Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks

Anders Giovanni Møller
IT University of Copenhagen
agmo@itu.dk

Jacob Aarup Dalsgaard
IT University of Copenhagen
jacd@itu.dk

Arianna Pera
IT University of Copenhagen
arpe@itu.dk

Luca Maria Aiello
IT University of Copenhagen
luai@itu.dk

Abstract

Obtaining and annotating data can be expensive and time-consuming, especially in complex, low-resource domains. We use GPT-4 and ChatGPT to augment small labeled datasets with synthetic data via simple prompts, in three different classification tasks with varying complexity. For each task, we randomly select a base sample of 500 texts to generate 5,000 new synthetic samples. We explore two augmentation strategies: one that preserves original label distribution and another that balances the distribution. Using a progressively larger training sample size, we train and evaluate a 110M parameter multilingual language model on the real and synthetic data separately. We also test GPT-4 and ChatGPT in a zero-shot setting on the test sets. We observe that GPT-4 and ChatGPT have strong zero-shot performance across all tasks. We find that data augmented with synthetic samples yields a good downstream performance, and particularly aids in low-resource settings, such as in identifying rare classes. Human-annotated data exhibits a strong predictive power, overtaking synthetic data in two out of the three tasks. This finding highlights the need for more complex prompts for synthetic datasets to consistently surpass human-generated ones.

1 Introduction

Natural language processing (NLP) has recently attracted significant attention due to the development of *large language models* (LLMs) such as OpenAI's ChatGPT and GPT-4. These models are highly versatile, with easy interaction interfaces and the ability to generate high-quality, human-like text, which has made them appealing to a broad audience, including non-experts. LLMs have demonstrated impressive performance across a range of tasks, including solving math problems and university entry exams, generating code, performing well on zero-shot classification tasks, and even writing emails, poems, and articles (Bubeck et al., 2023a).

Researchers have recently shown that LLMs can outperform human crowd workers in data labeling (Gilardi et al., 2023; He et al., 2023), highlighting their potential for automating labor-intensive annotation tasks and improving label quality. Automating the generation of high-quality labels creates an opportunity for cheaply producing training data for supervised learning. However, generating training sets by directly annotating data with LLMs may not be effective when rare classes are difficult to find or in privacy-sensitive tasks where data cannot be processed through the APIs of proprietary LLMs. This raises the question of whether LLMs can effectively *augment* existing training datasets.

In this study, we investigate the use of LLMs, specifically GPT-4 and ChatGPT, for generating synthetic training data by augmenting small sets of human-generated training samples. We experiment with three NLP classification tasks within the domain of computational social science, with increasing levels of complexity. We use the synthetic data to train a 110M parameter multilingual model (Wang et al., 2022a) using increasing amounts of synthetic data, and compare their performance with models trained on the same amount of human-labeled data. Additionally, we also evaluate GPT-4 and ChatGPT in a zero-shot classification setting, where only a brief or no explanation of the labels is provided.

Our contribution includes new experiments on the use of LLMs for data augmentation for obtaining high-quality labeled data in low-resource settings using simple prompts, and provides insights into the potential benefits of using synthetic data in such settings. We show that even with simple prompts, synthetic training data generated by LLMs can be used to train smaller in-house models that achieve comparable or better performance than large LLMs. However, human-annotated data

performs better than synthetic data in some tasks, indicating the need for further research into crafting prompts for the data augmentation process that can generate more diverse and informative examples.

Our findings have important implications for research units with limited resources, as our approach allows for the acquisition of high-quality data quickly and at a low cost. Specifically, in the domain of computational social science that we address in this work, collecting examples of linguistic expressions of complex psycho-social dimensions can pose a challenge since some of these dimensions occur rarely. In such cases, LLMs can play a critical role in augmenting diverse data synthetically, which can be used downstream for task-specific fine-tuning.

