ChatGPT for Text Annotation? Mind the Hype!

Étienne Ollion*, Rubing Shen*†, Ana Macanovic‡§, Arnault Chatelain*

October 4, 2023

Abstract

In the past months, researchers have enthusiastically discussed the relevance of zero- or few-shot classifiers like ChatGPT for text annotation. Should these models prove to be performant, they would open up new continents for research, and beyond. To assess the merits and limits of this approach, we conducted a systematic literature review. Reading all the articles doing zero or few-shot text annotation in the human and social sciences, we found that these few-shot learners offer enticing, yet mixed results on text annotation tasks. The performance scores can vary widely, with some being average and some being very low. Besides, zero or few-shot models are often outperformed by models fine-tuned with human annotations. Our findings thus suggest that, to date, the evidence about their effectiveness remains partial, but also that their use raises several important questions about the reproducibility of results, about privacy and copyright issues, and about the primacy of the English language. While we definitely believe that there are numerous ways to harness this powerful technology productively, we also need to harness it without falling for the hype.

*Centre de Recherche en Économie et de Statistiques (CREST), IP Paris, Paris, France
†Médialab, Sciences Po, Paris, France
‡Department of Sociology, Utrecht University/ICS, Utrecht, Netherlands
§Centre for Complex Systems Studies, Utrecht University, Utrecht, Netherlands

1 Introduction

The release of ChatGPT during the Fall of 2022 has thrust "assistant-type" large language models, a certain type of AI models, into the limelight. Programmed to interact with humans using only natural language, these models can help their users with a variety of tasks: they search for or synthesize information, they perform repetitive tasks, they translate and even assist in writing texts. These generative AIs of a particular kind can also classify large volumes of textual data.

Should they do this latter task efficiently, these models could mark an important shift in the way we use textual data. At the end of the 2010s, the advent of pre-trained language models using transfer learning had marked a radical change in text annotation. Because this technique required considerably fewer human annotations than previous methods while performing well on standard benchmarks (Minaee et al., 2022), LLMs were massively adopted as annotation tools. Using them, social scientists could now, in a reasonable amount of time, train an efficient text classifier on virtually any annotation task (Do et al., 2022).

Yet, in order to function properly, LLMs still require training data, and sometimes in volumes that limit their use. By contrast, the newer, assistant-type LLMs need very limited guidance. They are trained to perform tasks with as little information as a single instruction (often called a prompt) offered in natural language (this is what is called "zero-shot learning"), or with just a few examples ("few-shot" or "in-context learning").

In the past year, a growing number of scholars have argued that these assistant-type LLMs can be successfully used in research. Recent work suggests that ChatGPT outperforms crowd workers on some conventional text-annotation tasks (Gilardi et al., 2023), and even experts on others (Törnberg, 2023). Since then, these models have been tested on a variety of cases, with positive results.

As researchers who have been working with text for years, often using AI, we found these promises extremely alluring. Yet, upon reading through the literature using ChatGPT for text-annotation, we were also disconcerted. We sometimes found a disconnect between the claims made and the results. We also noticed that the articles resort to diverse, and sometimes contradicting, standards to declare the superiority of these models.

We thus conducted a literature review on these few-shot classifiers when used for human and social scientific tasks. Our main conclusion is that, while these classifiers have much to contribute to our disciplines, we should guard ourselves against

unfettered enthusiasm. Our analysis shows that the new models do not always yield the best, and sometimes not even very solid results. They also expose rather consistent patterns of bias. Adding to a series of concerns that the use of these models raises for scientific purposes, we advocate for an audacious yet cautious use of these technologies in science.

2 Ambivalent Results

In the face of a rapidly growing but diverging body of literature, we conducted a systematic analysis of papers performing text classification with zero or few shots classifiers in the human and social sciences. At the time of writing, just for these disciplines, our search yielded 20 published articles or preprints assessing GPT-4.0 or ChatGPT's performance and consistency[1]. In the papers, the models were compared to various benchmarks, on a total of 117 classification tasks[2].

