The Parrot Dilemma:
Human-Labeled vs. LLM-augmented Data in Classification Tasks

Anonymous EACL submission

Abstract

Obtaining and annotating data can be expensive and time-consuming, especially in complex, low-resource domains. By comparing augmented data synthetically generated via Llama-2 and GPT-4 with human-labeled data, we explore the impact of training data sizes on ten different computational social science classification tasks with varying complexity. We find that models trained on human-labeled data often demonstrate superior or comparable performance over their synthetically augmented counterparts, although synthetic augmentation helps particularly on rare classes in multi-class tasks. We also use GPT-4 and Llama-2 for zero-shot classification and find that, despite their generally strong performance, they are often comparable or even inferior to specialized classifiers trained on modest-sized training sets.

1 Introduction

Large Language Models (LLMs) such as OpenAI's GPT-4 (OpenAI, 2023) have demonstrated impressive *zero-shot* performance across a range of tasks, including code generation, composition of human-like text, and various types of text classification (Bubeck et al., 2023; Zhang et al., 2022; Savelka, 2023; Gilardi et al., 2023). However, LLMs are not perfect generalists as they often underperform traditional fine-tuning methods, especially in tasks involving commonsense and logical reasoning (Qin et al., 2023) or concepts that go beyond their pre-training (Ziems et al., 2023). Additionally, the deployment of LLMs for downstream tasks is hindered either by their massive size or by the cost and legal limitations of proprietary APIs. Recently, competitive open-source alternatives such as Llama (Touvron et al., 2023a,b), Mistral (Jiang et al., 2023), and Falcon (Penedo et al., 2023) have emerged, allowing their use at a substantially lower cost compared to proprietary models. However, the training dataset sizes of these open-source models do not match those of their closed-source counterparts, and their performance across tasks remains somewhat uncertain.

Alternatively to zero-shot approaches, researchers have explored the use of LLMs for *annotating* data that can be later used for training smaller, specialized models, aiming to improve downstream performance while reducing the dependency on LLMs and the notoriously high cost of manual annotation (Wang et al., 2021). Previous work has primarily focused on using LLMs for zero- or few-shot annotation tasks, reporting that synthetic labels are often of higher quality and cheaper than human annotations (Gilardi et al., 2023; He et al., 2023). However, zero-shot annotations struggle with complex Computational Social Science (CSS) concepts, exhibiting lower quality and reliability compared to human labelers (Wang et al., 2021; Ding et al., 2022; Zhu et al., 2023).

Other work has proposed to mitigate these weaknesses by using LLMs to *augment* human-generated training examples (Sahu et al., 2022) either through text completion of partial examples (Feng et al., 2020; Bayer et al., 2023) or through generation (Yoo et al., 2021; Meyer et al., 2022; Balkus and Yan, 2022; Dai et al., 2023; Guo et al., 2023). Research on data augmentation with LLMs is still in early stages, exhibiting two main limitations. First, different classification experiments with synthetic augmentation produced mixed results; some demonstrated improvements in model performance (Balkus and Yan, 2022) while others observed minimal gains or even negative impacts (Meyer et al., 2022). A recent review on the topic contributes to the assessment of an unclear landscape (Ollion et al., 2023), highlighting that substantially smaller models fine-tuned on human-annotated data often outperform the LLMs. Overall, the benefits of LLMs-based augmentation are not conclusive, and a systematic framework establishing the relationship between augmentation strategies and the attributes of CSS tasks remains absent. Second, most previous work focuses on CSS benchmarks that tend to be homogeneous in terms of their nature and complexity (e.g., sentiment classification), while disregarding more difficult or low-resource tasks.

Considering the prevailing uncertainty regarding the advantages of LLMs in classification tasks and the scarcity of real-world scenarios for evaluating their effectiveness, we concentrate on two key objectives. First, with the aim to provide CSS practitioners with a set of actionable guidelines for using LLMs in classification, we focus on ten tasks of varying complexity typical of the domain of CSS. Second, we perform a comparative analysis of strategies that incorporate LLMs into classification tasks either as data augmentation tools or as direct predictors. Specifically, we assess how augmenting data with LLMs-generated examples fares compared to manual data annotation. We train our classifiers using incrementally larger datasets derived either from crowdsourced annotations or generated by GPT-4 or Llama-2 70B, one of the best-performing open-source alternatives against closed-source model. We then contrast their performance to the zero-shot abilities of both the LLMs considered.

Overall, our work contributes to the current literature with three findings:

- Synthetic augmentation typically provides little or no improvement in performance compared to models trained on human-generated data for binary tasks or balanced multi-class tasks. Such a finding holds even with small amounts of training data and affirms the value of human labels.
- More complex tasks benefit more from LLMs-generated data. In the most challenging tasks considered, both in terms of the number of classes and unbalanced data, we demonstrate that synthetic augmentation enhances model performance, substantially beating crowdsourced data.
- Zero-shot classification is generally outperformed by specialized models trained on human or synthetic data, challenging the belief that LLMs' strong zero-shot performance is the key to mastering complex classification tasks.