2 Related work

In recent years, researchers have found that scaling *pre-trained language models* (PLMs) models dramatically improves their performance (Kaplan et al., 2020), leading to the development of billion-sized models, also known as *large language models* (LLMs). These models range from "small" 7B parameter-sized models like Llama (Touvron et al., 2023) and Alpaca (Taori et al., 2023), to mid-size and InstructGPT/ChatGPT (Ouyang et al., 2022) and GPT-3 (Brown et al., 2020) with 175B parameters, up to very large models like PaLM (Chowdhery et al., 2022) with a staggering 540B parameters. LLMs show impressive capabilities in in-context learning and zero-shot tasks.

One notable technique that has enabled the development of effective dialogue agents like ChatGPT is Reinforcement Learning with Human Feedback (RLHF) (Ziegler et al., 2020; Ouyang et al., 2022; Lambert et al., 2022). In this approach, human evaluations of LLM output responses are used to train a separate language model, called the reward model, which is then used in a reinforcement learning feedback loop to train the LLMs for optimal dialogue and produce output that aligns with human expectations.

The current research on LLMs is diverse and manifold. Microsoft call it the *spark of Artificial General Intelligence* after early experiments with GPT-4 (Bubeck et al., 2023a). They demonstrate remarkable performance in a wide range of tasks, and foresee models like GPT-4 (OpenAI, 2023) can greatly make major sectors more efficient, including healthcare, education, engineering, arts, and sciences. Recently we also saw BloombergGPT (Wu et al., 2023), a 50B parameter model, that has been specifically designed and trained for the financial market. In a joined effort with OpenResearch and University of Pennsylvania, OpenAI explored the use of LLMs and related technologies in the labor market and estimated that 80% of the U.S. workforce could have at least 10% of their work affected by the use of ChatGPT-like models (Eloundou et al., 2023). As large language models suffer from hallucination and toxic behavior as a consequence of the data used in the pre-training phase, a lot of effort is and will be put into reducing unintended output and bias (Bubeck et al., 2023b). A very recent method is red-teaming (Perez et al., 2022; Ganguli et al., 2022), an approach to find model vulnerabilities that potentially lead to unintended behavior. The idea and goal behind red-teaming are to create prompts that trigger the LLMs to generate harmful outputs, similar to adversarial attacks (Rajani et al., 2023).

Gilardi et al. (2023) use a sample of $2,382$ tweets to demonstrate how zero-shot classification with ChatGPT outperforms crowd-workers in four out of five tasks. He et al. (2023) propose AnnoLLM, a two-step '*explain-then-annotate*' framework for annotating text documents. In the first step, they create prompts for every demonstrated example, which is used to provide an explanation for the gold standard labels. Next, they construct few-shot chain-of-though (Wei et al., 2023) prompts with the generated explanation. Finally, these prompts are employed to annotate the data. These approaches differ from our work, as we investigate the use of LLMs to generate new data in low-resource settings, rather than annotating existing data.

3 Methods

We present an experiment spanning three tasks of increasing difficulty within the domain of computational social science: sentiment analysis of English text (Rosenthal et al., 2017), detection of offensive language in Danish (Sigurbergsson and Derczynski, 2023), and detection of social dimensions of online conversations (Choi et al., 2020). We can express the difficulty of tasks in multiple dimensions (Table 1). Data for all tasks is publicly available. The experiment aims at testing whether models trained on human annotations outperform models trained on annotated data generated by ChatGPT and GPT-

