Supervised Text Classification with LLM-Generated Training Labels

Anonymous ACL submission

Abstract

Computational social science practitioners often rely on human data to train supervised classifiers for text annotation. We assess the potential for researchers to augment or replace human-generated training data with synthetic training labels from generative large language models (LLMs). We introduce a recommended workflow and test this LLM application by measuring performance by replicating 14 classification tasks. We employ a novel corpus of English-language text classification data sets from recent computational social science articles in high-impact journals. Because these data sets are stored in password-protected archives, our analyses are less prone to issues of contamination. For each task, we compare supervised classifiers fine-tuned using GPT-4 labels against classifiers trained with human annotations and against GPT-4 few-shot labels. Our findings indicate that supervised classification models trained on LLM-generated labels perform comparably to models trained with labels from human annotators. Training models using LLM-generated labels is a fast, efficient and cost-effective method of building supervised text classifiers.

1 Introduction

Supervised text classification often relies on human-labeled text data for training and validation. Computational social scientists frequently use these types of supervised models to classify large quantities of text, ranging from news articles on the internet to government documents (Grimmer et al., 2022; Lazer et al., 2020). Collecting training and validation labels generated by humans for these tasks, however, is expensive, slow, and prone to a variety of errors (Grimmer and Stewart, 2013; Neuendorf, 2016).

To address these limitations, prior research suggests utilizing few-shot capabilities of generative large language models (LLMs) to annotate text data instead of human annotators (Brown et al., 2020; Gilardi et al., 2023; Wang et al., 2021; Ziems et al., 2023). Generative LLMs are faster and cheaper than human annotators and do not suffer from common human challenges such as limited attention span or fatigue. Although this approach has its limitations (Ollion et al., 2023) and generative LLMs do not excel at all text annotation tasks (Pangakis et al., 2023), past work illustrates that there are numerous circumstances where generative LLMs can produce high quality text-annotation labels.

Although past work suggests generative LLM few-shot annotation is highly effective, it may be cost prohibitive in many settings. Computational social science often involves classifying millions of documents or text samples. For example, a recent computational social science article studies a data set of 6.2 million tweets labeled on four dimensions (Hopkins et al., 2024), a task that would have cost over $25,000 if using GPT-4 alone. Using a knowledge distillation approach (Gou et al., 2021; Dasgupta et al., 2023), it may be possible to approximate the performance of a larger "teacher" model (e.g., GPT-4, estimated to have over 1.7T parameters (OpenAI, 2023)) with much smaller and cheaper task-specific "student" models (e.g., BERT Base, approximately 110 million parameters).

In this paper, we evaluate the feasibility of using generative LLMs to create synthetic labels for training downstream supervised classification models. Our approach involves first using a generative LLM to label a subset of text samples and then training a series of supervised text classifiers with the LLM-generated labels. We introduce a novel strategy to measure noise in LLM few-shot labels and isolate high quality labels for use as training data. Using our outlined approach, we assess performance across ten different models by replicating 14 classification tasks. In addition to a GPT-4 few-shot model, we assess performance between popular supervised classifiers (i.e., BERT, RoBERTa, and

DistilBERT) trained on varying quantities of either human-labeled samples or GPT-4-labeled samples.

A small number of studies have utilized similar approaches in related domains. Chen et al. (2023) use ChatGPT annotations to train various Graph Neural Networks (GNNs) for a fraction of the cost of human annotations. Golde et al. (2023) also harness ChatGPT to create synthetic text data that aligns with a specific valence (i.e., positive and negative) and then subsequently fine-tune a supervised classifier using the synthetic text. Most analogous to our approach here, Wang et al. (2021) train RoBERTa (Liu et al., 2019) and PEGASUS (Zhang et al., 2020) models on labels generated by GPT-3. Despite strong performance across their analyses, Wang et al. (2021), as well as the previously mentioned studies, exclusively evaluate closed-source models (i.e., GPT-3 and ChatGPT) on popular, publicly available NLP benchmark tasks (e.g., AG-News, DBPedia, etc), which are plausibly included in the training data for the generative LLM. As a result, these analyses provide an unclear indication of performance because their results plausibly suffer from contamination. Put otherwise, strong performance may reflect memorization, which casts doubt on the generalizability of the findings.

To compare supervised classifiers trained using LLM-generated labels against those trained with labels from human annotators, researchers must assess performance on tasks and data less likely to be affected by contamination. To this end, all 14 of the classification tasks we replicate are conducted on data sets stored in password-protected archives. Each of the classification tasks in our corpus are real applications in computational social science and contain human-labeled annotations that we consider as ground-truth.[1] Because our data come from non-public data sets from recently published academic journals, our findings are less prone to concerns of leakage and contamination.