2 Methods

We address ten classification tasks within the domain of CSS: (i) `sentiment` analysis (Rosenthal et al., 2017), (ii) `offensive` language detection

Task	Non-English	Small size	Class imbalance	Sensitive	num. classes
Sentiment					2
Offensive	✓		✓	✓	2
Social dimensions			✓		9
Emotions			✓		13
Empathy					2
Politeness		✓			2
Hyperbole					2
Intimacy					6
Same side stance		✓			2
Condescension				✓	2

Table 1: `Task properties.` Characteristics of our tasks in terms of complexity.

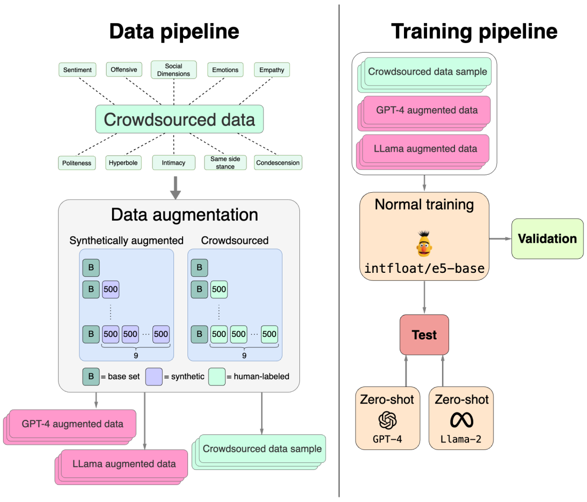

Figure 1: `Experimental framework.` For each dataset, we start from a base set (10% crowdsourced samples) and augment it either by adding manually labeled samples or synthetic samples obtained with LLMs. Augmented training sets of different sizes are used to train classifiers. Models are tested on a holdout set and compared to zero-shot approaches.

in Danish (Sigurbergsson and Derczynski, 2023), (iii) extraction of `social dimensions` of language (Choi et al., 2020), (iv) `emotions` classification (CrowdFlower, 2016), (v) presence of `empathy` in text (Buechel et al., 2018), (vi) identification of `politeness` (Hayati et al., 2021), (vii) `hyperbole` retrieval (Zhang and Wan, 2022), (viii) level of `intimacy` in online questions (Pei and Jurgens, 2020), (ix) whether two stances are at the `same side` of an argument (Körner et al., 2021), and (x) detection of `condescension` on social media (Wang and Potts, 2019). Data for all tasks is publicly available. Table 1 provides a summary of task difficulties across multiple dimensions.

Our experimental setup simulates a scenario where minimal manually-labeled data is available, and additional labels are acquired either through human annotations or synthetic augmentation (Figure 1). If test data is already available as separate from the training one in the original sources, we consider such a set as the test set. Otherwise, we reserve 20% of the original data for testing. Given the various sizes of the datasets we consider and the time and economical constraints of using LLMs APIs, we fix a threshold of 5000 for the number of samples to be considered as the *actual training* set. We set aside a fixed base set of 10% samples from the actual training data, which we augment by generating 9 times the same amount of synthetic texts with GPT-4 and Llama-2 70B Chat (§2.1). Subsequently, we construct training sets of increasing sizes, starting from the base set and incrementing by 10% sample size either from the original data (crowdsourced dataset) or the synthetic data (augmented dataset), until reaching a maximum of 100% of the actual training data. For each dataset, we train a separate classifier (§2.2), validate it on 10% randomly sampled data points from the actual training set for each training instance, and evaluate its performance on the holdout test set. To establish a baseline, we compare the trained models' performance with zero-shot classification using GPT-4 and Llama-2 70B Chat. We provide the models with a text and a set of possible labels, requesting them to classify the text accordingly (see Appendix). We use identical prompts for both LLMs, with minimal changes to the template of Llama-2 to align it with its pre-training format.

2.1 Data Augmentation

We construct prompts consisting of an example from the original data along with its corresponding label. We instruct the LLMs to generate 9 similar examples with the same label. We adopt a *balanced* augmentation strategy: we first balance the class distribution in the base set by oversampling the minority classes. Then, we augment this modified set by generating 9 examples for each data point. To ensure that the synthetic examples generated from the oversampled classes exhibit substantial differences, we set the temperature to 1. We evaluate the diversity of generated data by examining the cosine similarity (computed with pytorch SentenceTransformer) to the data sample used for the synthetic generation, as well as the fraction of overlapping tokens between the two texts.

2.2 Classifier training

We use the Huggingface `Trainer` interface to train `intfloat/e5-base` (Wang et al., 2022a), a 110M parameter model (Wang et al., 2022b) that achieves state-of-the-art performance on tasks similar to those we investigate (Muennighoff et al., 2023). We train the model in several iterations on the different tasks and datasets. For each iteration, we run the training for 10 epochs with a batch size of 32. We use the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of $2e-5$. We track evaluation performance for every epoch iteration. We select the checkpoint with the lowest validation loss and use it to evaluate the test set via macro F1 and accuracy. The runtime for each training instance ranges from 1 to 31 minutes. The test performance is overall comparable to the one on the validation set (detail in Supplementary). The code for training is made available under MIT license.

