Detecting ChatGPT Modification in Global Education Research: A Case Study Across Translated and Open Access Contexts

Jon Ball
Graduate School of Education
Stanford University
jonball@stanford.edu

Anna Saraiva
Department of Economics
Stanford University
annacds@stanford.edu

Sawal Acharya
ICME
Stanford University
sawal386@stanford.edu

Abstract

Education researchers around the world are increasingly relying on OpenAI's ChatGPT to write and edit their work. We leverage a word-distributional quantification framework to estimate that the proportion (α) of ChatGPT-modified text in open access Education article abstracts has exceeded a quarter ($\alpha = 0.28$) in Q1 2024. Translated abstracts show fewer signs of ChatGPT but greater signs of AI modification overall, due to Google Translate ($\alpha = 0.65$). Usage varies extremely across linguistic communities, with Indonesian Education research showing no signs of modification by the models we test. Overall, we conclude that Education researchers increasingly modify their work with ChatGPT, which can improve readability, and that ChatGPT complements Google Translate in multilingual contexts. Because these researchers are also teachers of teachers, we suggest that their large language model (LLM) usage presages greater usage and linguistic homogenization in the future.

1 Introduction

ChatGPT is changing how researchers write and publish their findings, as have previous language technologies. But it is challenging to isolate ChatGPT's influence on writing, let alone other technologies' for comparison. To this end, we present a case study focusing on a specific form of academic writing – open access, English-language, Education research article abstracts – and two use cases for large language models (LLMs): editing and translation. We utilize observed word distributions over human- and LLM-written abstracts in a discontinuity design to predict mixture in abstracts published after ChatGPT's public release (Nov. 30, 2022). We find that words relatively overused by Google Translate are ubiquitous in translated abstracts, while words relatively overused by ChatGPT are becoming more frequent in both translated and English-only abstracts. We argue that growth in the amount of peer-reviewed, LLM-edited Education research corresponds with increased (English-language) readability, which is crucial for contemporary scientific communication. However, we understand that LLMs also threaten research integrity, and so we temper our hope for improved communication of results across sociolinguistic boundaries with awareness of the potential for inauthentic scientific communication in an age of AI.

Educators, the professional community we focus on in this study, have observed vast increases in the share of students using AI tools on assignments (Westfall, 2023). LLMs may also have uses at different stages of the Education research process, for example, during data annotation and conversion (Ziems et al., 2024; Mittelstadt et al., 2023). Researchers in adjacent fields like Social Psychology consider using LLMs to model human behavior as a cheaper alternative to running real-life experiments (Kaddour et al., 2023). LLMs can also aid in scientific writing more generally, by reducing the time authors spend on editing and proofreading (Almarie et al., 2023).

Perhaps for these reasons, and despite widespread concerns about LLM usage in research, we continue to see increased adoption. Liang et al. (2024b) find a steady increase in the share of LLM-generated text in arXiv, bioRxiv, and Nature portfolio journal submissions. In an effort to catch up, journals have enacted changes allowing LLMs, with caveats. For instance, the *New England Journal of Medicine* has allowed researchers to employ LLMs in their submissions (Koller et al., 2024). However, the authors must take ownership of the content produced by these models and must honestly acknowledge their use. Relevant to this study, Elsevier explicitly allows authors to use generative AI to improve the language and readability of their work.

Rather than monitor or map ChatGPT usage across domains, we narrow our focus to a single domain - Education - and estimate the share of LLM-modified text in a corpus of Education research article abstracts. Our domain-specific focus also enables future work, for example, drawing field-specific comparisons with articles indexed by the Education Research Information Center (ERIC) or Education-focused preprint servers like EdArXiv and EdWorkingPapers. We make all data and models from this study accessible for this purpose.[1]

2 Related Work

A classical way to detect text as human / AI-written would be to train a classifier on the feature mappings of the text. Zellers et al. (2019) take this approach and train a linear classifier on the final layer outputs of their text generation model, Grover. Their experiments show that the classifier's performance is high on texts generated by Grover itself. Uchendu et al. (2020) take a step further by training a deep-learning model to classify text as human and AI-written. More recently, the focus has turned towards using zero-shot detection methods. A popular zero-shot text detection method is DetectGPT (Mitchell et al., 2023). The core idea behind the technique is that the texts generated by LLM models tend to lie in regions with high negative curvature (close to local maxima) of the logarithm of the probability function used in generating texts. By quantifying how far a text's representation is from the region with high curvatures, DetectGPT classifies the text as human or AI written.

The methods described above depend on access to source model weights, which may not be readily available. To get around this issue, Liang et al. (2024a) propose the **Distributional GPT Quantification** framework, which we describe in detail in Section 4.1. The authors applied this framework to estimate the proportion of ChatGPT-edited text in machine learning conference peer reviews. Likewise, Liang et al. (2024b) apply the framework to quantify the influence of ChatGPT on STEM preprints hosted by ArXiv and BioarXiv, as well as articles published in Nature Portfolio journals.

On the qualitative side, Lund et al. (2023) discuss how ChatGPT may impact academia and libraries specifically, noting the potential for improved academic search and discovery as well as threats to privacy. Kasneci et al. (2023) argue that widespread usage of LLMs underscores the importance of educators developing relevant competencies and literacies, while also reiterating the core importance of critical thinking and fact checking skills. Demszky et al. (2023) state various opportunities for using LLMs productively in Psychology, from data augmentation to experimentation and feedback, although they qualify by emphasizing that domain-specific data collection, benchmarking, and access agreements will be necessary to realize the supposed benefits of LLMs.

3 Data

For ChatGPT detection, our training data comprise human-written, English Education article abstracts published in calendar year 2021 ($n = 27,010$). We re-write each of these training abstracts using OpenAI's gpt-3.5-turbo-0613 (2023), in order to estimate the model's word distribution over comparable English-language outputs. Prompts used in all API calls are provided in the appendix. We validate on abstracts published during calendar year 2022 ($n = 43,479$), up until the cutoff of ChatGPT's Nov. 30 release. Finally, we estimate the proportion of LLM-modified text in a test set of Education research article abstracts published from January 2023 to March 2024 ($n = 66,425$).

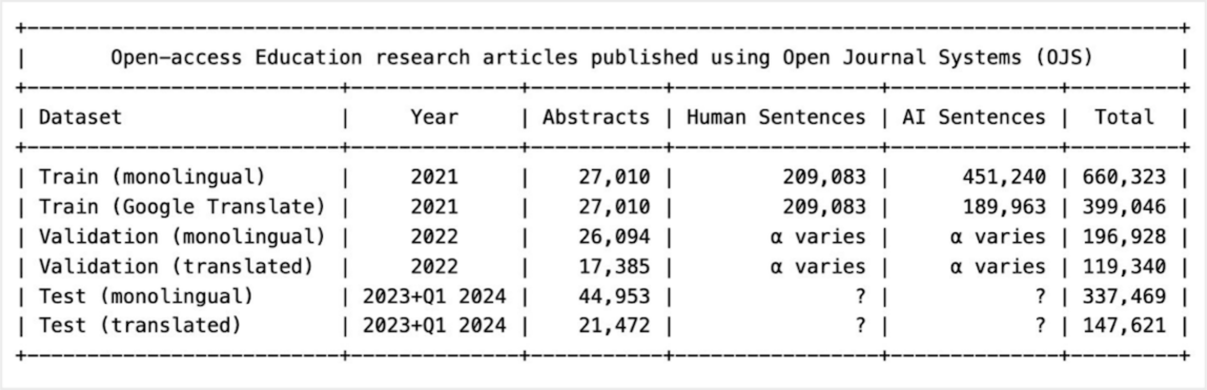
All articles analyzed in this study are open access articles published using the open source platform Open Journal Systems (OJS).[2] OJS automatically generates and stores OAI-PMH[3] metadata for each article published using the software, and these metadata form the basis for our study. Visualizations of OJS usage patterns over time by language and field of study are provided in the appendix. For context, academic journals using OJS are open access, skew new, and generally managed by small, independent publishers. But the fact that these articles are peer reviewed, albeit inconsistently, renders more surprising our findings about the extent of LLM modification therein.