Our main contributions are as follows:

1. Across 14 classifications tasks, supervised models trained with GPT-generated labels perform comparably to models trained with human-labeled data. Specifically, the median F1 performance gap between models trained using GPT-labels and models trained on human-labeled data is only 0.039. While supervised classifiers trained with LLM-generated labels perform slightly worse than classifiers trained with human labels, LLM-generated labels are a fast, efficient and cost-effective method to fine-tune supervised text classifiers.

2. Supervised models trained on GPT-generated labels perform remarkably close to GPT few-shot models, with a median F1 difference of only 0.006 across the classification tasks.

3. GPT few-shot models and supervised models trained on GPT-generated labels perform significantly better than all other models on *recall*, but noticeably worse on *precision*.

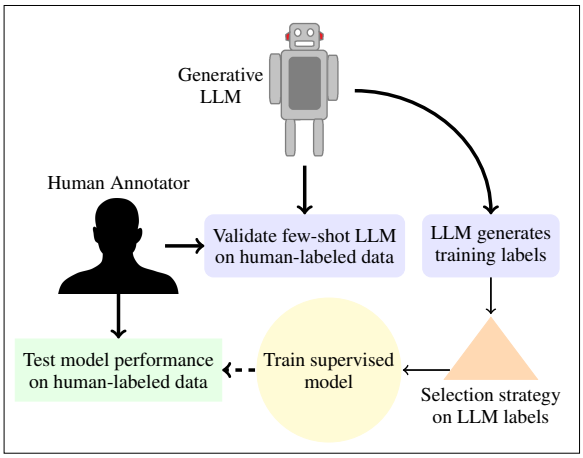

Figure 1: Supervised text classification with LLM-generated training labels.

2 Methodology

Figure 1 shows our five step workflow. First, we validate LLM few-shot performance against a small subset (n=250) of human-labeled text samples for each task. We provide GPT-4 with detailed instructions to label the text samples into conceptual categories outlined in the original study.[2] Because LLM few-shot annotation performance varies across tasks and data sets, validation is always necessary (Pangakis et al., 2023). We then fine-tune the prompt to optimize performance on this initial sample.[3] Using the validated prompt, the second step involves labeling an additional 1,250 text samples per task using the same generative LLM, which will later be used as training data for the supervised classifier.

[1] Table A1 and Table A2 include a full list of the data sets and classification tasks.

[2] We selected GPT-4 because it was the highest performing model at the time of our analyses.

[3] We include all prompt details in the supplementary material. Additional prompt tuning details and analyses are discussed in Appendix B.