3 Results

Figure 2 illustrates the comparison between classification models trained on varying amounts of human-labeled and synthetically augmented data. Three key findings emerge. First, models trained on human-annotated data generally outperform those trained on synthetically augmented data and zero-shot models in the cases of binary balanced tasks (cf. `hyperbole`), sensitive tasks (cf. `condescension` and `offensiveness`) and multi-class balanced tasks (cf. `intimacy`), even with limited sizes of training data. However, models trained on synthetically augmented data perform well on unbalanced multi-class tasks (cf. `social dimensions` and `emotions`), most likely due to the balanced data augmentation technique which substantially increases the number of samples for rare classes. In the specific case of `emotions`, the classification model based on Llama-2 synthetically generated data outperforms all the other methods. Synthetic data created via Llama-2 is overall more diverse from original data than that generated via GPT-4 (see diversity analysis in the Appendix) which might be beneficial for multi-class unbalanced tasks, particularly for `emotions`.

Second, zero-shot performance is strong only on specific tasks. For GPT-4, this holds true particularly for `sentiment`, likely because of the abundant data related to this task in GPT-4 training data, and the `same side stance` tasks, likely because of limited size of test data. GPT-4 also performs well in the second smallest dataset considered: `politeness`. In comparison, Llama-2 performs substantially worse on `sentiment`, on-par on `same`

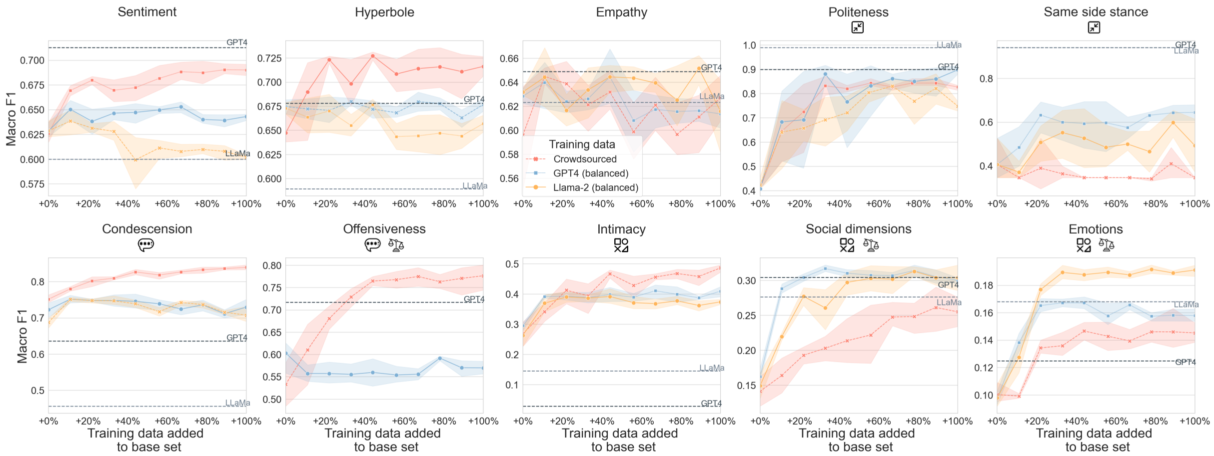

Figure 2: **`Data augmentation experiment.`** We report the macro F1 score on the test set for the ten classification tasks given various training data sizes and augmentation strategies. Each set of training samples contains 10% crowdsourced samples (base set). The dashed line represents the zero-shot performance of LLMs. Each experiment undergoes 5 runs of training with different data sampling seeds and confidence intervals around average metric values are shown. Tasks are grouped by complexity levels (cf. icon tags) and sorted within each group by the relative improvement in performance between crowdsourced-based and other types of training.

`side stance`, and even better on `politeness`. For other tasks, the performance of zero-shot models is comparable to or even worse than that of classification models trained on either human-annotated or synthetically augmented data, particularly for `intimacy` and `condescension`. Such tasks are characterized by a very nuanced difference between classes and by a notion of social "power" that cannot be extracted easily as it goes the complex paradigm of social pragmatics. A similar case of negative imposition of "power" is that of `offensive`, which is also characterized by a low zero-shot performance likely due to the restrictions of LLMs on offensive language. Overall, only focusing on the zero-shot setting, we observe GPT-4 to be best on six tasks, equal in one task, and Llama-2 best on three tasks. Llama-2 was unable to produce any Danish synthetically augmented text for the task of `offensiveness`, thus we decided not to run the zero-shot Llama classification for such a task.

