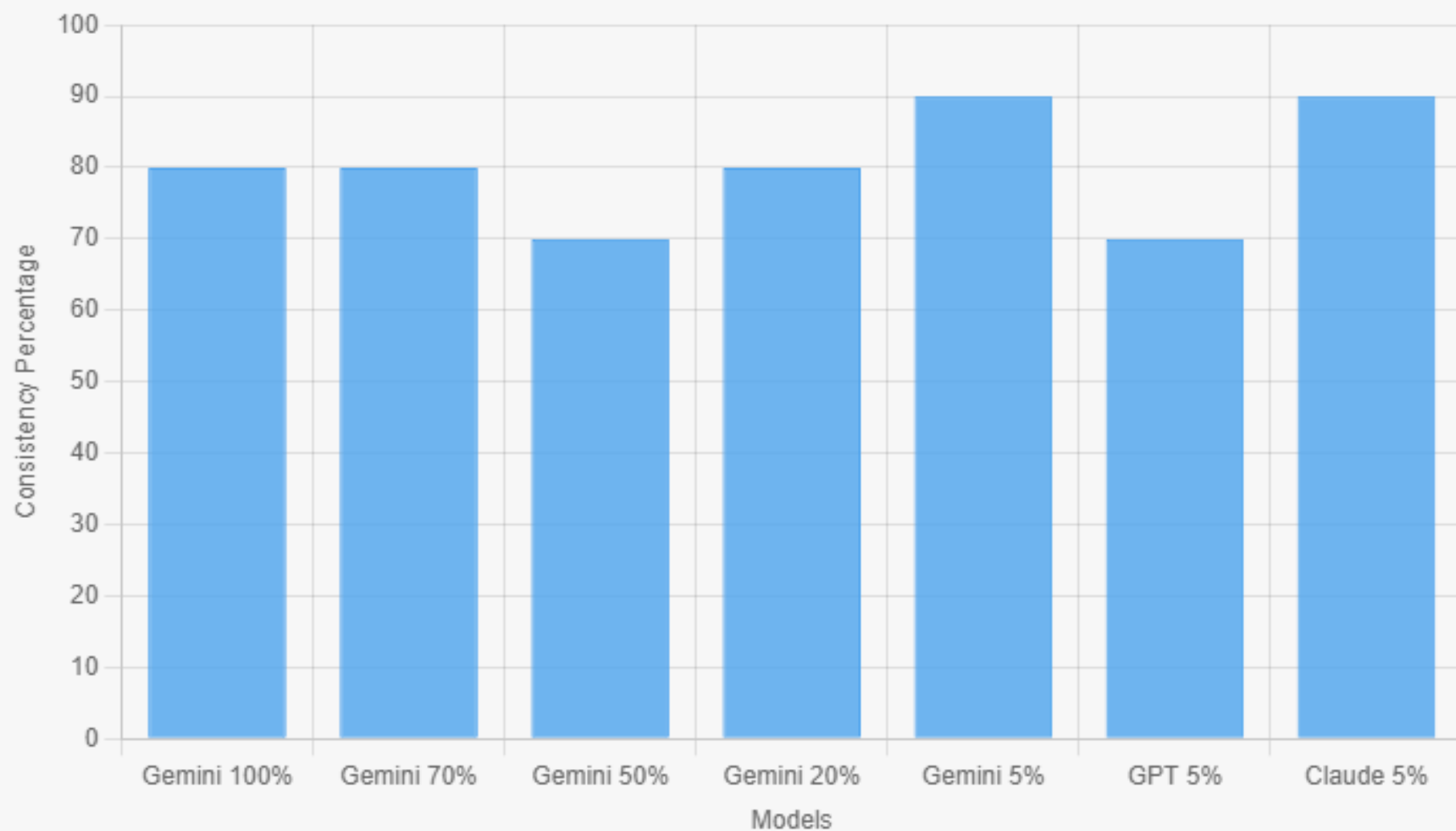
Figure 1: *ChatGPT zero-shot text annotation performance, compared to MTurk and trained annotators. ChatGPT’s accuracy outperforms that of MTurk for four of the five tasks. ChatGPT’s intercoder agreement outperforms that of both MTurk and trained annotators in all tasks. Accuracy means agreement with the trained annotators.*

Large Language Models Outperform Humans in Topic Modelling (Theme Generation and Classification)