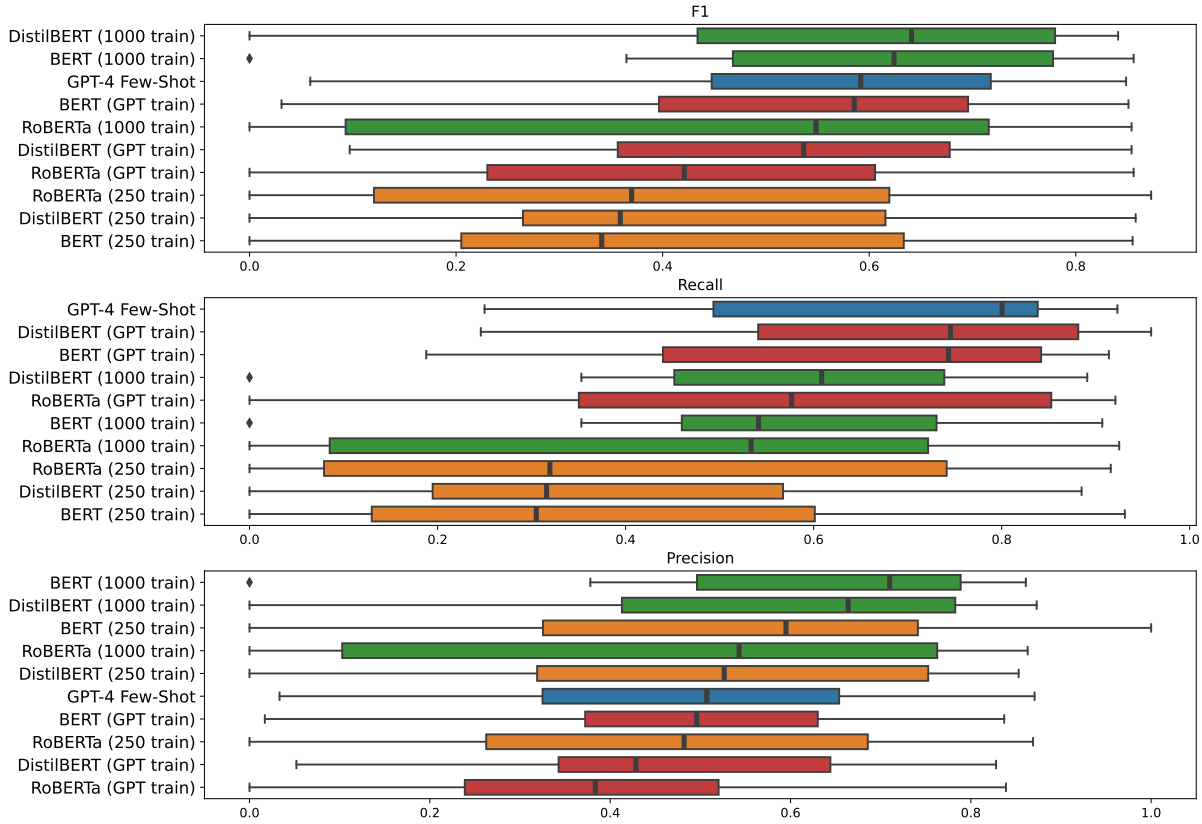

Figure 2: Box plots of performance on test data across 14 tasks. Thick vertical line denotes median.

.

Third, we implement a novel selection strategy to only sample training labels with the highest probability of correct classification (see also Bansal and Sharma, 2023). For our approach, we exploit the generative LLM's predicted token sampling process to identify higher confidence annotations. By inducing randomness in the LLM sampling process through the use of the temperature hyperparameter and by repeating an annotation task on the same text sample, we generate an empirical measure of uncertainty in the label that we deem a "consistency score."[4] Given a vector of classifications, C, with length l for a given classification task, *consistency* is measured as the proportion of classifications that match the modal classification: $\frac{1}{l}\sum_{i=1}^{l} C_i == C_{mode}$.

For our analyses, we classify every text sample five times at a temperature of 0.7 and only keep annotations with a consistency score of 1.0. Put otherwise, we only retain annotations where GPT-4 consistently labeled the same category across all iterations. Across all analyzed tasks, classifications with a consistency of 1.0 show significantly higher accuracy (19.4% increase), true positive rate (16.4% increase), and true negative rate (21.4% increase) compared to classifications with a consistency less than 1.0. Roughly 85% of classifications had a consistency of 1.0, which reduced our training set to slightly more than 1000 samples per task.

In the fourth and fifth steps, we trained a variety of supervised text classifiers and assessed performance against a held-out set of 1000 human-labeled samples. Our supervised text classification models include BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DistilBERT (Sanh et al., 2019). We select these models because of their frequent application in computational social science. For each task-specific supervised classifier, we conduct a grid search to optimize performance, training 18 models and select the combination of hyperparameters that yield the best F1 performance.[5] Ultimately, we compare performance between text classifiers trained on 1000 LLM-generated samples, 250 human-labeled samples, and 1000 human-labeled samples.

[4]Accessing token log probabilities, once available, will be an effective way to do the same type of selection approach.

[5]We optimize the learning rate, the batch size, and the number of epochs. We elaborate on this process in Appendix B.2.

Model	Training data	Accuracy	F1	Precision	Recall
GPT-4	Few shot	0.88	0.59	0.51	0.80
BERT	Human annotation: 250	0.89	0.34	0.59	0.30
	Human annotation: 1000	0.92	0.62	0.71	0.54
	GPT-4 annotation: 1000	0.87	0.59	0.50	0.74
DistilBERT	Human annotation: 250	0.89	0.36	0.53	0.32
	Human annotation: 1000	0.89	0.64	0.66	0.61
	GPT-4 annotation: 1000	0.85	0.54	0.43	0.75
RoBERTa	Human annotation: 250	0.88	0.37	0.48	0.32
	Human annotation: 1000	0.90	0.55	0.54	0.53
	GPT-4 annotation: 1000	0.84	0.42	0.38	0.58

Table 1: Comparison of classification performance on held-out validation data. Median performance across 14 tasks shown.

3 Results

Classification results are shown in Table 1. In Figure 2, each box plot displays the range of evaluation metrics across all 14 tasks for a given model/training data combination. The thick vertical line denotes the median performance metric. Across all 14 classification tasks, DistilBERT and BERT trained on 1000 human-samples are the highest performing models, with a median F1 score of 0.641 and 0.624, respectively.[6] Not far behind, however, is the GPT-4 few-shot model (0.592 median F1) and BERT trained on 1000 GPT-labeled samples (0.586 median F1). From this we draw two conclusions: First, models trained on few-shot synthetic labels from a generative LLM perform comparably to models trained on human labels. Despite a small performance gap, training supervised models on LLM-labeled data may be a quick, effective, and budget-friendly approach for constructing supervised text classifiers.

Second, models trained on synthetic labels from GPT-4 demonstrate very similar validation performance as few-shot labels with GPT-4. As each additional GPT-4 query incurs more expense, researchers can save resources by avoiding classifying an entire data set using a generative LLM and instead use them to create training labels for a supervised model.