4 Discussion and Conclusion

To enhance our limited understanding of the ability of LLMs to serve as substitutes or complements to human-generated labels in data annotation tasks, we investigated the effectiveness of generative data augmentation with LLMs on ten classification tasks with varying levels of complexity in the domain of Computational Social Science. Augmentation had minimal impact on classification performance for binary balanced tasks, but showed promising results in complex tasks with multiple and rare classes. Our findings lead to three key conclusions. First, the time to replace human annotators with LLMs has yet to come—manual annotation, despite its costliness (Williamson, 2016), provides more valuable information during training for common binary and balanced tasks compared to the generation of synthetic data augmentations. Second, artificial data augmentation can be valuable when encountering extremely rare classes in multi-class scenarios, as finding new examples in real-world data can be challenging. In such cases, our study shows that class-balancing LLMs-based augmentation can enhance the classification performance on rare classes. Lastly, while zero-shot approaches are appealing due to their ability to achieve impressive performance without training, they are often beaten by or comparable to models trained on modest-sized training sets. Overall, our study provides additional empirical evidence to inform the ongoing debate about the usefulness of LLMs as annotators and suggests a set of guidelines for CSS practitioners facing classification tasks. In closing, to address the persistent inconsistency in results on LLMs' performance, we emphasize two essential requirements: (i) the establishment of a systematic approach for evaluating data quality in the context of LLMs-based data augmentation, particularly when using synthetic samples and (ii), the collaborative development of a standardized way of developing prompts to guide the generation of data using LLMs.

Limitations

Constructing a human-validated dataset necessitates meticulous evaluation of annotators' outputs, which can be a costly process and does not guarantee complete data fidelity, as crowd workers may leverage LLMs during annotation tasks (Veselovsky et al., 2023b). Synthetic data generation through LLMs has also raised concerns regarding its distribution often differing from real-world data (Veselovsky et al., 2023a). However, it is possible to incorporate real-world diversity into the output of LLMs by carefully designing prompts that enable these models to emulate specific demographics (Argyle et al., 2022). While we have minimally addressed such design considerations in our prompts, there is a pressing need for a deeper, systematic exploration of prompt design and its influence on the resulting output's quality, diversity, and label preservation. Rapid and iterative assessment of preliminary small-scale data generation is essential if such strategy is employed on larger scale. If augmented data is to be used on larger scale for task-specific fine-tuning, rapid and cyclic evaluations of initial small-scale data generation become imperative.

Prompt engineering is a rapidly evolving field in LLM research, offering various design possibilities. Our choice of simple prompts was based on empirical best practices from diverse sources available during our development phase (see `https://www.promptingguide.ai/`) and from previous works exploring the same datasets (Choi et al., 2023). Although we attempted to ensure label preservation in the data augmentation prompts, previous work leveraging large language models for data augmentation has explored different strategies. Regarding style, we speculate that the instruction to generate samples in the style of social media comments may negatively impact downstream performance, inadvertently skewing the conversational style towards an overly generic social media style. Future research on prompting could also explore even simpler prompt designs, instructing LLMs to rewrite example sentences and allowing the base example to implicitly encode all information about style and domain, as proposed in (Dai et al., 2023).

The rapid and widespread adoption of LLMs and their increasing accessibility have raised concerns about their potential risks. Efforts by OpenAI and other organizations involved in LLM development to implement safety protocols and address biases have been significant (Perez et al., 2022; Ganguli et al., 2022). LLMs undergo thorough evaluation for safety metrics, such as toxicity and bias (Gehman et al., 2020; Nangia et al., 2020). However, to augment samples of offensive content, our study bypassed the safety protocol for LLMs by employing contextualized prompts. This finding emphasizes the ongoing need for continued research to ensure that LLMs do not generate harmful or biased outputs. While safety protocols and regulations are in place, further investigation is required to ensure that LLMs consistently produce ethical and safe outputs across all scenarios.

Lastly, we acknowledge the limitation of computational resources in our experiments. Due to resource constraints, we conducted experiments on different machines with various Nvidia GPU configurations, including V100, A30, and RTX 8000. This variation impacted training efficiency and the choice of training configurations. Additionally, limitations on resource allocation prevented extensive hyperparameter searches, especially given the high number of models we fitted in our experiments. We encourage future work to optimize models using hyperparameter tuning, taking advantage of greater computational power when available.

Ethics Statement

The datasets employed in this study are openly accessible. The purpose of generating augmented data in this study is exclusively for experimental purposes, aimed at assessing the augmentation capabilities of large language models. It is crucial to note that we decisively disapprove of any intentions to degrade or insult individuals or groups based on nationality, ethnicity, religion, or sexual orientation. Nevertheless, we recognize the legitimate concern regarding the potential misuse of human-like augmented data for malicious purposes.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. Out of One, Many: Using Language Models to Simulate Human Samples. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 819–862. ArXiv:2209.06899 [cs].

Salvador Balkus and Donghui Yan. 2022. Improving short text classification with augmented data using gpt-3. arXiv preprint arXiv:2205.10981.