OJS users are mainly based in the global South,

[1] Github link to be provided. (MIT license)

[2] OJS is an open source publishing software used worldwide and maintained by the Public Knowledge Project (PKP). A journal-level dataset with updated summary statistics on global OJS usage can be accessed at https://doi.org/10.7910/DVN/OCZNVY.

[3] https://www.openarchives.org/pmh/



```
+----------------------------------------------------------------------------------------+
|       Open-access Education research articles published using Open Journal Systems (OJS)       |
+--------------------------+-------------+-----------+-----------------+--------------+---------+
| Dataset                  |     Year     | Abstracts | Human Sentences | AI Sentences |  Total  |
+--------------------------+-------------+-----------+-----------------+--------------+---------+
| Train (monolingual)      |     2021     |    27,010 |         209,083 |      451,240 | 660,323 |
| Train (Google Translate) |     2021     |    27,010 |         209,083 |      189,963 | 399,046 |
| Validation (monolingual) |     2022     |    26,094 |        α varies |     α varies | 196,928 |
| Validation (translated)  |     2022     |    17,385 |        α varies |     α varies | 119,340 |
| Test (monolingual)       | 2023+Q1 2024 |    44,953 |               ? |            ? | 337,469 |
| Test (translated)        | 2023+Q1 2024 |    21,472 |               ? |            ? | 147,621 |
+--------------------------+-------------+-----------+-----------------+--------------+---------+
```

Figure 1: OJS Article Statistics

with a plurality of users in Indonesia. We do not filter for publications from specific countries on the basis of ChatGPT availability, mainly because the authors of this corpus are international, and we only have journal top-level domains to draw assumptions from. Less than half of OJS editors publish articles exclusively in English, while others use the software to publish in 59 additional languages (Khanna et al., 2022).

To predict language labels for abstracts, we use fastText's pretrained language classifier (Joulin et al., 2017). To predict field of study labels, we use AllenAI / Semantic Scholar's pretrained s2_fos model.[4] The s2_fos model was trained on English-language abstracts mapped to disciplinary labels (both human- and GPT-4-assigned), hence why we narrow to articles for which there is an English version of the abstract. We predict field of study over title, English abstract, and name of publisher for each record. We only include abstracts in our training, validation, and test sets if Education is the predicted field of study with confidence greater than 0.9. Sampling and subsequent manual review revealed no articles falling outside the (admittedly wide) purview of Education.

Because we want to model a mixture of human-written and LLM-generated text, we focus first on article records for which *only* an English-language abstract is provided ($n{=}44,953$). We test separately on English-language abstracts for which there are multiple language versions ($n = 21,472$). We observe a bimodal distribution over the length in tokens of translated abstracts, consistent with LLM modification by Google Translate, and our ChatGPT detection models additionally trained on these translated abstracts perform worse on our validation set. The rationale for not training on, and predicting separately over translated abstracts is therefore empirical. It also aligns with our study's objective of distinguishing between editing and translation use cases.

Because we exclusively focus on Education research and analyze a more or less consistent set of publications within this domain (those published in journals using OJS), we have reason to believe that core topics are consistent across the 2021-2022 and 2023-2024 time slices. The exception, of course, is ChatGPT as a topic.

4 Methods

4.1 Distributional GPT Quantification

We describe the method presented in Liang et al. (2024b) and Liang et al. (2024a). The framework describes a document, x, as a mixture of human and machine-generated text. Let P and Q denote the probability distributions of documents written by humans and ChatGPT [5], respectively. The distribution of the document is,

$$D(x) = \alpha Q(x) + (1 - \alpha)P(x) \qquad (1)$$

The main variable of interest is α, which is estimated using a maximum likelihood (MLE) approach. Before using MLE, one needs to estimate the distributions P and Q. Liang et al. (2024a) propose a simple yet effective scheme for doing so. Let X_{human} denote the corpus consisting of human-written texts only. If t is any token in set T,

[4] https://github.com/allenai/s2_fos

[5] The framework is readily extendable to other LLMs as well, as our analysis of Google Translate usage shows.

define

$$\hat{p}(t) = \frac{\text{\# documents in } X_{human} \text{ containing t}}{\text{\#documents in } X_{human}} \tag{2}$$

Then, for each human-written document, $x_i \in X_{human}$, define

$$P(x_i) = \prod_{t \in x_i} \hat{p}(t) \prod_{t \notin x_i} (1 - \hat{p}(t)) \tag{3}$$

$Q(x_i)$ and $\hat{q}(t)$ can be defined analogously on the set of ChatGPT-written documents, X_{GPT}. Then, α is given by,

$$\hat{\alpha} = \arg\max_{\alpha \in [0,1]} \sum_{i=1}^{n} \log((1-\alpha)P(x_i) + \alpha Q(x_i)), \tag{4}$$

In the above equation, n is the total documents in our target corpus. The distribution, $\hat{p}(t)$ can be estimated empirically from a sample of human-written documents published before the release of ChatGPT. To estimate $\hat{q}(t)$, we follow the two-stage approach presented in Liang et al. (2024b). The first step involves generating a bullet-point summary of the abstract using ChatGPT. In the second step, we prompt the ChatGPT to write a paper based on the summary. As Liang et al. (2024b) point out, this process closely mirrors how researchers use LLMs in their writings.

4.2 Implementation

We use the code[6] provided by Liang et al. (2024b) to estimate α. To validate the estimates of alpha yielded by the framework, we employ the scheme outlined in Algorithm 1

Algorithm 1 Validation Scheme

```
Create $X_{train}$ and $X_{val}$ from the combination of $X_{human}$ and $X_{GPT}$
Estimate the distributions P and Q using $X_{train}$
for $\alpha \in \Omega$ do
    Create a synthetic corpus, $X_{synthetic}$ by sampling $\alpha|X_{val}|$ ChatGPT-written documents and $(1-\alpha)|X_{val}|$ human written documents from $X_{val}$
    Estimate $\hat{\alpha}$ on $X_{synthetic}$ using equation 4
end for
```

In our case, we set Ω = {0, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225, 0.25}. The comparison between the α estimated by the model and ground truth is made in Figure 5. The base error rate for our model, $\hat{\alpha}$ when ground truth is 0, is around 6%. The figure also shows that the model performance improves as ground truth α increases.

4.3 Robustness Checks

Our "gpt-3.5-turbo-0613"-trained model also detects the much larger "gpt-4-turbo-0613" model on validation within a 1% margin of error (0.06 vs. 0.066, respectively, with CI 0.005). We train this model only on paired human-written and AI-rewritten abstracts ($n = 27,010$) from 2021. Adding paired human-written and AI-rewritten abstracts from 2020 ($n = 18,587$) had no significant effect on validation performance, resulting in wasted compute. Adding an additional ($n = 22,755$) human-written abstracts from 2000-2019 worsened performance. Adding (LLM-)translated abstracts from 2020-2021 ($n = 33,791$) also worsened performance. By contrast, our best Google Translate detection model was trained on paired human-written and specifically Google Translated abstracts from both 2020 and 2021 ($n = 33,791$).

When evaluating our ChatGPT detection model on the Google Translate validation set, and vice versa, we observe a nearly uniform distribution over predicted alphas (0.06 for ChatGPT $\rightarrow$ Google Translate, and 0.075 for the opposite). As ground truth α increases all the way from 0 to 0.25 in each validation set, however, we do observe a marginal but statistically significant increase in the predicted α. Some words are therefore favored by both ChatGPT and Google Translate relative to our Education corpus.