A secondary finding is that GPT few-shot models and supervised models trained on GPT-generated labels produce remarkably high performance on recall. GPT-4 few-shot (0.8 median recall) as well as DistilBERT and BERT trained on GPT-labels (both with 0.746 median recall) achieve significantly better median recall than any model trained with human labels. The opposite is true for precision: BERT trained on human-labels achieved the highest precision of the models tested, which was 0.214 higher than median precision for BERT models trained on GPT-4 labels. Therefore, using sythetic training labels may be better suited for tasks where recall is prioritized over precision.

4 Discussion

We demonstrate that synthetic labels from generative LLMs offer a viable strategy for training task-specific supervised classifiers. These models can achieve high performance with minimal resources relative to other options. Future work should explore the performance of additional models by including open-source LLMs (e.g., LLaMa (Touvron et al., 2023)), larger supervised models (e.g., Falcon[7]), and fine-tuned generative LLMs.[8]

A few points of caution are worth emphasizing. There are numerous cases where GPT-4 fails to accurately label the underlying text data. While advancements in LLM technology and additional prompt engineering could mitigate these concerns, it is essential that researchers validate and optimize generative LLM performance against ground-truth human-labeled data. Thus, while generative LLMs can improve the entire classification workflow, their application must always remain human-centered.

[6] We use F1 as our primary evaluation criteria due to class imbalance. Full results are shown in Table A3.

[7] Documentation here: https://huggingface.co/blog/falcon

[8] For example, see here: https://openai.com/blog/gpt-3-5-turbo-fine-tuning-and-api-updates.

5 Limitations

Here, we identify three main limitations of our analysis. First, as discussed in Section 4 and shown in full detail in Table A3, there are various circumstances where supervised models trained on LLM-generated labels fail to produce satisfactory results. This may be due to inaccurate annotations from GPT-4, poor performance from the supervised classifier, or both. While it is possible that additional prompt engineering or hyperparameter tuning could improve performance, it is essential to stress that each of these optimization strategies rely on human labels for comparison. As a result, we argue that it is essential to center human judgement as ground truth when optimizing models and adjudicating between models.

A second, related limitation refers to understanding the errors in the model outputs. Specifically, it is possible that errors from a GPT-trained model produces correlated but unobservable errors. Building a supervised classifier on top of GPT-4 labels would magnify, rather than offset, any such biases. This, too, underscores the importance of human validation and error analysis. It is, of course, also essential to minimize bias by human annotators. For instance, recruiting human annotators from varying demographic backgrounds when conducting an annotation project may diminish the potential for correlated errors across annotators.

Finally, treating human labels as ground truth is an additional limitation. Although most data sets in our analysis employed multiple human coders, it is of course possible that these annotators made correlated errors. As a result, some disagreements between human ground truth labels and synthetic GPT-4 labels may stem from human error. Such errors could bias performance metrics downward for any of the models assessed. Because our primary interest is making comparisons across models, however, we are mainly interested in their relative performance. Because each model would suffer from the same errors in the human labeled data, we do not see this as a significant concern for this analysis.

For the analysis in this paper, our reliance on text classification tasks and data from peer-reviewed research in high-impact journals helps to mitigate concerns about data annotation quality. The annotation procedures in each of these tasks received IRB approval and was assessed by independent reviewers to be of quality enough for publication in a high-impact journal. Still, it is important to acknowledge that applied researchers should invest in high-quality human labels, even if only to validate generative LLM annotation performance.

6 Ethics Statement

Our research complies with the ACL Ethics Policy. Specifically, our research positively contributes to society and human well-being by providing tools that can aid computational social scientists studying the social world. Using the methods we introduce and test will help scientists better understand a wide range of complicated social problems. Because the techniques proposed and assessed in this article require dramatically less resource expenditure than alternatives, our results can help address inequities in resources across researchers.

Due to the inherent risks of deploying biased models, we stress the necessity of human validation throughout our paper. Given the ease and efficiency gains of using generative LLMs to train supervised classifiers, we believe it is essential to build rigorous testing and evaluation standards that are human-centered. This is why we took great efforts to center our analyses on data sets less prone to contamination risks.

Moreover, our research and data analysis does not cause any harm while also respecting privacy and confidentiality concerns. As we discuss in our data collection procedures in Appendix A, we conformed to each data repository's usage and replication policies. Each of the original studies received IRB approval and our analyses conformed to the same safety protocols. All collected data was anonymized by the original authors. Appendix B.3 provides additional details on human annotation protocols, which were all conducted by the original studies and received IRB approval.

References

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec

Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences of the United States of America*, 31.

Zhikai Chen, Haitao Mao, Hongzhi Wen, Haoyu Han, Wei Jin, Haiyang Zhang, and Hui Liuand Jiliang Tang. 2023. Label-free node classification on graphs with large language models (llms).

Sayantan Dasgupta, Trevor Cohn, and Timothy Baldwin. 2023. Cost-effective distillation of large language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7346–7354, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Le, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, page 4171–4186.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks.