1. Yu, S. (2024). Using Large Language Models In Short Text Topic Modeling: Model Choice And Sample Size (No. mqk3r_v1). *Center for Open Science*.
2. Akash, P. S., & Chang, K. C. C. (2024). Enhancing Short-Text Topic Modeling with LLM-Driven Context Expansion and Prefix-Tuned VAEs. *arXiv preprint* arXiv:2410.03071.
3. Mu, Y., Dong, C., Bontcheva, K., & Song, X. (2024). Large language models offer an alternative to the traditional approach of topic modelling. *arXiv preprint* arXiv:2403.16248.
4. Invernici, F., Curati, F., Jakimov, J., Samavi, A., & Bernasconi, A. (2024). Capturing research literature attitude towards Sustainable Development Goals: an LLM-based topic modeling approach. *arXiv preprint* arXiv:2411.02943.
5. Rizvi, S. A. M., Tasneem, F. B., Ahmad, R., Khan, M. U., Jamshed, N., & Ahmad, T. (2024, November). Exploring the Efficacy of Large Language Models in Topic Modelling: A Comparative Analysis. In *2024 3rd Edition of IEEE Delhi Section Flagship Conference (DELCON)* (pp. 1-6). IEEE.



Consistency Percentage Across Models



Large Language Models Match Human Experts in Scientific Feedback

1. Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., ... & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJMAI*, 1(8), Aloa2400196.
2. Zhuang, Z., Chen, J., Xu, H., Jiang, Y., & Lin, J. (2025). Large language models for automated scholarly paper review: A survey. *arXiv preprint* arXiv:2501.10326.
3. Liu, R., & Shah, N. B. (2023). Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint* arXiv:2306.00622.
4. Ye, R., Pang, X., Chai, J., Chen, J., Yin, Z., Xiang, Z., ... & Chen, S. (2024). Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *arXiv preprint* arXiv:2412.01708.



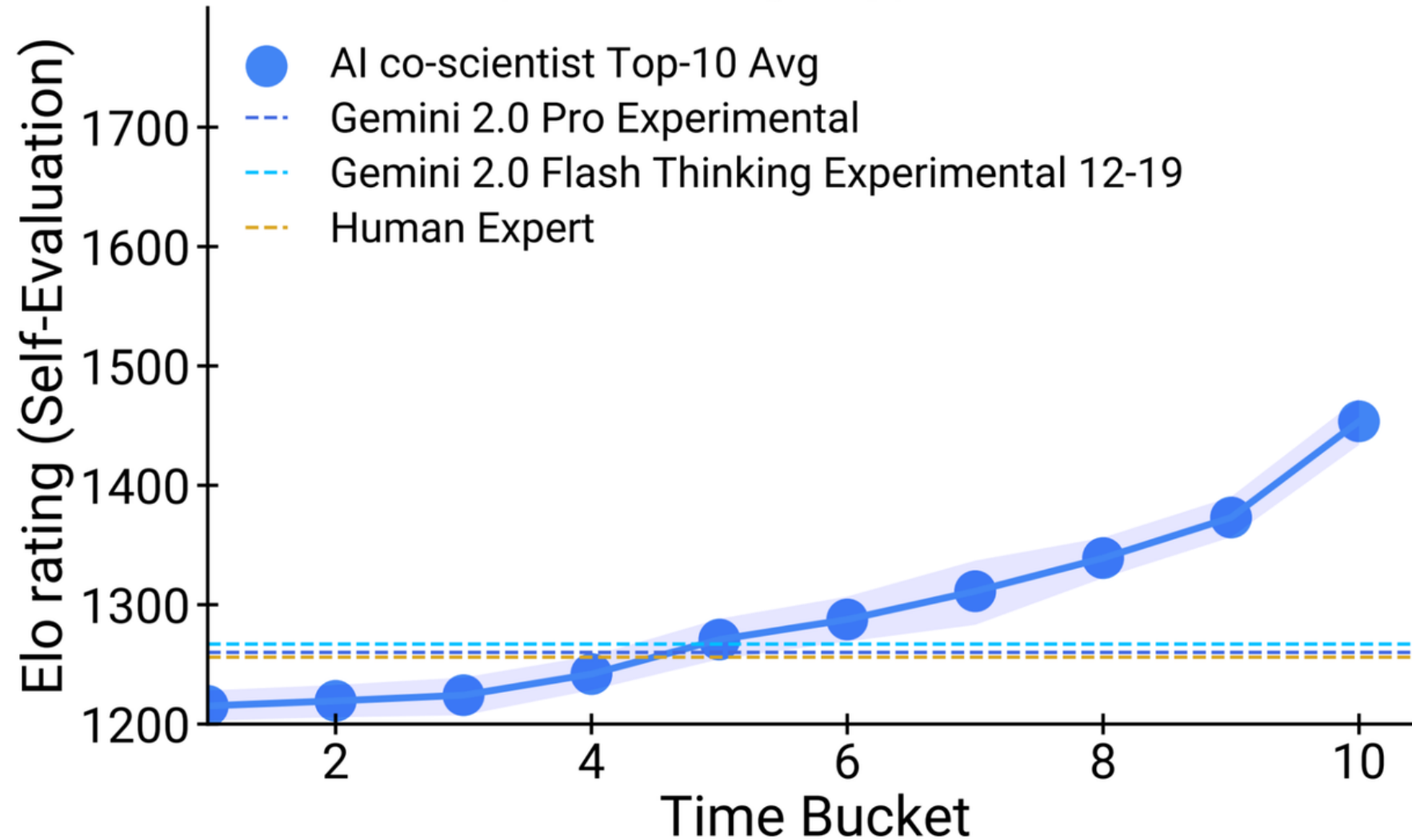
Large language models surpass human experts in predicting scientific results