5 Results

Our estimates for the proportion of ChatGPT-modified text in the monolingual English 2023-2024 test corpus grow from 5.5% in January 2023 (below our error rate of 6.0%) to 29.0% in March 2024 4. We estimate systematically lower proportions of ChatGPT-modified text in the corpus of translated abstracts, although this may be due to model error when predicting over Google Translate-modified text, which is highly prevalent in translated abstracts ($\bar{\alpha} = 0.674$). Based on validation, these are likely underestimates.

In addition to estimating the temporal trend in α, we also run experiments to see how other fac-

[6] https://github.com/Weixin-Liang/Mapping-the-Increasing-Use-of-LLMs-in-Scientific-Papers (MIT license)

tors, including author publication frequency, readability of abstract, and number of authors relate to ChatGPT usage. We describe the relation between α and these factors in the sub-sections below.

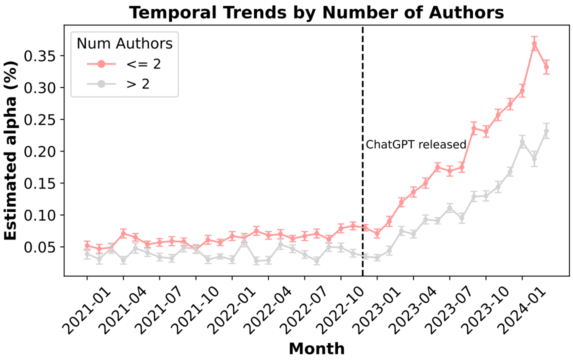


Figure 2: Evolution of estimated α over time for papers with 2 or fewer authors and more than 2 authors. α is higher for the former group.

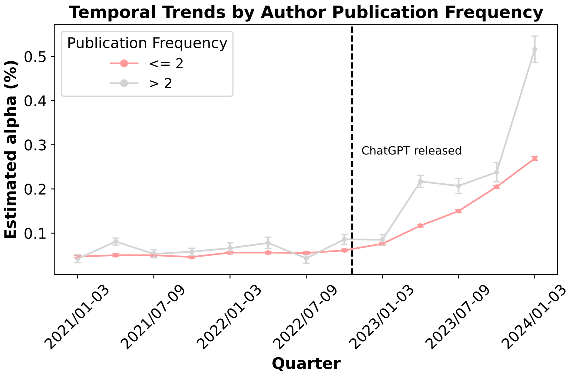


Figure 3: Evolution of estimated α over time for papers by (first) authors with 2 or fewer publications in our dataset and authors with more. Abstracts of papers by (first) authors who publish more frequently have a higher proportion of ChatGPT-modified sentences

5.1 Number of Authors

For a given month and year, we categorize published articles into two groups: those with 2 or fewer authors and those with more than 2 authors. We pick the threshold of 2 because it is the median number of authors per article. Figure 2 shows that alpha is higher in articles with 2 or fewer authors. One possible explanation for this observation is that the workload per author is higher in articles with fewer authors. To manage their workloads, especially as deadlines approach, single and pairs of authors may use LLMs. Additionally, peer pressure against using LLMs may be more pronounced on larger research teams.

5.2 Author Publication Frequency

We group the articles published in a given quarter into two groups. The first consists of articles written by first authors who had 2 or less publications in that quarter. The second group includes articles written by first authors with more than 2 publications in that period. Figure 3 shows that alpha is higher for abstracts written by authors who have higher publication frequency. Like the case for number of authors, deadline effects are more pronounced in authors who publish more. Therefore, to improve writing efficiency, frequently publishing authors may use LLMs more often. Similar finding is also observed in (Liang et al., 2024b). Computer Science publications by researchers who publish more often have higher proportion of ChatGPT modified texts.

5.3 Readability measures

To explore how the usage of ChatGPT contributes to the readability of the text, we calculate the readability scores of the abstracts. We consider three measures:

- Automated Readability Index (Smith and Senter, 1967):

$$ARI = 4.71\left(\frac{\#characters}{\#words}\right) + 0.5\left(\frac{\#words}{\#sentences}\right) - 21.43 \quad (5)$$

- Flesh-Kincaid Reading Level (Kincaid et al., 1975):

$$FKRL = 0.39\left(\frac{\#words}{\#sentences}\right) + 11.8\left(\frac{\#syllables}{\#words}\right) - 15.59 \quad (6)$$

- Coleman–Liau Index (Coleman and Liau, 1975): Let L denote the average number of letters per 100 words and S denote the average number of sentences per 100 words. Then,

$$CLI = 0.0588 \cdot L - 0.296 \cdot S - 15.8 \quad (7)$$

All three measures approximate the (US) grade level needed to comprehend the text. The results for ARI are displayed in Figure 13, and

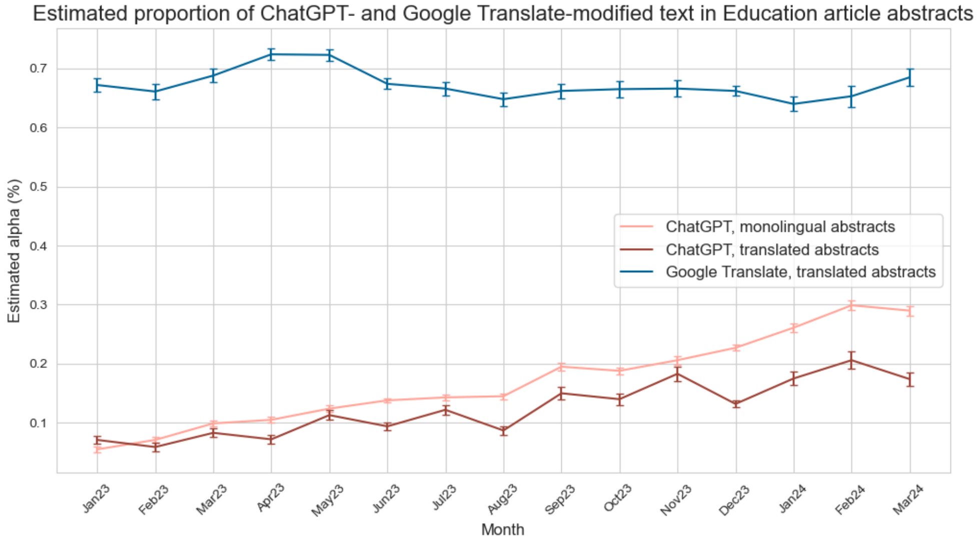


Figure 4: Evolution over time of estimated α for each model. Monolingual abstracts refer to papers where a single English abstract was provided. Translated refers to the group where both English and non-English language abstracts were present. The steady increase in α provides evidence that the usage of ChatGPT in Education research has increased over time.

show sentences with a reading level closer to graduate writing have a higher proportion of ChatGPT modified text. The findings are consistent across both Coleman-Liau and Flesh-Kincaid indexes (see Appendix). Additionally, we examined differential impact based on the average word length and number of words in each abstract. A higher share of AI-modified text was observed in abstracts with longer words, as well as in abstracts with an average word count below the median. Taken together, this evidence suggests that ChatGPT generates shorter sentences and longer words relative to the authors of our corpus, modifying text in a manner that approximates college reading levels. Qualitatively, this may correspond with fewer typos and malformed sentences, such as "run-on" sentences.