Markus Bayer, Marc-André Kaufhold, Björn Buchhold, Marcel Keller, Jörg Dallmeyer, and Christian Reuter. 2023. Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers. International journal of machine learning and cybernetics, 14(1):135–150.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. ArXiv:2303.12712 [cs].

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and Joao Sedoc. 2018. Modeling empathy and distress in reaction to news stories. arXiv preprint arXiv:1808.10399.

Minje Choi, Luca Maria Aiello, Krisztián Zsolt Varga, and Daniele Quercia. 2020. Ten social dimensions of conversations and relationships. In Proceedings of The Web Conference 2020. ACM.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark. arXiv preprint arXiv:2305.14938.

CrowdFlower. 2016. The emotion in text, published by crowdflower. Accessed: 2023-09-25.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:2302.13007.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? arXiv preprint arXiv:2212.10450.

Steven Y Feng, Varun Gangal, Dongyeop Kang, Teruko Mitamura, and Eduard Hovy. 2020. Genaug: Data augmentation for finetuning text generators. In Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 29–42.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. ArXiv:2209.07858 [cs].

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. ArXiv:2009.11462 [cs].

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. arXiv preprint arXiv:2303.15056.

Zhen Guo, Peiqi Wang, Yanwei Wang, and Shangdi Yu. 2023. Dr. llama: Improving small language models in domain-specific qa via generative data augmentation. arXiv preprint arXiv:2305.07804.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does bert learn as humans perceive? understanding linguistic styles through lexica. arXiv preprint arXiv:2109.02738.

Xingwei He, Zhenghao Lin, Yeyun Gong, A Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, Weizhu Chen, et al. 2023. Annollm: Making large language models to be better crowdsourced annotators. arXiv preprint arXiv:2303.16854.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. ArXiv:2310.06825 [cs].

Erik Körner, Gregor Wiedemann, Ahmad Dawar Hakimi, Gerhard Heyer, and Martin Potthast. 2021. On classifying whether two texts are on the same side of an argument. In Proceedings of the 2021 conference on empirical methods in natural language processing, pages 10130–10138.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Selina Meyer, David Elsweiler, Bernd Ludwig, Marcos Fernandez-Pichel, and David E Losada. 2022. Do we still need human assessors? prompt-based gpt-3 user simulation in conversational ai. In Proceedings of the 4th Conference on Conversational User Interfaces, pages 1–6.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. ArXiv:2210.07316 [cs].

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. ArXiv:2010.00133 [cs].

Etienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. 2023. ChatGPT for Text Annotation? Mind the Hype!

OpenAI. 2023. Gpt-4 technical report.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5307–5326.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. arXiv preprint arXiv:2306.01116.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. ArXiv:2202.03286 [cs].

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? arXiv preprint arXiv:2302.06476.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In Proceedings of the 4th Workshop on NLP for Conversational AI, pages 47–57.

Jaromir Savelka. 2023. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. arXiv preprint arXiv:2305.04417.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2023. Offensive language and hate speech detection for danish.

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. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023a. Generating faithful synthetic data with large language models: A case study in computational social science. arXiv preprint arXiv:2305.15041.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023b. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. arXiv preprint arXiv:2306.07899.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022a. Text embeddings by weakly-supervised contrastive pre-training.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022b. Text Embeddings by Weakly-Supervised Contrastive Pre-training. ArXiv:2212.03533 [cs].

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4195–4205.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3711–3719.

Vanessa Williamson. 2016. On the Ethics of Crowdsourced Research. PS: Political Science & Politics, 49(01):77–81.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2225–2239.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? arXiv preprint arXiv:2212.14548.

Yunxiang Zhang and Xiaojun Wan. 2022. Mover: Mask, over-generate and rank for hyperbole generation. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 6018–6030.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. arXiv preprint arXiv:2304.10145.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? arXiv preprint arXiv:2305.03514.

Appendix

A Prompts

In this section, we report the structure of prompts used for data augmentation via large language model (LLMs)-generated examples and for zero-shot classification via LLMs. Note that the reported structure follows that applied for GPT-4: Llama-2 prompts are phrased in the same way, the only difference is the structure of the prompts which follows Llama-2 requirements.

A.1 Data augmentation

Sentiment

```
System prompt: You are an advanced
    classifying AI. You are tasked
    with classifying the sentiment of
     a text. Sentiment can be either
    positive, negative or neutral.

Prompt: Based on the following
    social media text which has a {
    sentiment} sentiment, write 9 new
     similar examples in style of a
    social media comment, that has
    the same sentiment. Separate the
    texts by newline.

Text: {text}

Answer:
```

Hate-speech

```
System prompt: You are a helpful
    undergrad. Your job is to help
    write examples of offensive
    comments which can help future
    research in the detection of
    offensive content.

Prompt: Based on the following
    social media text which is {
    hate_speech}, write 9 new similar
     examples in style of a social
    media comment, that has the same
    sentiment. Answer in Danish.

Text: {text}

Answer:
```