Overall, the papers find that the assistant-type LLMs classify text well. They often outperform old (pre-LLM) automatic annotation methods, and regularly come on a par with humans. Even considering a likely publication bias, whereby a larger share of positive results gets published, the results are undeniable. ChatGPT and kin models can annotate data with some success. How much? Here are the main lessons we draw from our review (see Table 1).

1. The first one is that **these few-shot classifiers are not, to date, the best models**. In most cases, models fine-tuned with dozens or hundreds of annotations are. When ChatGPT is compared to fined-tuned transformers, it performs better on 10 tasks (23%), equally on 3 (7%), and worse on 30 (70%). This should come as no surprise, since providing more examples to a classifier will, in all likelihood, improve the model performance. This fact is, in fact, often acknowledged by the proponents of assistant-type LLMs for text annotation. They rightfully argue that assistant LLMs offer a good trade-off given the limited investment they require. This, nonetheless, means that if performance is what one is after, only using assistant-LLMs may not be the best option: using fine-tuned transformers still is.

2. A second salient conclusion is that **the articles exhibit widely different performance scores**: some are very good, but some are only average, and quite

[1]We did so by searching for "LLM annotation", "ChatGPT annotation", "generative AI annotation" in Google Scholar, and curated a list of papers quantitatively evaluating ChatGPT on text classification tasks.

[2]Tasks assessed on different datasets are counted multiple times.

a few are really low. In a large-scale analysis, Pangakis et al. (2023) find that ChatGPT's success rates in annotating text vary between 0.06 and 0.97 (see also Ziems et al., 2023). The results displayed in Table 1 confirm these extreme variations. This suggests, at the very least, that these models cannot be used without a thorough validation procedure.

3. Another interesting pattern stems from this review. When detailed performance scores are reported (see Table 1), for 42 out of 59 relevant labels (71%), **chatGPT and kin models yield a higher recall than precision**, indicating a tendency to output more false positives than false negatives. This could suggest the presence of an acquiescence bias (or "yea-saying") in ChatGPT's annotation. Once again, this calls for a thorough human validation, but also points towards an opportunity: these models could be used to create a first set of annotations which, after being reviewed by humans, could help fine-tune supervised models in a rapid and efficient way.

4. Our final finding is that **the metrics used to evaluate ChatGPT vary greatly from text to text**. Some focus on model performance and others on consistency between annotators, almost each time with different indicators. While this diversity makes it possible to capture different dimensions of the model's behavior, it also limits our ability to compare between experiments, and it could, in the worst case, favor cherrypicking – the selection of the most favorable output. Collectively, we would certainly benefit from harmonizing the way we present these results.

We would also gain from comparing our models to qualified annotators, and not to crowdworkers, as is often the case in the papers we reviewed. In the social sciences, the use of gig workers has become common in the last years, as they allow researchers to outsource menial yet time-consuming tasks. Yet, the quality of the annotation of these often untrained and meagerly paid crowd workers has been increasingly questioned (Marquardt et al., 2017). While some authors took measures to ensure the quality of their human annotations, not all did. Maybe it is time to raise our standards.

3 Zero-Shot, but at What Cost?

Overall, our review suggests that we should be careful before we declare assistant-type LLMs fit for automatic text classification. These are not the only aspects that call for caution when using ChatGPT for text annotation. Using these models raise a series of concerns we need to be aware of.