6 Discussion

LLM-modified scholarship arouses strong feelings with regard to academic integrity, and for good reason. We therefore suggest differentiating between LLM use cases and across contexts, specifically with regard to integrity. Accusations of cheating rest on assumptions about how people use AI – lazily, incompetently, to copy and fake. This is a matter of inauthentic literature. However, diligent usage of LLMs to translate article abstracts from non-English languages and then proofread them before publishing may not violate norms of academic integrity. This is rather a matter of facilitating the communication of research products. Insofar as we continue to see evidence of widespread LLM modification in scholarship, we are obliged to think critically about how AI usage maps back to core values of objectivity and rigor in science.

In this study, we draw a primary distinction between editing and translating, noting that both these actions can improve research communication. We conclude that ChatGPT usage lies mainly in writing, re-writing, and editing, rather than translating. We see consistent and relatively higher usage of Google Translate across translated contexts, with the exception of Indonesian contexts, where we see no evidence of AI tooling within the scope of our analysis (neglecting, for example, usage of DeepL for translation). That usage of ChatGPT and Google Translate appear to be complementary in the Education domain is perhaps unsurprising, given that educators may be more inclined to try out new or different AI tools after learning about them from students. However, the task-differentiated usage of LLMs we observe does contradict an assumed tendency to adopt a general-purpose, instruction-tuned AI model across use cases.

We use outputs from "gpt-3.5-turbo-0613" and "gpt-4-turbo-0613" to estimate ChatGPT's word distribution, because these now-deprecated mod-

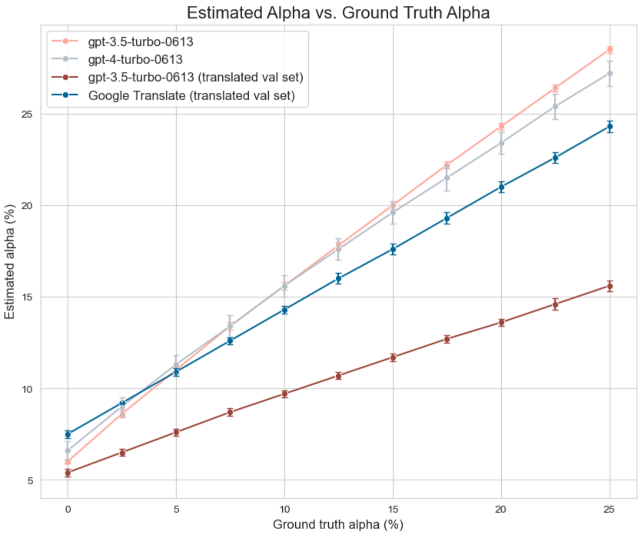


Figure 5: Comparison between ground truth and estimated values of α for each model

els were released June 13, 2023 and therefore correspond to the period covered by our test data. Unlike ChatGPT, Google Translate's development schedule and versioning are unclear. We made calls to the Google Translate API in June, 2024 and were mildly surprised to see the effectiveness of extending the present model's distribution back through the prior year. But we presume that training datasets for well-established machine learning products do not change so rapidly.

Crowd-sourced theories hold that ChatGPT's predilection for certain words (e.g., "delve"), which we confirm on a distributional basis, may be an artefact of reinforcement learning from human feedback (RLHF). However, Google Translate's relative overweighting of the words "jurisdiction" and "bibliographical" in our study are presumably not artefacts of RLHF. Distributional methods for estimating LLM mixture depend on the entire distribution of training data ingested by a model. With that said, we understand and sympathize with why people are drawn to discussing how direct human feedback shapes AI model behavior.

7 Limitations

Regarding data, we only train, validate, and test on the full text of research article abstracts. Abstracts are a standard and valued form of academic writing but necessarily do not convey the full scope of a given research article. However, if authors do in fact use ChatGPT and other LLMs to make their abstracts easier to read and comprehend – to improve the main "pitch" for their projects – then it is reasonable to assume that traces of these models can be found elsewhere in their articles. Moreover, policies intended to boost research output in Indonesia (Irawan et al., 2021) and Chile (Troncoso et al., 2022), for example, offer financial incentives for authors to have their work indexed by Web of Science, Scopus, and SciELO. This places a premium on research accessibility, of which writing and disseminating grammatically correct English-language abstracts is an important part.

Regarding methods, we observe that our models consistently underpredict on validation data at high ground truth α. Our results therefore likely underestimate the true proportion of ChatGPT- and Google Translate-modified text in the Education corpus. Furthermore, our predictions are noisier when we include possibly (LLM-)translated abstracts in the training data, suggesting that distributional quantification frameworks will become less robust as LLMs become more widely used and "human" data become more contaminated.

Regarding our findings, we reiterate that α represents a structural estimate of the proportion of LLM-modified text in a corpus of abstracts. It does not represent the proportion of abstracts (and therefore articles) modified by LLMs. This is a crucial distinction because we observe ChatGPT to be ridiculously verbose, such that AI-rewritten abstracts in our training and validation datasets are significantly longer than corresponding human-written abstracts. We see no evidence that Liang et al. (2024a,b)'s distributional quantification framework is sensitive to this imbalance.

Regarding the researchers whose work we test for AI modification, we wonder whether their continued co-writing and editing with LLMs might result in them learning to write more like LLMs (Li and Wang, 2023). In multilingual contexts, it is certainly possible that non-native English speakers will learn how to write in English as they co-create with chatbots (Kostka and Toncelli, 2023). Perhaps we will see *people* adopt habits of language use we assume in this study to be representative of AI.

Finally, and with full conviction as to the scientific character of research efforts supported by OJS and PKP around the world, we must also draw attention to the sheer variability in usage that comes with the globally distributed adoption of an open source publishing software. Relatively lower-resourced studies conducted in countries where OJS usage is highest are dogged by accusations of research impropriety, inadequate rigor, and non-reproducibility (Khanna and Willinsky, 2022). Acknowledging the fraught addition of LLMs to this dynamic may buoy claims made by detractors. We certainly do not want to fault researchers for using LLMs to improve the readability and accessibility of their work. However, there is a stark normative distinction between using LLMs to proofread or translate research, versus generating it altogether. We invite further studies that operationalize these kinds of usage distinctions.

8 Conclusion

When Google Translate was released to the learned public, did it constitute a threat to science in the way ChatGPT is typically assumed to? Or did it constitute a bridge between research cultures? If AI translation services, which still generate unverifiable information wholesale, are not assumed to threaten research integrity, then what is it about ChatGPT and generative AI that upsets researchers' ethical sensibilities? We suspect that readers will have intuitive answers to these questions, as we do. And with regard to those intuitions, we ask: how do they take use cases and context into consideration?

The proportion of AI-modified research in academic collections is growing, in Education and across the world. At the same time, the days in which a study can effectively isolate the influence of a specific AI model or two may already be waning. Understanding and making reasonable assumptions about use cases and contexts – narrowing focus, in other words, to specific domains and time periods and communities – will be vital to assessing impacts in the future.

We show only a temporary snapshot of LLMs' impact on open access Education research and are heartened, at least, that AI shows promise for improving scholarly communication across longstanding sociolinguistic boundaries. The corollary of improved communication in the style of English favored by contemporary academia, however, is linguistic homogenization. This is an issue of immense cultural and methodological importance. After having had the pleasure of working with such wonderfully multilingual and multicultural data, we wonder how authors in the future will balance their concerns about linguistic maintenance against pressures to publish in the dominant language of science and LLMs.

9 Acknowledgements

We thank Weixin Liang, Jing Huang, Diyi Yang, and John Willinsky for their comments and feedback. We thank Michael Felczak and Janio Sarmento at the Public Knowledge Project (PKP) for providing OJS data access. AUTHOR gratefully acknowledges support from the Stanford Institute for Economic Policy Research.

10 Authorship Statement

All authors worked together to design the research questions and experiments, analyze the data, and write the paper. AUTHOR retrieved the OJS data and liaised with PKP. Together, we take responsibility for the information of this article. We performed the analysis in good faith and did not use sources other than those listed in the bibliography.