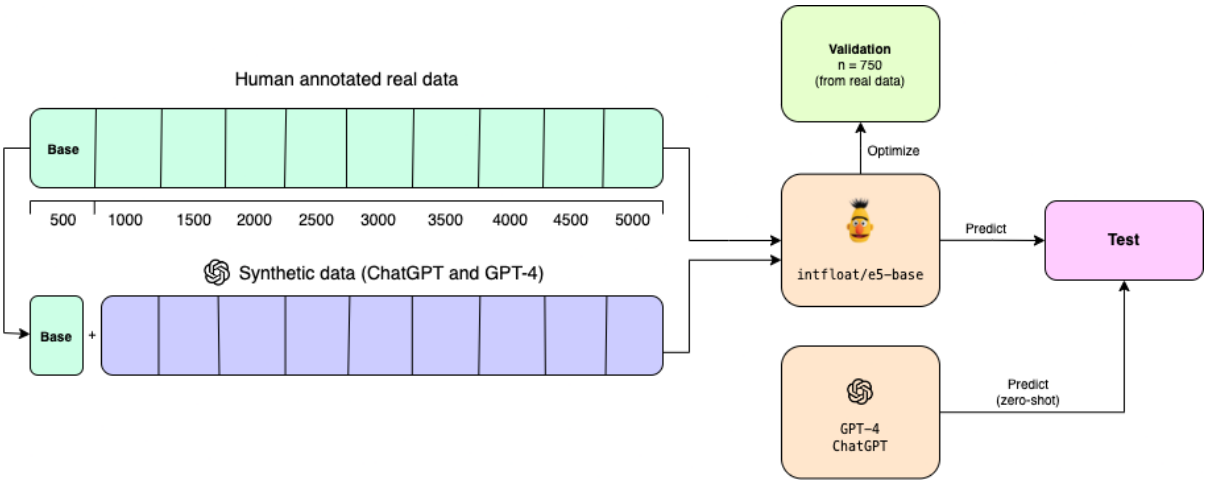

Figure 1: **`Experiment framework.`** We use three tasks that vary in complexity. To evaluate our models' performance, we reserve 20% of the original data in each task for testing purposes. Next, we randomly select a maximum of $5,000$ samples from each dataset and set aside 500 samples as a fixed base set. Using this base set, we generate $5,000$ new synthetic texts using GPT-4 and ChatGPT models, employing two distinct strategies. The first strategy is proportional, based on retaining the original label distribution, and the second is balanced, based on sampling data to obtain an approximately equal label distribution. We end up with five different datasets for each task, and we evaluate each of them by training `intfloat/e5-base` with various sample sizes. Specifically, we begin with a 500-sample dataset containing only the base set and gradually increase the sample size by 500 in each subsequent experiment. We choose the model with the lowest evaluation loss and then evaluate its performance on the reserved test set. For the zero-shot classification task using OpenAI's GPT-4 and ChatGPT models, we present the model with a text and a set of possible labels and ask it to classify the text accordingly.

4, as well as assessing the zero-shot performance of ChatGPT and GPT-4. We propose an experiment framework (Figure 1) that uses LLMs to generate a large amount of labeled data from a small base dataset by prompting the models to generate new training examples that resemble examples from the base set. Additionally, we create prompts for zero-shot classification based on the tasks. We report all the prompts we used in the Appendix.

Task	Language	Complex latent variable	Imbalance	Low resource	Problem size
Sentiment analysis					
Offensive Language	✓		✓	✓	
Social dimension		✓	✓	✓	✓

Table 1: **`Task difficulties.`** We describe the complexity of tasks according to a range of dimensions that complicate the classification task. Each task is marked if a certain dimension applies to the task. *Language*, whether the task is non-English. *Complex latent variable*, whether the task involves detecting complex pragmatics. *Imbalance*, whether the class distribution is heavily imbalanced. *Low resource*, whether task-specific data is scarce. *Problem size*, whether the task has many classes e.g. more than 5. This categorization of the task leads to a taxonomy of task difficulty.

3.1 Data sources

Below we briefly describe the data used in the project. For all tasks, we randomly sample 20% of the data to use as a test set or use the splits made by the authors of the datasets when available. We initially sample 500 texts to use as base data and, in the training phase, we randomly select 750 texts from the remaining training data to be used for validation. We use fixed seeds for consistency and reproducibility.

Sentiment analysis. The data used for the sentiment analysis task is the SemEval-2017 Task 4: Sentiment Analysis in Twitter (Rosenthal et al., 2017). We limit such data to only include the English task and we randomly sample a training set of 5000 examples. We sample the base set from this training set.

Hate speech. We collect DKHATE (Sigurbergsson and Derczynski, 2023), a Danish collection of user-generated texts containing offensive language from various online social platforms. The data is publicly released with train/test splits consisting of $2,960$ samples for training and 329 for testing. The label distribution is highly unbalanced with only

13% of the texts being offensive. This distribution applies to both the training and test splits. Despite this dataset having only $2,960$ samples, we still augment synthetic data up to $5,000$ samples for consistency.