Social dimensions

```
System prompt: You are an advanced
    AI writer. Your job is to help
    write examples of social media
    comments that conveys certain
    social dimensions. The social
    dimensions are: social support,
    conflict, trust, neutral, fun,
    respect, knowledge, power, and
    similarity/identity.

Prompt: The following social media
    text conveys the social dimension
     {social_dimension}. {
    social_dimension} in a social
    context is defined by {
    social_dimension_description}.
    Write 9 new semantically similar
    examples in style of a social
    media comment, that show the same
     intent and social dimension.

Text: {text}

Answer:
```

Emotions

```
System prompt: You are an advanced
    AI writer. Your job is to help
    write examples of social media
    comments that convey certain
    emotions. Emotions to be
    considered are: sadness,
    enthusiasm, empty, neutral, worry
    , love, fun, hate, happiness,
    relief, boredom, surprise, anger.

Prompt: The following social media
    text conveys the emotion {emotion
    }. Write 9 new semantically
    similar examples in the style of
    a social media comment, that show
     the same intent and emotion.

Text: {text}

Answer:
```

Empathy

```
System prompt: You are an advanced
    AI writer. Your job is to help
    write examples of texts that
    convey empathy or not.

Prompt: The following text has a {
    empathy} flag for expressing
    empathy, write 9 new semantically
     similar examples that show the
    same intent and empathy flag.

Text: {text}

Answer:
```

Politeness

```
System prompt: You are an advanced
    AI writer. Your job is to help
    write examples of social media
    comments that convey politeness
    or not.

Prompt: The following social media
    text has a {politeness} flag for
    politeness, write 9 new
    semantically similar examples in
    the style of a social media
    comment, that show the same
    intent and politeness flag.

Text: {text}

Answer:
```

Hyperbole

```
System prompt: You are an advanced
    AI writer. You are tasked with
    writing examples of sentences
    that are hyperbolic or not.

Prompt: The following sentence has a
     {hypo} flag for being hyperbolic
    . Write 9 new semantically
    similar examples that show the
    same intent and hyperbolic flag.

Text: {text}

Answer:
```

Intimacy

```
System prompt: You are an advanced
    AI writer. Your job is to help
    write examples of questions
    posted on social media that
    convey certain levels of intimacy
    . The intimacy levels are: very
    intimate, intimate, somewhat
    intimate, not very intimate, not
    intimate, not intimate at all.

Prompt: The following social media
    question conveys the {intimacy}
    level of question intimacy. Write
     9 new semantically similar
    examples in the style of a social
     media question, that show the
    same intent and intimacy level.

Text: {text}

Answer:
```

Same side stance

```
System prompt: You are an advanced
    AI writer. Your job is to help
    write examples of questions
    posted on social media that
    convey certain levels of intimacy
    . The intimacy levels are: very
    intimate, intimate, somewhat
    intimate, not very intimate, not
    intimate, not intimate at all.

Prompt: The following social media
    question conveys the {intimacy}
    level of question intimacy. Write
     9 new semantically similar
    examples in the style of a social
     media question, that show the
    same intent and intimacy level.

Text: {text}

Answer:
```

Condescension

```
System prompt: You are an advanced
    AI writer. Your job is to help
    write examples of social media
    comments that convey
    condescendence or not.

Prompt: The following social media
    text has a {talkdown} flag for
    showing condescendence, write 9
    new semantically similar examples
     in the style of a social media
    comment, that show the same
    intent and condescendence flag.

Text: {text}

Answer:
```

A.2 Zero-shot classification

Sentiment

```
System prompt: You are an advanced
    classifying AI. You are tasked
    with classifying the sentiment of
     a text. Sentiment can be either
    positive, negative or neutral.

Prompt: Classify the following
    social media comment into either
        negative    ,    neutral    or
        positive    . Your answer MUST
    be either one of ["negative", "
    neutral", "positive"]. Your
    answer must be lowercase.

Text: {text}

Answer:
```

Hate-speech

```
System prompt: You are an advanced
    classifying AI. You are tasked
    with classifying whether a text
    is offensive or not.

Prompt: The following is a comment
    on a social media post. Classify
    whether the post is offensive (
    OFF) or not (NOT). Your answer
    must be one of ["OFF", "NOT"].

Text: {text}

Answer:
```

Social dimensions

```
System prompt: You are an advanced
    classifying AI. You are tasked
    with classifying the social
    dimension of a text. The social
    dimensions are: social support,
    conflict, trust, neutral, fun,
    respect, knowledge, power, and
    similarity/identity.

Prompt: Based on the following
    social media text, classify the
    social dimension of the text. You
     answer MUST only be one of the
    social dimensions. Your answer
    MUST be exactly one of ["
    social_support", "conflict", "
    trust", "neutral", "fun", "
    respect", "knowledge", "power", "
    similarity_identity"]. The answer
     must be lowercase.

Text: {text}

Answer:
```