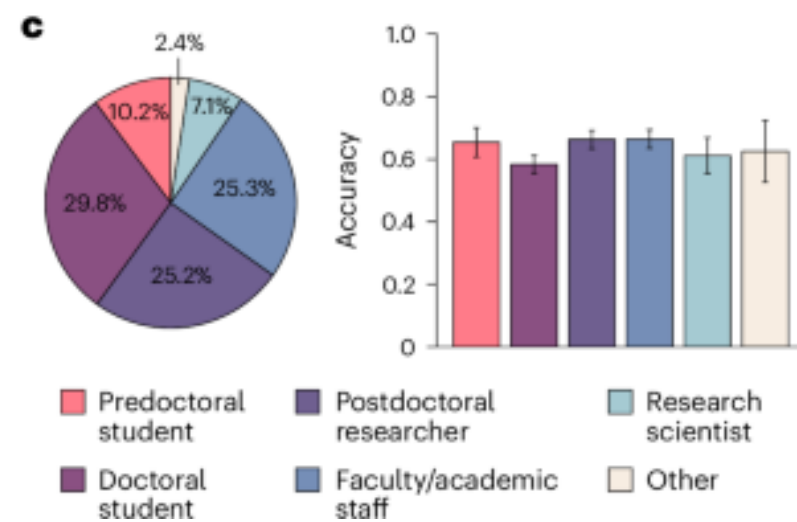
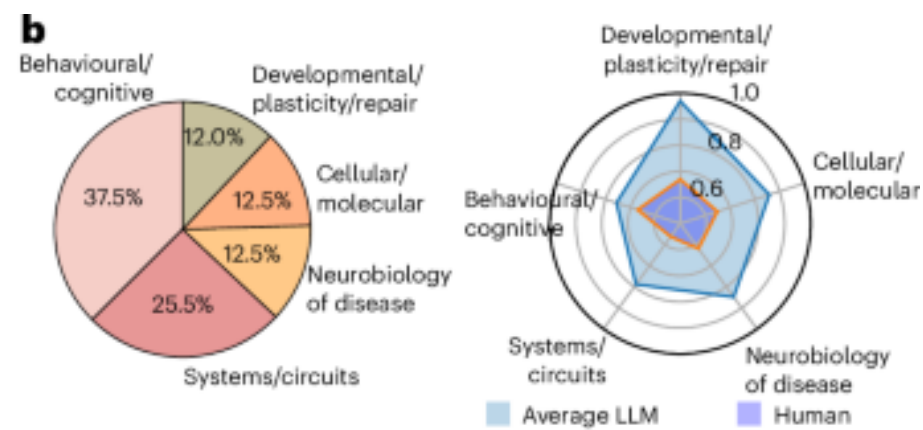
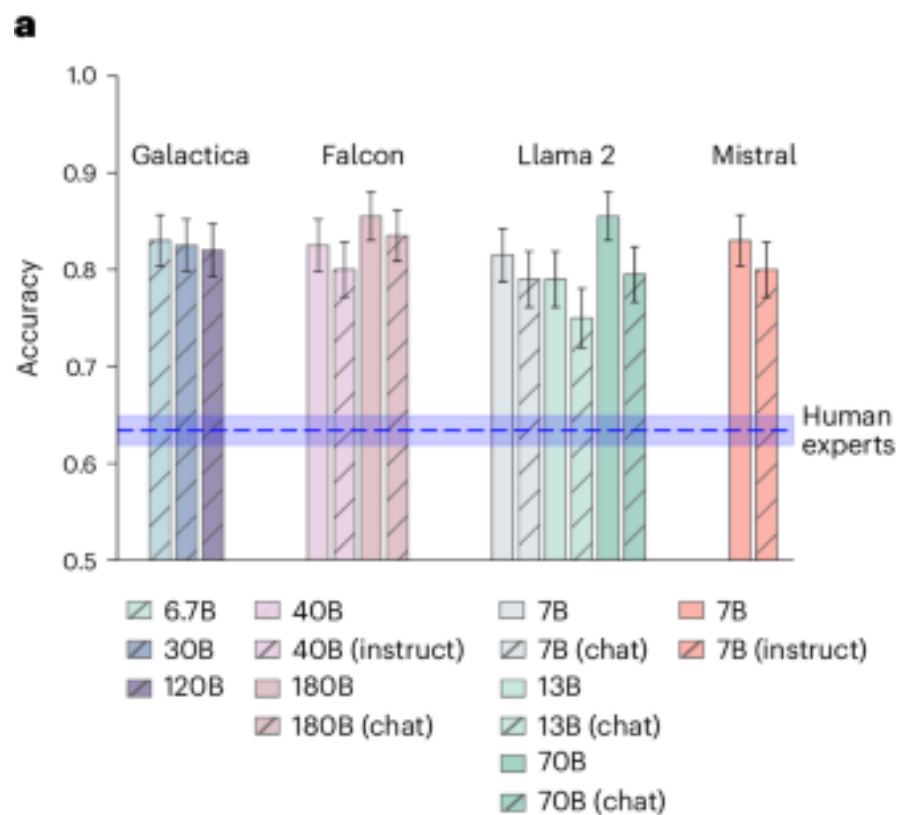
Luo, X., Rechart, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., ... & Love, B. C. (2024). Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, 1-11.