Social dimensions. A dataset of $7,855$ texts from online social media, that multiple annotators have annotated with one or more complex "social dimensions" that represent conversational archetypes (e.g., whether a text conveys social support) (Choi et al., 2020). This naturally defines a multilabel and multiclass task. For the sake of simplicity, we transform the task into a multiclass problem by assigning a given label to a text if two or more annotators have assigned such a class. If a text is assigned to n classes with $n > 1$ according to two or more annotations, we replicate the text n times and assign each replica to one of the n classes; else, we assign a text to a *neutral* class. We also remove the *romance* label and merge *similarity* and *identity* as their semantics is naturally similar. This leaves us with nine highly unbalanced classes.

3.2 Data augmentation

We use the OpenAI API to send requests to GPT-4 and ChatGPT. We use the `langchain` LLM wrapper for python as our framework for prompting. As such, we construct prompts that take an example and its corresponding label and instruct the LLMs to generate 10 similar examples exhibiting the same label. We design the augmentation prompts to be as minimal as possible. However, in the task of social dimensions, we include in the prompts: *i)* a short description of the labels, as defined by the original authors (Choi et al., 2020), and *ii)* an instruction to write the examples in the style of social media comments.

We explore two different augmentation strategies: *proportional* and *balanced*. The proportional strategy generates the data in proportion to the class distribution from the base datasets, i.e. for each example in the base dataset, we generate 10 synthetic examples. We use standard hyperparameters for generation except for temperature, which is set to 0 for reproducibility purposes. The balanced strategy balances the class distribution during augmentation. As we do not have sufficient data in all classes to generate enough samples, we oversample minority classes to get a uniform distribution of labels. Then we use the generation procedure with a temperature of 1 to ensure that the synthetic examples generated from the oversampled classes will be substantially different. The data augmentation process results in four augmented datasets per task, one for combination of model (ChatGPT, GPT-4) and augmentation strategy (proportional, balanced).

3.3 Training Classifier

We use the Huggingface `Trainer` interface to train `intfloat/e5-base` (Wang et al., 2022b), a multilingual model that achieves state-of-the-art performance on a variety of tasks (Muennighoff et al., 2023), in several iterations on the different tasks and datasets. We train the model for 10 epochs using a batch size of 32. We use AdamW (Loshchilov and Hutter, 2019) as optimizer with a learning rate of 2^{-5}. We track evaluation performance for every epoch iteration, select the checkpoint with the lowest validation loss, and use to evaluate the test set. The test set is evaluated using macro F1 and accuracy. Details on all training runs and zero-shot performances are available in our Weights & Bias project[1]. All code can be found in our Github repository[2].

4 Results

In the `sentiment` task (Figure 2), synthetic data underperforms significantly the human-annotated data. As the original data is close to being evenly balanced, all the synthetic datasets achieve comparable performance with each other. The crowdsourced data performs close to equal with ChatGPT after being trained on 2000 training samples. GPT-4 achieves the highest performance with a macro F1 score and accuracy of 0.71. We observe little to no improvement for the crowdsourced model beyond a sample size of 2000 and only a slight correlation between sample size and performance for the synthetic data models. In Table 2 in the Appendix, we present the classification report of GPT-4 on the test set. The negative label has the highest F1-score of 0.76, primarily due to a high recall of 0.85. However, correctly classifying the positive label, which is the least represented in the test set, is the most challenging task for GPT-4, achieving a precision of 0.65.