Table 1: Review of papers assessing ChatGPT's capability for text classification

Paper	Field of Study	Nb. Tasks	Assessment	Evaluation Scores		Acq. Bias	Comparing to	Better *#/tasks*	On par *#/tasks*	Worse *#/tasks*
Amin et al. (2023)	Psychology	3	Performance	*ACC*	0.45–0.93		f.-t. transf.		1/3	2/3
				UAR	0.48–0.91		SVM		3/3	
Gilardi et al. (2023)	Political Science	11	Performance	*ACC*	0.38–0.87		Crowdworkers	8/11		3/11
			Consistency	α	0.50–1.00		Crowdworkers	11/11		
Heseltine & Clemm Von Hohenberg (2023)	Political Science	16	Performance	*ACC*	0.72–0.95					
				F1	0.72–0.95					
Huang et al. (2023)	Computational Social Science	1	Performance	*ACC*	0.80					
Kuzman et al. (2023)	Linguistics	3	Performance	*ACC*	0.67–0.91		f.-t. transf.	1/3		2/3
				F1	0.56–0.91					
Li et al. (2023)	Communication	3	Performance	*ACC*	0.75–0.87	4/6				
				F1	0.39–0.61					
			Consistency	α	0.90–0.98					
Mellon et al. (2023)	Political Science	1	Performance	*ACC*	0.86–0.95		RAs			1/1
							SVM	1/1		
Mets et al. (2023)	Computational Social Science	1	Performance	*F1*	0.65		f.-t. transf.		1/1	
Møller et al. (2023)	Computational Social Science	3	Performance	*ACC*	0.38-0.86	8/9	f.-t. transf.	1/3	1/3	1/3
				F1	0.32-0.72					
Pangakis et al. (2023)	Political Science, Psychology	27	Performance	*ACC*	0.67–0.98	20/27				
				F1	0.06–0.97					
Rathje et al. (2023)	Psychology	14	Performance	*F1*	0.30–0.78	5/8	f.-t. transf.			13/13
							Naïve Bayes	1/1		
Reiss (2023)	Communication	1	Consistency	α	0.50–0.95					
Rytting et al. (2023)	Political Science	3	Performance	*ACC*	0.79		Crowdworkers		2/3	1/3
				ICC	-0.51–0.74		SVM			1/1
				JPA	0.47–0.63					
				κ	0.48–0.61					
Savelka et al. (2023)	Law	1	Performance	*ACC*	0.29–0.55		Experts		1/1	
				F1	0.37–0.57					
				α	0.19–0.53					
Törnberg (2023)	Political Science	1	Performance	*ACC*	0.90–0.94		Experts & Crowdworkers	1/1		
			Consistency	α	0.95–0.98		Experts & Crowdworkers	1/1		
Wu et al. (2023)	Political Science	1	Performance	r	0.92–0.96					
			Consistency	r	1.00					
Yang & Menczer (2023)	Computational Social Science	1	Performance	ρ	0.51-0.62					
				AUC	0.89					
				F1	0.63-0.73					
Yu et al. (2023)	Linguistics	1	Performance	*ACC*	0.50 *chatgpt* 0.84-0.93 *bing*		RAs			1/1
Ziems et al. (2023)	Computational Social Science	20	Performance	*F1*	0.15–0.72		f.-t. transf.	8/20		12/20
				κ	0.01–0.64					
Zhu et al. (2023)	Computational Social Science	5	Performance	*ACC*	0.57–0.65	5/9				
				F1	0.55–0.65					

Notes: For each paper, we provide (columns): the main field of study; the number of classification tasks on which a generative classifier is assessed (tasks assessed on different datasets are counted multiple times); which aspect of the annotation is assessed (performance/consistency); reported evaluation scores in various classification metrics (Accuracy (ACC), F1-score (F1), Unweighted Average Recall (UAR), Area Under the Curve (AUC)), agreement metrics (Kripendorff's alpha (α), Fleiss' kappa (κ), Inter-Class Correlation (ICC), Joint-Probability of Agreement (JPA)) and correlation metrics (Pearson's r (r), Spearman's rho (ρ)); the presence of acquiescence bias (acq. bias): fractions represent number of labels of interest for which recall is higher than precision (blank: detailed scores not reported); results of comparison with human annotators (crowdworkers, experts, research assistants (RAs)and with other annotation models (fine-tuned transformers (f.-t. transf.), naïve bayes, support vector machine (SVM)). Fractions represent the proportions of tasks on which ChatGPT is reported to perform better than/equal with/worse than what it is compared to (blank: no such comparison was made).

Towards a Reproducibility Crisis?

One possible problem is the limited robustness of these methods. Some indicate that minor changes in the wording of the prompt or the requested annotation format (e.g., requesting a binary label versus a probability), can cause large inconsistencies in ChatGPT's performance (Reiss, 2023; Savelka et al., 2023; Li et al., 2023). Yet, others find that changes to instructions minimally affect model performance and reliability (Pangakis et al., 2023; Rytting et al., 2023). At best, this mixed evidence suggests that researchers need to be very mindful of, and transparent about, their prompting strategies.