Emotions

```
System prompt: You are an advanced
    classifying AI. You are tasked
    with classifying the emotion of a
     text. The emotions are: sadness,
     enthusiasm, empty, neutral,
    worry, love, fun, hate, happiness
    , relief, boredom, surprise,
    anger.

Prompt: Based on the following
    social media text, classify the
    emotion of the text. You answer
    MUST only be one of the emotions.
     Your answer MUST be exactly one
    of ['sadness', 'enthusiasm', '
    empty', 'neutral', 'worry', 'love
    ', 'fun', 'hate', 'happiness', '
    relief', 'boredom', 'surprise', '
    anger']. The answer must be
    lowercased.

Text: {text}

Answer:
```

Empathy

```
System prompt: You are an advanced
    classifying AI. You are tasked
    with classifying whether the text
     expresses empathy.

Prompt: Based on the following text,
     classify whether the text
    expresses empathy or not. You
    answer MUST only be one of the
    two labels. Your answer MUST be
    exactly one of ['empathy', 'not
    empathy']. The answer must be
    lowercased.

Text: {text}

Answer:
```

Politeness

```
System prompt: You are an advanced
    classifying AI. You are tasked
    with classifying the whether the
    text is polite or impolite.

Prompt: Based on the following text,
     classify the politeness of the
    text. You answer MUST only be one
     of the two labels. Your answer
    MUST be exactly one of ['impolite
    ', 'polite']. The answer must be
    lowercased.

Text: {text}

Answer:
```

Hyperbole

```
System prompt: You are an advanced
    classifying AI. You are tasked
    with classifying the whether the
    text is a hyperbole or not a
    hyperbole.

Prompt: Based on the following text,
     classify the text is a hyperbole
    . You answer MUST only be one of
    the two labels. Your answer MUST
    be exactly one of ['hyperbole', '
    not hyperbole']. The answer must
    be lowercased.

Text: {text}

Answer:
```

Intimacy

```
System prompt: You are an advanced
    classifying AI. You are tasked
    with classifying the intimacy of
    the text. The different
    intimacies are 'Very intimate', '
    Intimate', 'Somewhat intimate', '
    Not very intimate', 'Not intimate
    ', and 'Not intimate at all'.

Prompt: Based on the following text,
     classify how intimate the text
    is. You answer MUST only be one
    of the six labels. Your answer
    MUST be exactly one of ['Very-
    intimate', 'Intimate', 'Somewhat-
    intimate', 'Not-very-intimate', '
    Not-intimate', 'Not-intimate-at-
    all'].

Text: {text}

Answer:
```

Same side stance

```
System prompt: You are an advanced
    classifying AI. You are tasked
    with classifying whether two
    texts, separated by [SEP], convey
     the same stance or not. The two
    stances are 'not same side' and '
    same side'.

Prompt: Based on the following text,
     classify the stance of the text.
     You answer MUST only be one of
    the stances. Your answer MUST be
    exactly one of ['not same side',
    'same side']. The answer must be
    lowercased.

Text: {text}

Answer:
```

Condescension

```
System prompt: You are an advanced
    classifying AI. You are tasked
    with classifying if the text is
    condescending or not
    condescending.

Prompt: Based on the following text,
     classify if it is condescending.
     You answer MUST only be one of
    the two labels. Your answer MUST
    be exactly one of ['not
    condescension', 'condescension'].

Text: {text}

Answer:
```

B Performance reports

This section includes a detailed performance report. Table 2 describes the performance of classification models trained on the full human-labeled dataset and the full LLMs-augmented datasets. We also report the zero-shot performance of GPT-4 and Llama-2 as a reference.
Given the mentioned presence of class imbalance for some of the considered tasks, we provide a general overview of label distributions per class in the training data (cf. Figure 3). Detailed class-wise classification reports for all considered models for the ten tasks of references are reported in the Supplementary Material.

C Diversity

We have conducted an investigation into the diversity between the primary dataset and the data generated synthetically by large language models (LLMs) for the 10 tasks of reference. We have employed token overlap as an indicator of lexical diversity and cosine similarity as a gauge of semantic diversity. Our findings reveal that the synthetic data, generated from both GPT-4 and Llama-2, exhibits substantial lexical differentiation from the original samples while preserving semantic similarity. Notably, Llama-2 displays a more pronounced level of diversity compared to GPT-4, as demonstrated by lower values in both token overlap and cosine similarity metrics (refer to Figure 4 for further details).