Jonas Golde, Patrick Haller, Felix Hamborg, Julian Risch, and Alan Akbik. 2023. Fabricator: An open source toolkit for generating labeled training data with teacher llms.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, page 1789–1819.

Justin Grimmer, Margaret E. Roberts, and Brandon Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

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

Daniel J. Hopkins, Yphtach Lelkes, and Samuel Wolken. 2024. The rise of and demand for identity-oriented media coverage. *American Journal of Political Science*.

David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, and Helen Margetts. 2020. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Wes McKinney. 2011. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9.

Stefan Müller. 2022. The temporal focus of campaign communication. *The Journal of Politics*, 84(1):585–590.

Kimberly A. Neuendorf. 2016. *The Content Analysis Guidebook*. Sage Publications.

Etienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2023. Chatgpt for text annotation? mind the hype!

OpenAI. 2023. Gpt-4 technical report.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, and Trevor Killeen et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*.

Hao Peng, Daniel M. Romero, and Emőke Ágnes Horvát. 2022. Dynamics of cross-platform attention to retracted papers. *Proceedings of the National Academy of Sciences*, 119(25):585–590.

Punyajoy Saha, Narla, Komal Kalyan, and Animesh Mukherjee. 2023. On the rise of fear speech in online social media. *Proceedings of the National Academy of Sciences of the United States of America*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison and´ Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. *In Proceedings of EMNLP*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *International Conference on Machine Learning*, page 11328–11339.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? Working paper.

A Appendix: Data sets

In this section, we elaborate on the data sets used in our analysis. Our corpus includes 14 classification tasks across five data sets representing recent applications in computational social science. To avoid the potential for contamination, we rely exclusively on data sets stored in password-protected data archives (e.g., Dataverse). We draw from research published in outlets across a spectrum of disciplines ranging from interdisciplinary publications (e.g., *Proceedings of the National Academy of Sciences*) to high-impact field journals in social science (e.g., *American Journal of Political Science*). To find these articles, we searched journals for articles related to computational social science that implemented some type of manual annotation procedure. The human-labeled data from the original study is treated as the ground truth. We discuss the human annotation procedures in the original studies at greater length in Appendix B.3.

Table A1 and Table A2 contain the full details for every task and data set. Overall, our data encompass diverse degrees of class imbalance: Across tasks, the mean positive class frequency is 16.2%, the minimum is 0.04%, and the maximum is 61%. The sources of labels are representative of common approaches to annotation: 42.9% of tasks were annotated by crowdsourced workers, 28.6% by experts, and 28.6% by research assistants.

Our repications involve fine-tuning supervised classifiers using manually annotated data from the replication data sets. For every replication classification task, we conformed to each data repository's replication policies. Each of the original studies received IRB approval and our analyses conformed to the same safety protocols, including full anonymization and agreeing to not publicly post the raw data without permission. As such, our replication of each data set is compatible with its intended usage.

Although all of the data sets were anonymized before our replications, we manually reviewed each data set to confirm privacy protections. One of the data sets (Saha et al., 2023) contains hate speech, but this is because it is a central part of the research question from the original study. As a result, we include examples of hate speech in that particular replication. From manual review, no other data set contained offensive material.

B Appendix: Additional methodological details

B.1 Prompt tuning

As discussed in Section 2, for every task, we optimized each GPT-4 prompt on a subset of 250 text samples labeled by humans. To do this, we tested generative LLM performance on the subset of data and then, if relevant, made iterative human-in-the-loop updates to the codebook to optimize the prompt for accurate annotations. To harmonize the diverse range of annotation tasks into a common framework for evaluation, we treat every dimension as a separate binary annotation task. Thus, if an article includes a classification task with three potential labels, we split the annotation process into three discrete binary classification tasks. In the supplementary material, we include each LLM prompt instruction as a .txt file. We also include our code to query the GPT-4 API.

Figure A3 shows the distributions of change in performance metrics after updating the LLM prompt and re-annotating the same text samples. This analysis demonstrates whether and how prompt optimization affects LLM annotation, holding constant the data and conceptual categories. In most cases, prompt optimization led to modest improvement in accuracy and F1—although recall decreased in more cases than improved after updating the prompts. While the magnitude of improvement was generally small, researchers experiencing subpar LLM annotation performance can use human-in-the-loop prompt optimization to ensure that their instructions are not the cause of poor performance.