More problematic is the lack of control by researchers over the tool they use to annotate, which raises serious questions when it comes to replicating the results. There is, of course, the classic criticism of the "black box" these models constitute, because they lack transparency on key parameters (such as training data, weights, model architecture, code), or in the additional safety mechanisms proposed in the chat environment (Liesenfeld et al., 2023). Worse, in our view, is the fact that different versions of a model yield different results as the models are updated. This situation is all the more critical because companies such as OpenAI tend to deprecate older models, making reproducibility virtually impossible[3].

Privacy and Copyright Issues

A second matter of concern is that only certain data can be analyzed using ChatGPT or a kin commercial solution, as this practice raises questions about privacy and intellectual property. Arguably, OpenAI, the firm that commercializes ChatGPT, claims that it does not "use content that you provide to or receive from our API [...] to develop or improve services" [4]. But this does not mean that they won't do so in the future, or in another way.

If the data one wants to classify is protected by intellectual property laws, it should not be transmitted to the platform in the first place. In fact, the authors of a large-scale study that used articles from the *New York Times* were forced to conduct it on the title only, as the rest of the text was "not available in the public data" (Rytting et al., 2023, p.11). The texts we need to annotate can also raise privacy issues. In the social sciences, they can consist of open-ended questions in surveys containing potentially identifying or personal information, they can describe medical conditions.

[3] https://platform.openai.com/docs/deprecations (Retrieved on August, 26th 2023).
[4] https://openai.com/policies/terms-of-use (Retrieved on August, 26th 2023).

This only furthers the recent calls for open-source generative AI models (Spirling, 2023).

Do We Want an Even More English-Centered Research?

One last concern has to do with the English bias of these LLMs. As researchers who sometimes work in languages different from English, we cannot but notice variations in the performances of the models across languages. Several papers report that assistant-LLMs perform best in English (Lai et al., 2023; Ahuja et al., 2023) and display rather poor performance in some low-resource languages (Kuzman et al., 2023; Rathje et al., 2023). Others confirm this tendency by suggesting to either prompt the model in English first, or ask it to translate the prompt into English in order to get better results (Zhao et al., 2023; Etxaniz et al., 2023). This situation will certainly evolve in the future, as LLMs get trained into specific languages. Yet such an observation is puzzling, as the inequalities between languages will likely persist given the differential investments made by companies or governments. This could in turn lead to an increased attention to English corpora, at the expense of other objects and sites of study.

4 Conclusion

Let us be clear: we are excited about the current technological developments, and we do use LLMs (including assistant-type LLMs) in our research. We are also optimistic that they could help reduce some inequalities in science by offering affordable ways to annotate text, thus granting access to textual resources to more researchers across the globe.

The dazzling progress made by these models should nonetheless not conceal their potential flaws and limitations. Being oblivious about them would even backfire, as articulated by Bail (2023). As we explore the possibilities they afford, we should heed the lessons of the past. There have, indeed, been other moments of excitement for textual data. The lessons are many, but we should always keep in mind that using quantitative analysis should not dispense with reading our corpora, comparing methods, and validating our results (Grimmer & Stewart, 2013). We also need to collectively design procedures that allow us to fruitfully use these technologies without outsourcing our judgment to them entirely.

Acknowledgement

The authors would like to thank Chris Bail, Julien Boelaert, Måns Magnusson, Patrick Präg for helpful comments.

References

Ahuja, K., Diddee, H., Hada, R., Ochieng, M., Ramesh, K., Jain, P., Nambi, A., Ganu, T., Segal, S., Axmed, M., Bali, K., & Sitaram, S. (2023). MEGA: Multilingual evaluation of generative ai. *(preprint) arXiv:2303.12528*.

Amin, M. M., Cambria, E., & Schuller, B. W. (2023). Will affective computing emerge from foundation models and general ai? a first evaluation on chatgpt. *(preprint) arXiv:2303.03186*.

Bail, C. A. (2023). Can generative AI improve social science? *(preprint) SocArXiv:rwtzs*.

Do, S., Ollion, É., & Shen, R. (2022). The augmented social scientist: Using sequential transfer learning to annotate millions of texts with human-level accuracy. *Sociological Methods & Research*, (pp. 00491241221134526).