11 Ethics Statement

Our usage of Open Journal Systems (OJS) data is compliant with the open access terms stipulated by journals using OJS. We have permission from the Public Knowledge Project (PKP) to download and analyze article records and their contents. We share all our results with colleagues at PKP, with the hope of informing services and outreach. We make no claims about individual researchers or their articles, nor statements about the quality thereof. We understand that our findings on the usage of LLMs in scholarship can influence public perception of academic integrity, as well as the integrity of these publications. Our work aims to highlight how Education researchers may use ChatGPT to improve their writing in concert with translation services and other AI tooling. We hope to further the discussion about how best to incorporate these tools into academic research and writing, with an understanding that there are significant qualitative differences between use cases like writing, co-writing, editing, and translating.

References

Bassel Almarie, Paulo E P Teixeira, Kevin Pacheco-Barrios, Carlos Augusto Rossetti, and Felipe Fregni. 2023. Editorial - the use of large language models in science: Opportunities and challenges. *Princ Pract Clin Res*, 9(1):1–4.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60:283–284.

Dorottya Demszky, Diyi Yang, David S. Yeager, Christopher J. Bryan, Margarett Clapper, Susannah Chandhok, Johannes C. Eichstaedt, Cameron A. Hecht, Jeremy P. Jamieson, Meghann Johnson, Michaela Jones, Danielle Krettek-Cobb, Leslie Lai, Nirel JonesMitchell, Desmond C. Ong, Carol S. Dweck, James J Gross, and James W. Pennebaker. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2:688 – 701.

Dasapta Erwin Irawan, Juneman Abraham, Rizqy Amelia Zein, Ilham Akhsanu Ridlo, and Eric Kunto Aribowo. 2021. Open access in indonesia. *Development and Change*, 52(3):651–660.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Saurabh Khanna, Jon Ball, Juan Pablo Alperin, and John Willinsky. 2022. Recalibrating the scope of scholarly publishing: A modest step in a vast decolonization process. *Quantitative Science Studies*, 3(4):912–930.

Saurabh Khanna and John Willinsky. 2022. What those responsible for open infrastructure in scholarly communication can do about possibly predatory practices. In *Predatory Practices in Scholarly Publishing and Knowledge Sharing*, pages 147–166. Routledge.

J. Peter Kincaid, Robert P. Jr. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Daphne Koller, Andrew Beam, Arjun Manrai, Euan Ashley, Xiaoxuan Liu, Judy Gichoya, Chris Holmes, James Zou, Noa Dagan, Tien Y. Wong, David Blumenthal, and Isaac Kohane. 2024. Why we support and encourage the use of large language models in <i>nejm ai</i> submissions. *NEJM AI*, 1(1):AIe2300128.

Ilka Kostka and Rachel Toncelli. 2023. Exploring applications of chatgpt to english language teaching: Opportunities, challenges, and recommendations. *TESL-EJ*, 27(3):n3.

Huiting Li and Yakun Wang. 2023. The empowerment and impact of chatgpt technology on foreign language education in colleges and universities. In *Proceedings of the 2023 6th International Conference on Educational Technology Management*, pages 29–34.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024a. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D Manning, and James Y. Zou. 2024b. Mapping the increasing use of llms in scientific papers.

Brady D. Lund, Ting Wang, Nishith Reddy Mannuru, Bing Nie, Somipam Shimray, and Ziang Wang. 2023. Chatgpt and a new academic reality: Artificial intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 74(5):570–581.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature.

Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. To protect science, we must use llms as zero-shot translators. *Nature Human Behaviour*, 7(11):1830–1832.

E. A. Smith and R. J. Senter. 1967. Probabilistic topic models. page 1–14.

Elizabeth Troncoso, Francisco Ganga-Contreras, and Margarita Briceño. 2022. Incentive policies for scientific publications in the state universities of chile. *Publications*, 10(2):20.

Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. 2020. Authorship attribution for neural text generation. In *Conference on Empirical Methods in Natural Language Processing*.

Chris Westfall. 2023. Educators battle plagiarism as 89% of students admit to using open ai's chatgpt for homework. *Forbes*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *ArXiv*, abs/1905.12616.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Appendix

A Open Journal Systems(OJS) Statistics

Languages	**2021**	**2022**	**2023**	**2024**
English	37.85%	39.25%	42.22%	45.72%
Multilingual (?? → EN)	25.49%	25.34%	23.50%	19.29%
Indonesian	12.55%	14.37%	14.33%	13.25%
Portuguese	9.45%	6.91%	5.46%	5.47%
Spanish	5.74%	5.04%	4.01%	3.55%
Russian	1.39%	1.26%	1.84%	2.10%
German	0.98%	1.22%	0.75%	0.84%
Multilingual (non-EN)	0.80%	0.85%	0.83%	0.66%
Uzbek	0.08%	0.85%	2.11%	3.44%
Chinese	0.96%	0.49%	0.64%	0.97%
Total Articles	**2,200,151**	**2,134,850**	**2,055,881**	**400,324**

Table 1: Proportion of OJS-published articles in a specific language, by year

Discipline	**2021**	**2022**	**2023**	**2024**
Medicine	11.54%	11.25%	10.82%	11.39%
Education	8.68%	9.47%	10.49%	10.35%
Business	4.59%	5.29%	5.92%	6.14%
Environmental Science	5.53%	5.56%	5.54%	5.66%
Engineering	3.30%	3.03%	3.12%	3.84%
Computer Science	2.41%	2.36%	2.62%	2.90%
Law	2.76%	2.83%	3.10%	2.86%
Agricultural and Food Sciences	3.29%	3.21%	3.14%	2.82%
History	3.67%	3.48%	3.14%	2.65%
Economics	2.22%	2.42%	2.44%	2.62%
Sociology	2.95%	2.97%	2.82%	2.50%
Linguistics	1.79%	1.91%	2.11%	2.01%
Political Science	2.13%	2.13%	2.06%	1.88%
Psychology	1.37%	1.56%	1.56%	1.45%
Philosophy	1.53%	1.49%	1.34%	1.23%
Art	1.47%	1.47%	1.32%	1.17%
Mathematics	1.01%	1.10%	1.17%	0.97%
Materials Science	0.75%	0.76%	0.76%	0.68%
Biology	0.76%	0.72%	0.63%	0.50%
Chemistry	0.55%	0.54%	0.53%	0.44%
Physics	0.40%	0.41%	0.47%	0.36%
Geology	0.25%	0.20%	0.24%	0.28%
Geography	0.38%	0.42%	0.36%	0.28%
Total Articles	**2,200,151**	**2,134,850**	**2,055,881**	**400,324**

Table 2: Proportion by fields of study of OJS-published articles with English abstracts, by year