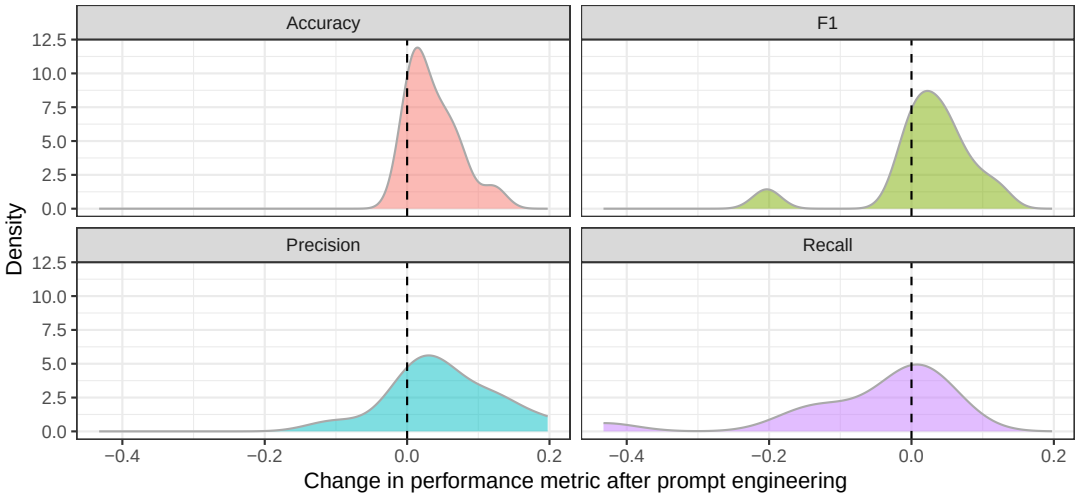

Figure A3: Change in LLM annotation performance on training data after one round of prompt optimization

B.2 Hyperparameter tuning, evaluation, and compute details

Our experiment involved varying the training data used to fine-tune supervised classifiers (i.e., 250 human samples, 1000 human samples, and 1000 GPT-labeled samples). To select each supervised classifier, we implemented a grid search over 18 possible hyperparameter combinations. In particular, we optimized learning rate (1e-5, 2e-5, and 5e-5), batch size (8 and 16), and epochs (2, 4, 6). We conducted our search on a subsample of 250 text samples per task and retained the best hyperparameters (in terms of highest F1) across each task. We subsequently used the best-performing combination of hyperparameters for all applications of a specific model (see best-performing hyperparameter configurations in Table A4). Despite not adopting a more exhausting approach to hyperparameter tuning approach, we observe strong performance across our classification tasks, with a few exceptions. Table A5 displays additional model hyperparameters that remained constant across tasks, as well as basic information about each model's architecture. We selected the chosen pretrained models (i.e., BERT, RoBERTa, and DistilBERT) because of their ease of usage, low cost, and popularity among computational social scientists.

For all 14 tasks, evaluation was conducted on a test set of 1000 held-out text samples that had previously been labeled by human annotators. As is standard in classification evaluation, we report accuracy, F1, precision, and recall for every task and model. Table A3 displays the full classification results across all tasks and models.

All of our supervised training analyses were implemented in Python 3.10.12 with HuggingFace's Transformers (Wolf et al., 2020) and PyTorch libraries (Paszke et al., 2019). We conducted all data preprocessing in Python Pandas (McKinney, 2011). Our computing infrastructure was Google Colab, where we used 215 T4 GPU compute units (roughly 421.4 GPU hours). As with our model selection, we chose this computing environment due to its low cost and ease of application. Any computational social scientist could conduct the same analyses. In the supplementary material, we include all code to run our supervised training procedures.

B.3 Additional details on human annotation procedures

We introduce a novel corpus of labeled text data for annotations. To create this data set, we compile labeled data from recent studies, as detailed in A1. As a result, we did not work with annotators to generate any original data. We adopted materials from these original studies instead. While we do not report the instructions given to each study's human annotators, we do provide the prompt instructions that were used to query GPT-4 in the supplementary material. These instructions were taken directly from the original study's human annotator instructions. All additional details on the annotation procedures (e.g., how they were recruited, payment, consent, and demographic characteristics) can be found in the original studies' supplementary material.

While we do not describe each study's procedures in detail, we manually selected our annotation studies due to their high-quality human label-

ing practices. All of the replicated studies were approved by an IRB. These studies all deployed either expert coders or numerous non-expert coders of varying backgrounds. Because all of the human annotation text is part of the peer-review process in high-impact journals and due to the strict annotation guidelines and principles these studies adhered to, we conclude that the human annotations are of high-quality.

C Appendix: Miscellaneous additional information

Additional sources:

- Robot image (used in Figure 1): https://commons.wikimedia.org/wiki/File:Grey_cartoon_robot.png
- Human silhouette image (used in Figure 1): https://commons.wikimedia.org/wiki/File:SVG_Human_Silhouette.svg

Author(s)	Title	Journal	Year
Card et al.	Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration	PNAS	2022
Hopkins, Lelkes, and Wolken	The Rise of and Demand for Identity-Oriented Media Coverage	American Journal of Political Science	2024
Müller	The Temporal Focus of Campaign Communication	Journal of Politics	2021
Peng, Romero, and Horvat	Dynamics of cross-platform attention to retracted papers	PNAS	2022
Saha et al.	On the rise of fear speech in online social media	PNAS	2022

Table A1: Replication data sources.