Zhou, Y., Liu, H., Srivastava, T., Mei, H., & Tan, C. (2024). Hypothesis generation with large language models. arXiv preprint arXiv:2404.04326.



Top-10 average hypothesis Elo





Using APIs for Text Analysis

Example 1:

You would like to know how consumers feel about grotesque advertising. You show them three grotesque ads and ask them to write down their feelings.

You use large language models (LLMs) to identify different types of feelings: shock, disgust, fear, creativity, and liking.

Example 2:

Imagine you have collected luxury brands' tweets. You want to know whether communicating sustainability on social media can increase social media engagement (the number of likes, retweets, and comments).

You use LLMs to label each tweet: [sustainability-related tweets] or [non-sustainability-related tweets].



How to identify main topics in a text?

Step 1: Randomly split your dataset into several subsets (e.g., subset 1, subset 2)

Step 2: Copy-paste all the text in subset 1 in LLM 1 and fine tune the prompt to obtain Topic List 1.

Step 3: Copy-paste all the text in subset 2 in LLM 2 and fine-tune the prompt to obtain Topic List 2.

Step 4: Compare Topic List 1 and Topic List 2 for the similarity (e.g., rouge 1 score, consistency)

Step 5: Repeat this process for subset 2 and compare the similarity of the topic list.

Step 6: Decide the final topic list.



ChatGPT

Developer: OpenAI
Performance: ★★★★★
Max length: 128,000 tokens
Price: Free/Subscription



Claude

Developer: Anthropic
Performance: ★★★★★
Max length: 200,000 tokens
Price: Free/Subscription...



Gemini

Developer: Google
Performance: ★★★★★
Max length: 32,000 tokens
Price: Free/Subscription...



Le Chat

Developer: Mistral
Performance: ★★★★★
Max length: 32,000 tokens
Price: Free...



Kimi

Developer: Moonshot
Performance: ★★★★★
Max length: 200,000 tokens
Price: Free...



Gemini 1.5

Developer: Google [Trial]
Performance: ★★★★★
Max length: 31 million tokens
Price: Free



How to use APIs to label the text?

Preparation

Step 1: register an account at an LLM provider (e.g., OpenAI, Gemini, Watsonx)

Step 2: Obtain the API key

Step 3: Prepare a training dataset (e.g., use humans to code 100 random entries)

Step 4: Open a related script at GAIforResearch.com

Automated labeling

Step 5: Modify the script (import data, change variable names, paste the API key)

Step 6: Fine-tune the prompt and adjust the parameters using the training set according to the F1 score.

Step 7: Start the automated coding process for the main dataset (this can take a long time)

Step 8: Draw a random sample to see the accuracy of the automated coding (F1 score or inter-coder reliability)