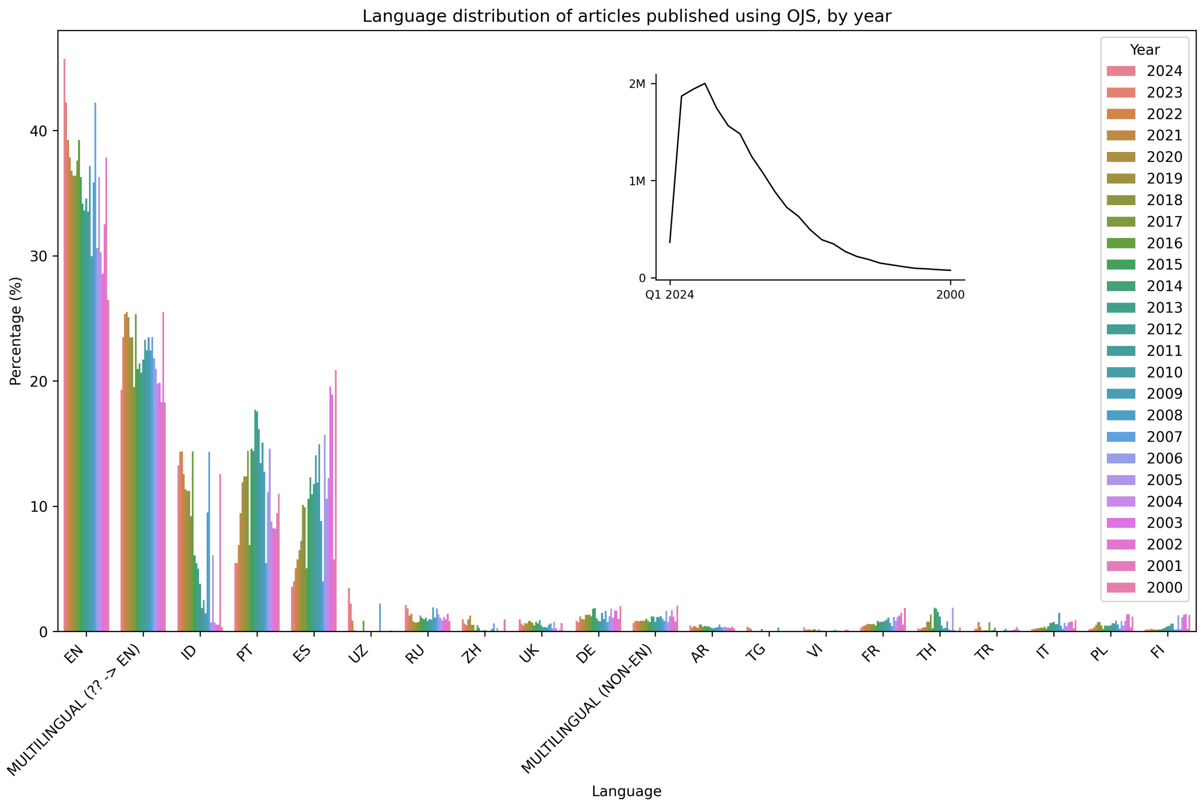


Figure 6: Percentage of OJS-published article abstracts in a given language, by year

* OJS went live in 2002, hence the articles dated 2000-2001 are digitized back issues.

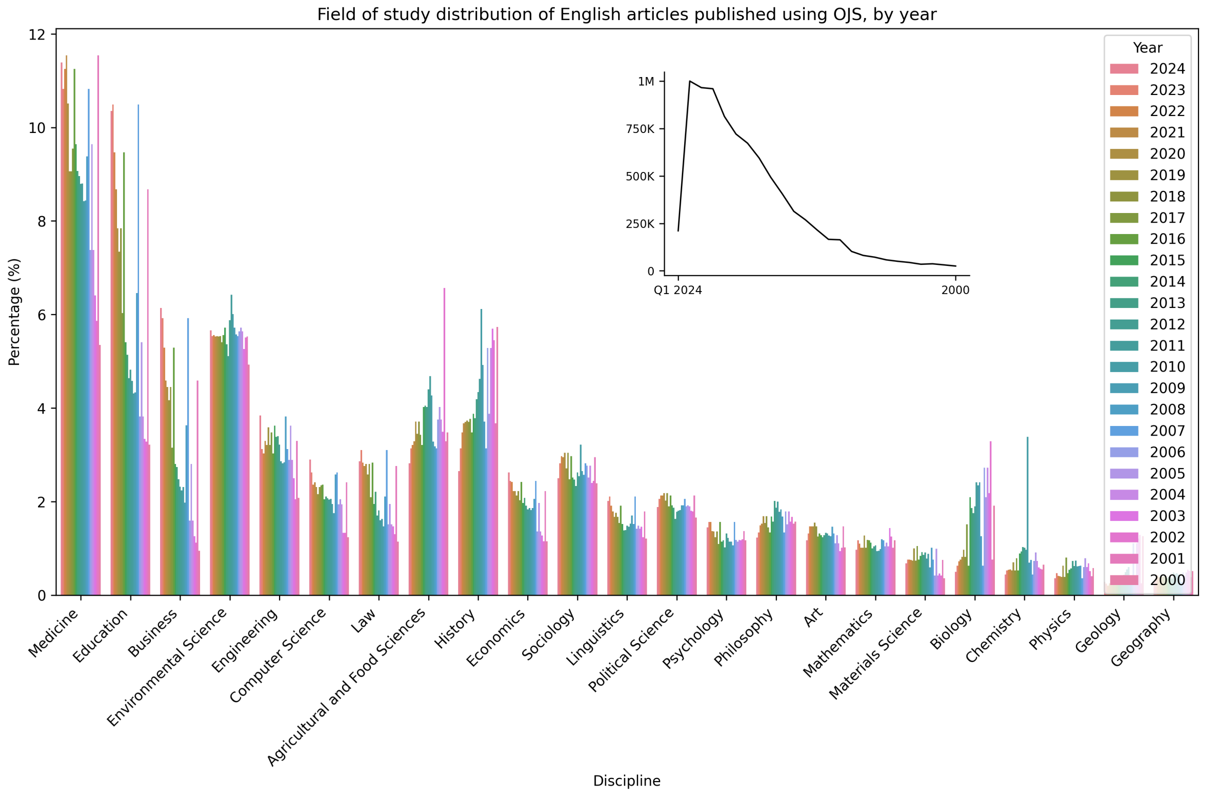


Figure 7: Percentage of OJS-published English articles by field of study, by year