Study	# of tasks	Annotation source	Classification tasks
Card et al. (2022)	4	Research assistants	Classify US congressional speeches to identify whether the speech discussed immigration or immigration policy, along with an accompanying tone: pro-immigration, anti-immigration, or neutral.
Hopkins, Lelkes, and Wolken (2024)	4	Crowd	Classify headlines, Tweets, and Facebook share blurbs to identify references to social groups defined by a) race/ethnicity; b) gender/sexuality; c) politics; d) religion.
Müller (2021)	3	Expert	Classify sentences from political party manifestos for temporal direction: past, present, or future.
Peng, Romero, and Horvat (2022)	1	Expert	Classify whether Tweets express criticism of findings from academic papers.
Saha et al. (2020)	2	Crowd	Classify social media posts into fear speech, hate speech, both, or neither.

Table A2: Descriptions of replication classification tasks.

Data set	Task	Model	Training data: Few shot Ac.	Few shot F1	Few shot Pr.	Few shot Re.	Human: 250 Ac.	Human: 250 F1	Human: 250 Pr.	Human: 250 Re.	Human: 1000 Ac.	Human: 1000 F1	Human: 1000 Pr.	Human: 1000 Re.	GPT: 1000 Ac.	GPT: 1000 F1	GPT: 1000 Pr.	GPT: 1000 Re.
Card et al.	Cat: Neg	GPT-4	0.85	0.65	0.54	0.83												
		BERT					0.88	0.58	0.74	0.48	0.87	0.56	0.65	0.49	0.81	0.56	0.47	0.72
		RoBERTa					0.85	0.51	0.59	0.45	0.84	0.48	0.55	0.42	0.78	0.57	0.43	0.82
		DistilBERT					0.86	0.56	0.61	0.51	0.86	0.58	0.61	0.55	0.81	0.58	0.47	0.74
	Cat: Imm	GPT-4	0.81	0.81	0.74	0.90												
		BERT					0.85	0.84	0.79	0.89	0.86	0.86	0.81	0.91	0.84	0.83	0.76	0.91
		RoBERTa					0.86	0.85	0.80	0.92	0.85	0.84	0.77	0.92	0.82	0.82	0.74	0.92
		DistilBERT					0.85	0.84	0.80	0.88	0.84	0.84	0.79	0.89	0.82	0.82	0.73	0.92
	Cat: Neut.	GPT-4	0.83	0.26	0.27	0.25												
		BERT					0.80	0.35	0.29	0.44	0.85	0.36	0.38	0.35	0.87	0.38	0.44	0.34
		RoBERTa					0.88	0.30	0.46	0.23	0.88	0.00	0.00	0.00	0.84	0.33	0.33	0.34
		DistilBERT					0.85	0.28	0.32	0.25	0.85	0.36	0.37	0.35	0.86	0.38	0.40	0.36
	Cat: Pro	GPT-4	0.88	0.50	0.55	0.46												
		BERT					0.86	0.33	0.44	0.27	0.84	0.44	0.42	0.46	0.87	0.45	0.51	0.40
		RoBERTa					0.87	0.37	0.51	0.30	0.84	0.37	0.41	0.34	0.85	0.41	0.43	0.39
		DistilBERT					0.87	0.29	0.55	0.19	0.83	0.38	0.38	0.37	0.84	0.35	0.40	0.31
Hopkins et al.	Political	GPT-4	0.88	0.43	0.30	0.79												
		BERT					0.95	0.32	0.60	0.22	0.96	0.62	0.71	0.54	0.82	0.34	0.21	0.82
		RoBERTa					0.84	0.37	0.23	0.85	0.96	0.62	0.73	0.54	0.84	0.37	0.23	0.85
		DistilBERT					0.94	0.29	0.50	0.20	0.96	0.63	0.72	0.56	0.83	0.34	0.22	0.80
	Gender	GPT-4	0.95	0.74	0.68	0.82												
		BERT					0.91	0.20	0.46	0.13	0.96	0.80	0.86	0.74	0.94	0.72	0.62	0.85
		RoBERTa					0.91	0.08	0.44	0.04	0.95	0.73	0.78	0.68	0.92	0.67	0.54	0.87
		DistilBERT					0.94	0.52	0.83	0.38	0.97	0.81	0.87	0.75	0.93	0.71	0.59	0.88
	Race	GPT-4	0.96	0.57	0.41	0.92												
		BERT					0.97	0.00	0.00	0.00	0.98	0.56	0.71	0.46	0.98	0.64	0.54	0.77
		RoBERTa					0.97	0.00	0.00	0.00	0.97	0.00	0.00	0.00	0.97	0.59	0.45	0.85
		DistilBERT					0.97	0.00	0.00	0.00	0.99	0.71	0.77	0.65	0.97	0.54	0.46	0.65
	Religion	GPT-4	0.98	0.61	0.47	0.88												
		BERT					0.98	0.21	1.00	0.12	0.99	0.73	0.75	0.71	0.98	0.61	0.48	0.82
		RoBERTa					0.98	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.98	0.00	0.00	0.00
		DistilBERT					0.98	0.00	0.00	0.00	0.99	0.69	0.67	0.71	0.97	0.53	0.37	0.94
Müller	Future	GPT-4	0.82	0.85	0.87	0.83												
		BERT					0.83	0.85	0.88	0.84	0.82	0.85	0.85	0.85	0.81	0.85	0.84	0.87
		RoBERTa					0.84	0.87	0.87	0.88	0.82	0.85	0.86	0.85	0.82	0.86	0.84	0.87
		DistilBERT					0.83	0.86	0.85	0.86	0.81	0.84	0.87	0.82	0.82	0.85	0.83	0.88
	Past	GPT-4	0.91	0.74	0.66	0.84												
		BERT					0.94	0.83	0.74	0.93	0.95	0.83	0.80	0.85	0.93	0.79	0.71	0.89
		RoBERTa					0.94	0.80	0.81	0.79	0.95	0.85	0.79	0.92	0.85	0.00	0.00	0.00
		DistilBERT					0.94	0.79	0.77	0.80	0.94	0.80	0.79	0.82	0.93	0.79	0.68	0.96
	Present	GPT-4	0.82	0.62	0.64	0.60												
		BERT					0.83	0.65	0.66	0.64	0.83	0.65	0.64	0.66	0.81	0.61	0.63	0.58
		RoBERTa					0.84	0.66	0.71	0.61	0.84	0.68	0.68	0.67	0.83	0.61	0.68	0.56
		DistilBERT					0.83	0.64	0.69	0.59	0.83	0.65	0.66	0.64	0.82	0.59	0.66	0.54
Peng et al.	Critical	GPT-4	0.85	0.54	0.48	0.63												
		BERT					0.87	0.43	0.59	0.34	0.91	0.63	0.76	0.54	0.79	0.43	0.35	0.56
		RoBERTa					0.88	0.44	0.61	0.34	0.87	0.62	0.54	0.73	0.78	0.43	0.34	0.59
		DistilBERT					0.83	0.43	0.42	0.44	0.86	0.54	0.50	0.58	0.77	0.41	0.33	0.56
Saha et al.	CV	GPT-4	0.97	0.06	0.03	0.25												
		BERT					1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.94	0.03	0.02	0.25
		RoBERTa					1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.93	0.05	0.03	0.50
		DistilBERT					1.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.94	0.10	0.05	0.75
	HD	GPT-4	0.88	0.35	0.28	0.45												
		BERT					0.91	0.17	0.24	0.13	0.92	0.41	0.45	0.38	0.90	0.21	0.24	0.19
		RoBERTa					0.92	0.24	0.35	0.19	0.92	0.47	0.43	0.52	0.91	0.20	0.26	0.16
		DistilBERT					0.91	0.26	0.32	0.22	0.91	0.40	0.38	0.42	0.91	0.28	0.33	0.25