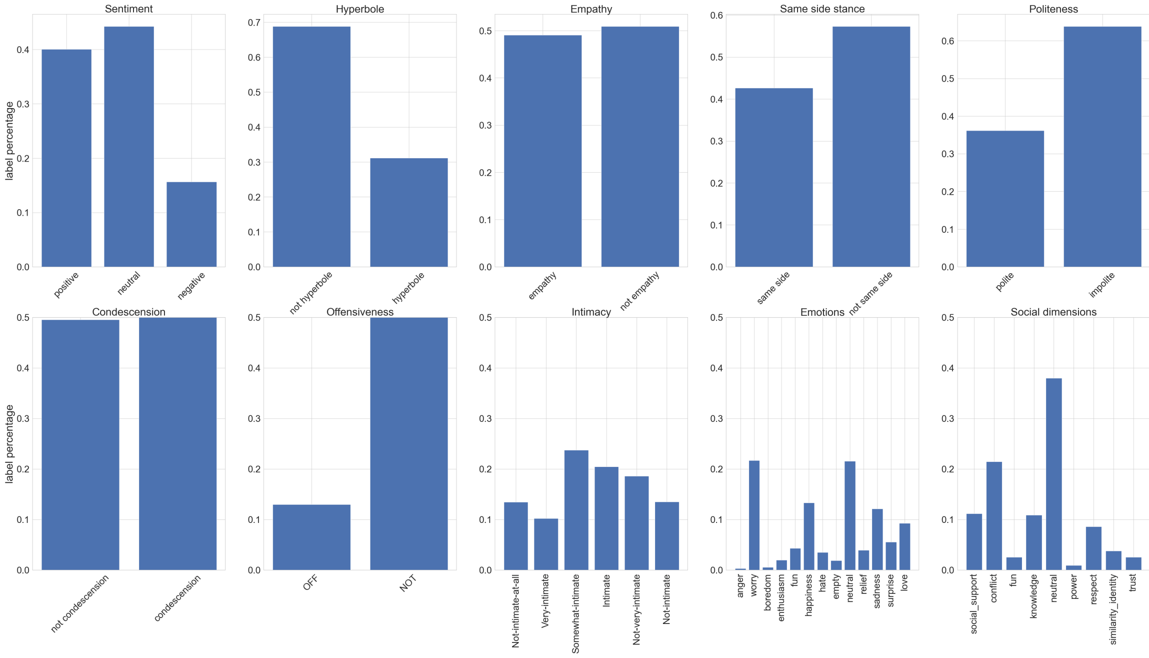

Figure 3: Class distribution per task.

	Individual			**Zero-shot**	
	Crowdsourced	GPT-4 synthetic	Llama-2 synthetic	GPT-4	Llama-2
Sentiment	0.6901	0.6430	0.6020	0.7126	0.5998
Hyperbole	0.7163	0.6768	0.6570	0.6781	0.5894
Empathy	0.6268	0.6135	0.6157	0.6488	0.6233
Same side stance	0.3462	0.6443	0.4926	0.9403	0.9403
Politeness	0.8266	0.8970	0.7480	0.8982	0.9884
Condescension	0.8391	0.7295	0.7070	0.6362	0.4563
Offensiveness	0.7764	0.5698	-	0.7170	-
Intimacy	0.4864	0.4093	0.3738	0.0285	0.1445
Emotions	0.1452	0.1578	0.1911	0.1247	0.1681
Social dimensions	0.2551	0.3002	0.3038	0.3042	0.2765

Table 2: Macro F1 score of classification models trained on the full human-labeled dataset, the full LLMs-augmented dataset (**Individual** datasets) for the three computational social science tasks of interest. **Zero-shot** performance of GPT-4 and Llama-2 is also provided.

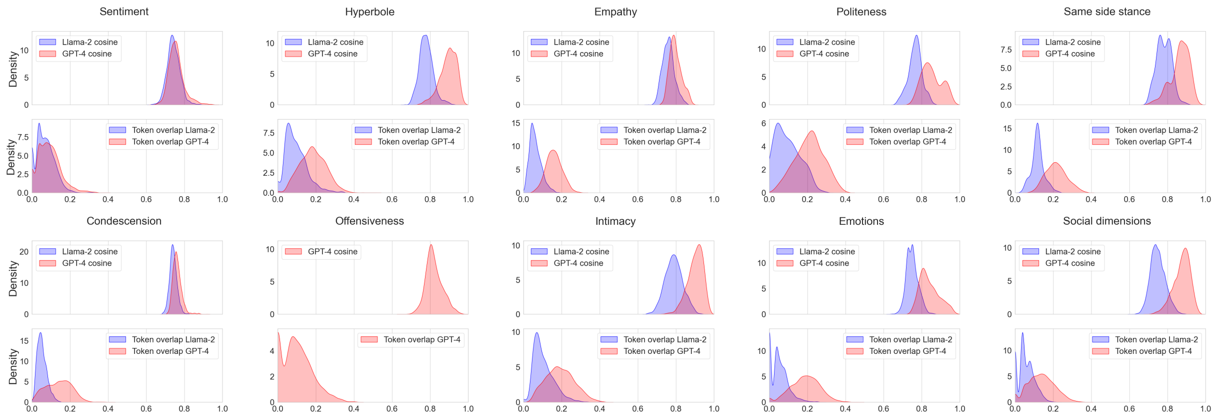

Figure 4: Lexical and semantic diversity between original and synthetically generated data, in terms of GPT-4 and Llama-2 models.