B Additional Results

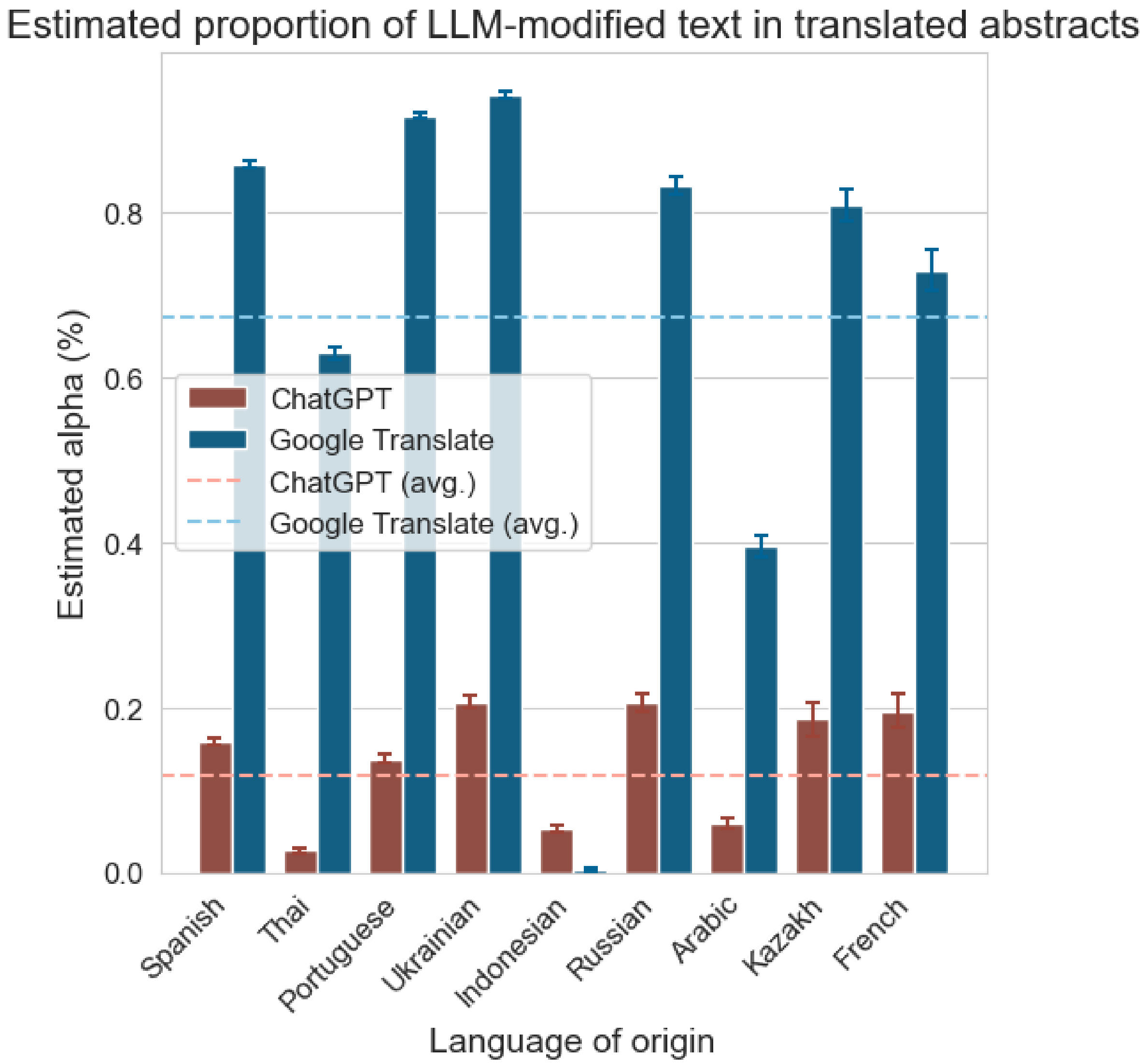


Figure 8: Estimated α by abstract language, comparing ChatGPT and Google Translate

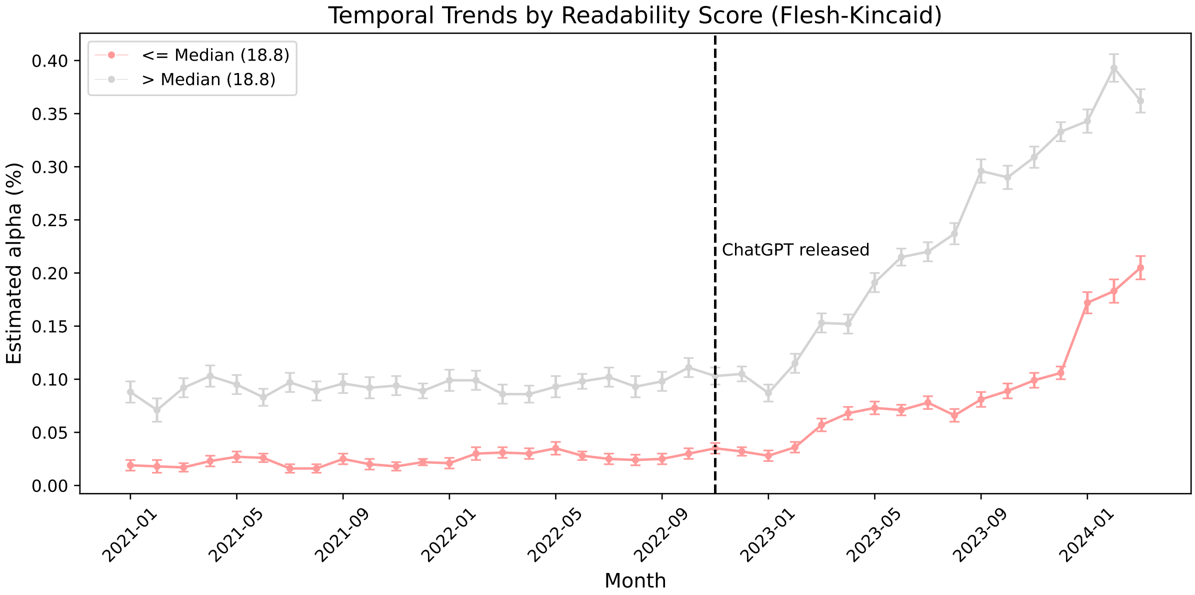


Figure 9: Evolution over time of estimated α for abstracts with a Flesh-Kincaid Index below and above the median.

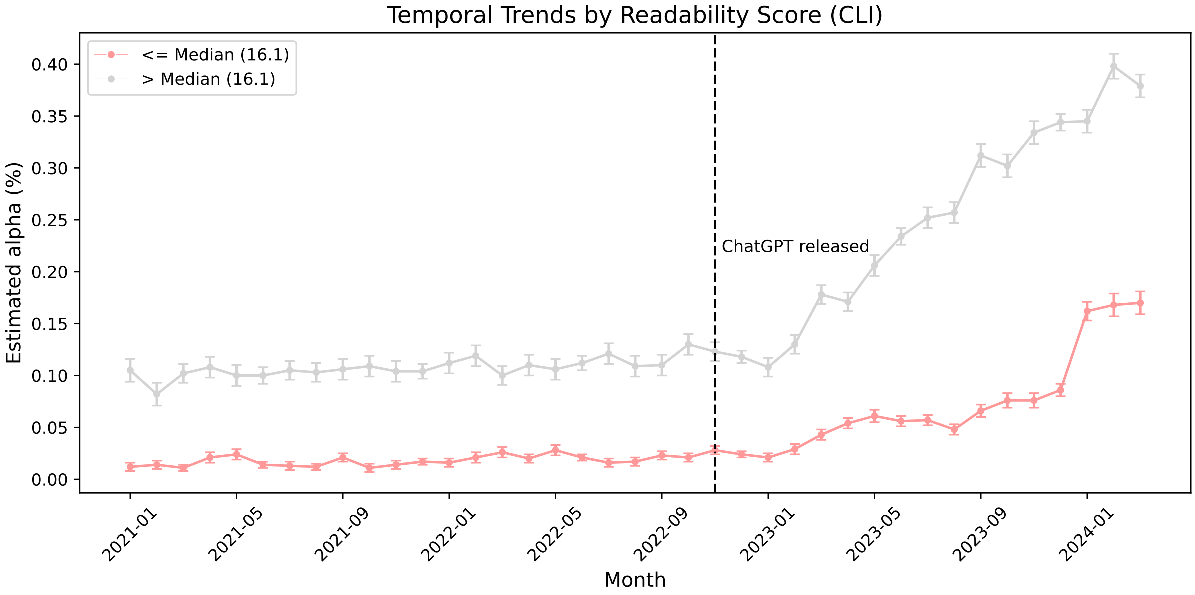


Figure 10: Evolution over time of estimated α for abstracts with a Coleman-Liau Index below and above the median.