Table A3: Complete task-by-task classification performance results. Ac., Pr., and Re. refer to accuracy, precision, and recall, respectively.

Study	Task	Hyperparameters
Card et al.	Classify immigration speeches	learning rate (5e-05), batch size (8), epochs (4)
	Classify pro-immigration speeches	learning rate (5e-05), batch size (16), epochs (6)
	Classify anti-immigration speeches	learning rate (5e-05), batch size (8), epochs (6)
	Classify neutral immigration speeches	learning rate (5e-05), batch size (8), epochs (4)
Hopkins et al.	Classify race/ethnicity	learning rate (2e-05), batch size (8), epochs (4)
	Classify gender	learning rate (5e-05), batch size (8), epochs (6)
	Classify political groups	learning rate (5e-05), batch size (16), epochs (6)
	Classify religious groups	learning rate (5e-05), batch size (8), epochs (6)
Müller	Classify past	learning rate (5e-05), batch size (8), epochs (4)
	Classify present	learning rate (5e-05), batch size (8), epochs (4)
	Classify future	learning rate (2e-05), batch size (8), epochs (6)
Peng et al.	Classify criticism	learning rate (5e-05), batch size (8), epochs (6)
Saha et al.	Classify fear speech	learning rate (5e-05), batch size (8), epochs (6)
	Classify hate speech	learning rate (5e-05), batch size (8), epochs (4)

Table A4: Hyperparameter settings per task.

	BERT-base	RoBERTa-base	DistilBERT
# parameters	110m	125m	66m
# attention heads	12	12	12
Hidden dim.	768	768	768
Feedforward dim.	3072	3072	3072
Activation	GELU	GELU	GELU
Dropout	0.1	0.1	0.1
Optimizer	Adam	Adam	Adam
Weight decay	0.01	0.01	0.01

Table A5: Model architectures and additional hyperparameters.