Etxaniz, J., Azkune, G., Soroa, A., de Lacalle, O. L., & Artetxe, M. (2023). Do multilingual language models think better in English? *(preprint) arXiv:2308.01223*.

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.

Grimmer, J. & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3), 267–297.

Heseltine, M. & Clemm Von Hohenberg, B. (2023). Large language models as a substitute for human experts in annotating political text. *(preprint) SocArxiv:cx752*.

Huang, F., Kwak, H., & An, J. (2023). Is ChatGPT better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *(preprint) arXiv:2302.07736*.

Kuzman, T., Ljubešić, N., & Mozetič, I. (2023). ChatGPT: beginning of an end of manual annotation? use case of automatic genre identification. *(preprint) arXiv:2303.03953*.

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). ChatGPT beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *(preprint) arXiv:2304.05613*.

Li, L., Fan, L., Atreja, S., & Hemphill, L. (2023). "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *(preprint) arXiv:2304.10619*.

Liesenfeld, A., Lopez, A., & Dingemanse, M. (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* (pp. 1–6). Eindhoven Netherlands: ACM.

Marquardt, K. L., Pemstein, D., Sanhueza, C., Petrarca, B. S., Wilson, S. L., Bernhard, M., Coppedge, M., & Lindberg, S. I. (2017). Experts, coders, and crowds: An analysis of substitutability. *V-Dem Working Paper*, 53.

Mellon, J., Bailey, J., Scott, R., Breckwoldt, J., & Miori, M. (2023). Does GPT-3 know what the most important issue is? using large language models to code open-text social survey responses at scale. *SSRN preprint: 4310154*.

Mets, M., Karjus, A., Ibrus, I., & Schich, M. (2023). Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media. *(preprint) arXiv:2305.13047*.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2022). Deep learning-based text classification: a comprehensive review. *ACM Computing Surveys*, 54(3), 1–40.

Møller, A. G., Dalsgaard, J. A., Pera, A., & Aiello, L. M. (2023). Is a prompt and a few samples all you need? using GPT-4 for data augmentation in low-resource classification tasks. *(preprint) arXiv:2304.13861*.

Pangakis, N., Wolken, S., & Fasching, N. (2023). Automated annotation with generative AI requires validation. *(preprint) arXiv:2306.00176*.

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., & Van Bavel, J. J. (2023). GPT is an effective tool for multilingual psychological text analysis. *(preprint) PsyArXiv:sekf5.*

Reiss, M. V. (2023). Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *(preprint) arXiv:2304.11085.*

Rytting, C. M., Sorensen, T., Argyle, L., Busby, E., Fulda, N., Gubler, J., & Wingate, D. (2023). Towards coding social science datasets with language models. *(preprint) arXiv:2306.02177.*

Savelka, J., Ashley, K. D., Gray, M. A., Westermann, H., & Xu, H. (2023). Can GPT-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *(preprint) arXiv:2306.13906.*

Spirling, A. (2023). Why open-source generative AI models are an ethical way forward for science. *Nature*, 616(7957), 413–413.

Törnberg, P. (2023). ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *(preprint) arXiv:2304.06588.*

Wu, P. Y., Nagler, J., Tucker, J. A., & Messing, S. (2023). Large language models can be used to scale the ideologies of politicians in a zero-shot learning setting. *(preprint) arXiv:2303.12057.*

Yang, K.-C. & Menczer, F. (2023). Large language models can rate news outlet credibility. *(preprint) arXiv:2304.00228.*

Yu, D., Li, L., & Su, H. (2023). Using LLM-assisted annotation for corpus linguistics: A case study of local grammar analysis. *(preprint) arXiv:2305.08339.*

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., & Wen, J.-R. (2023). A survey of large language models. *(preprint) arXiv:2303.18223.*

Zhu, Y., Zhang, P., Haq, E.-U., Hui, P., & Tyson, G. (2023). Can ChatGPT reproduce human-generated labels? a study of social computing tasks. *(preprint) arXiv:2304.10145.*

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z., & Yang, D. (2023). Can large language models transform computational social science? *(preprint) arXiv:2305.03514.*