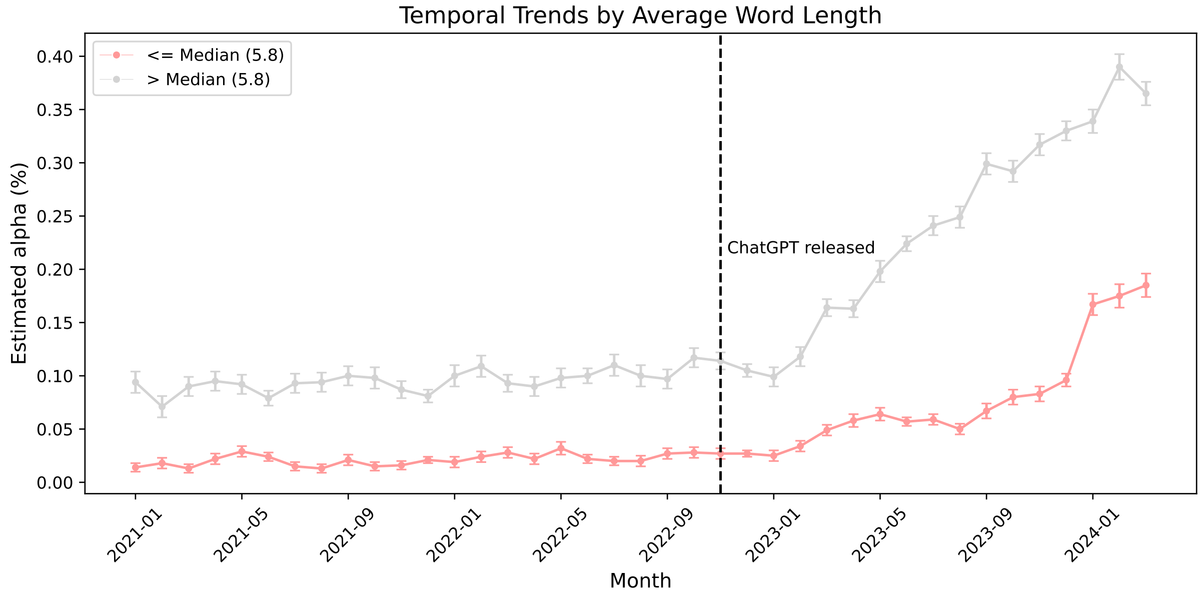


Figure 11: Evolution over time of estimated α for abstracts with an average word length below and above the median.

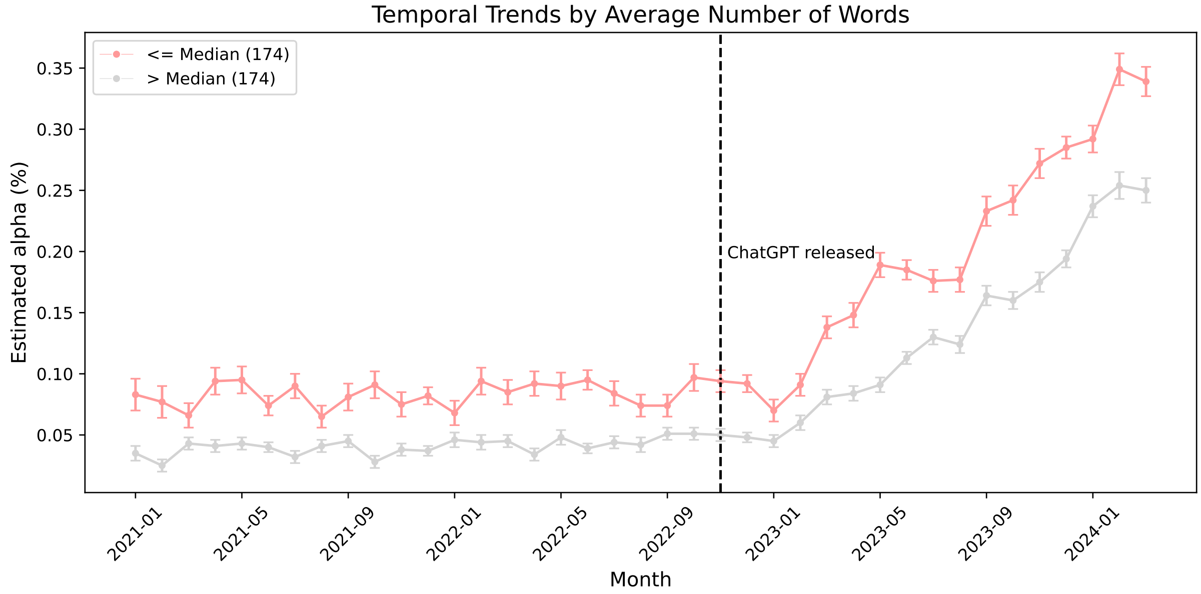


Figure 12: Evolution over time of estimated α for abstracts with a number of words below and above the median.

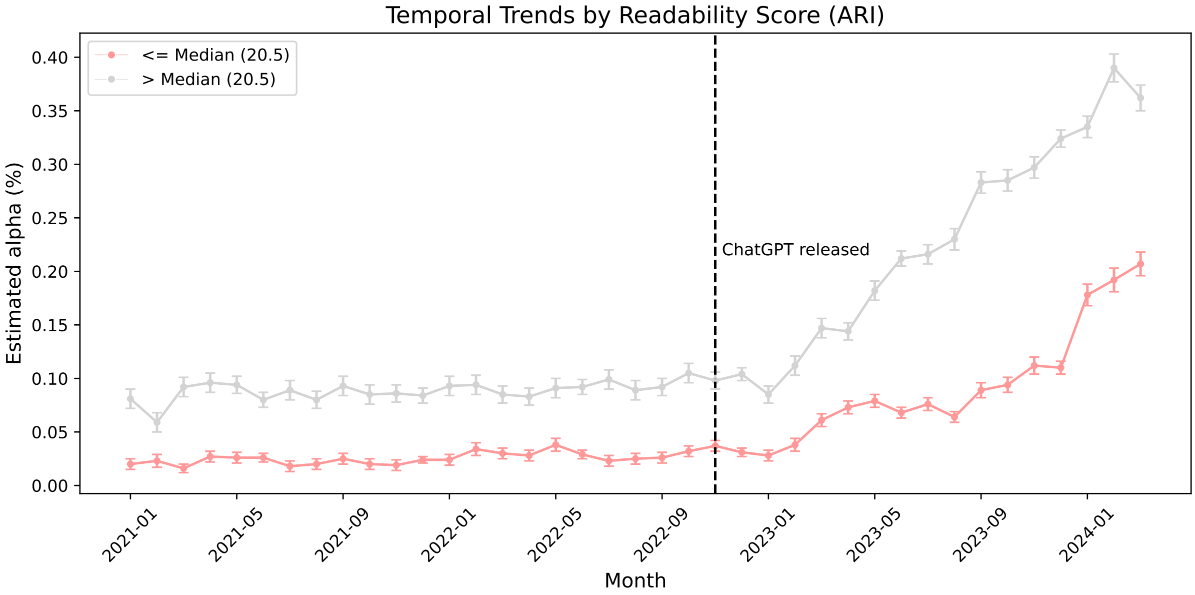


Figure 13: Evolution over time of estimated α for papers with an Automated Readability Index below and above the median. Abstracts of papers rated for reading levels of higher grades have a higher proportion of ChatGPT-modified sentences

Jinja LLM Prompting Templates

bullet.jinja

```
The aim here is to reverse-engineer the author’s writing process by taking the
    abstract from a paper and compressing it into a more concise form.
This process simulates how an author might distill their thoughts and key points
    into a structured, yet not overly condensed outline.
Here is the abstract:
{{human_abstract}}

Now, as a first step, summarize the goal of the article based on its abstract     e.
    g., what are its main scholarly or methodological contributions?
Then, refer to the text of the abstract itself, and reverse-engineer it into a list
    of concise bullet points:
```

write.jinja

```
Having reverse-engineered the author’s writing process by compressing the abstract
    of a paper, now we turn to rewriting the abstract.
Expand upon the concise, bulleted version you wrote:
{{bullets}}

This simulates how an author elaborates on the distilled thoughts and key points,
    arranging them into a detailed, structured narrative.
Given the concise outline above, develop it into a fully fleshed-out text abstract
    of no more than 400 words.
Write ONLY the full, grammatically correct text of the expanded abstract below:
```

translate.jinja

```
Translate the following Education research article abstract from ’{{language_code}}’
     to ’en’:
‘‘‘{{human_abstract}}‘‘‘

Given the abstract above, translate it into a fully fleshed-out English abstract of
    no more than 400 words.
Write ONLY the full, grammatically correct text of the English abstract below:
```