The `hate-speech` task is more challenging than the `sentiment` task due to the Danish language, the

[1] https://wandb.ai/cocoons/worker_vs_gpt/
[2] https://github.com/AGMoller/worker_vs_gpt/

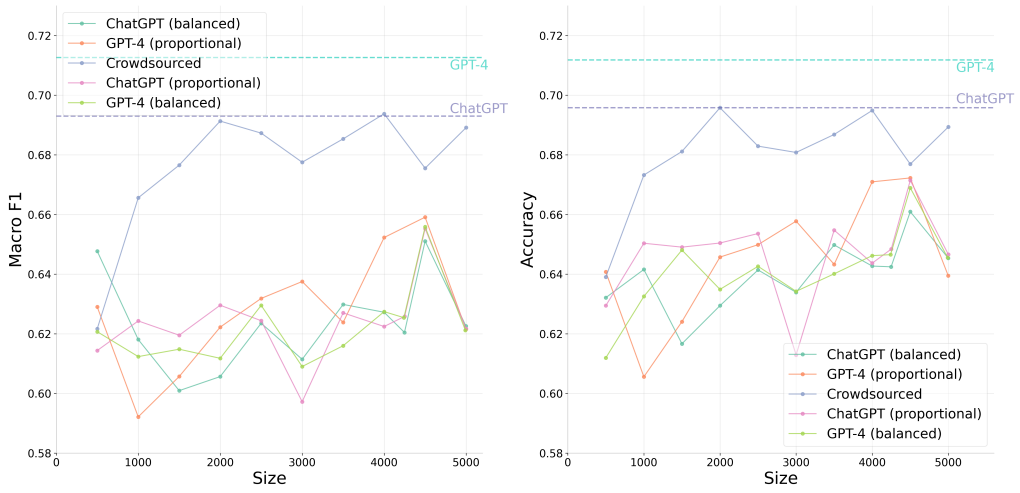

Figure 2: **`Sentiment.`** The Figure displays the macro F1 score and accuracy on the test set for various models trained on different sample sizes. The dashed lines represent the zero-shot performance on the test set for two models, ChatGPT and GPT-4. We observe the superior performance using the zero-shot approaches, specifically GPT-4 outperforms the other approaches. When training the models, a substantial gap between Crowdsourced and the augmented approaches is observed.

relatively small training set size ($2,460$ samples), and its unbalanced label distribution. Training the model on $2,000$ samples from the crowdsourced annotated data results in a macro-F1 score of 0.76 and an accuracy of 0.92 (Figure 3). However, due to the large class imbalance, correctly classifying the offensive label remains particularly challenging. Our model achieves high precision of 0.895 in predicting offensive language but suffers from the low recall of 0.415, resulting in an F1-score of 0.567 on the 41 samples in the test set. For the synthetic data, we observe little improvement beyond the first 1000 samples except for the proportional ChatGPT data. The zero-shot performance with GPT-4 showed a competitive macro-F1 score of 0.72. However, compared to the trained models, GPT-4 slightly underperforms alla other models in accuracy. This is due to ChatGPT's tendency to predict texts as containing offensive language, resulting in a high recall of 0.854 but a very low precision of 0.324. ChatGPT is even more likely to predict text as being offensive, which lowers its performance even further.

The `social dimension` classification is the most challenging problem due to the large number of complex and highly unbalanced classes. In this task, ChatGPT achieves the highest macro F1 score of 0.321, closely followed by GPT-4 with a score of 0.304 (Figure 4). Notably, both zero-shot approaches achieve higher macro F1 scores than any trained model, primarily due to their ability to correctly classify the *power* class, which is the most underrepresented class in the task and only occurs 13 times in the test set. ChatGPT correctly classifies only 2 out of the 13 examples, but since the macro-F1 score does not weigh by occurrence, it has a considerable impact on the overall score. We present the classification report of ChatGPT in Table 4 in the Appendix. The balanced data strategies are significantly more sample-efficient for training the models, as they increase the F1 considerably up to the first $2,000$ samples. The proportional augmentation and the crowdsourced data fill the performance gap at around $5,000$ samples. While F1 seems to stabilize for the balanced strategies, it is not clear whether more than $5,000$ samples could yield further improvements. For accuracy, the trend is very similar, but the balanced strategies perform slightly worse. Moreover, we observe that the combination of $4,500$ synthetic texts from both GPT-4 and ChatGPT outperforms the accuracy of zero-shot approaches while using $5,000$ samples of crowdsourced data yields the highest average accuracy of 0.405.

5 Discussion

Our study demonstrates the potential of using LLMs to generate synthetic data in sparse and low-

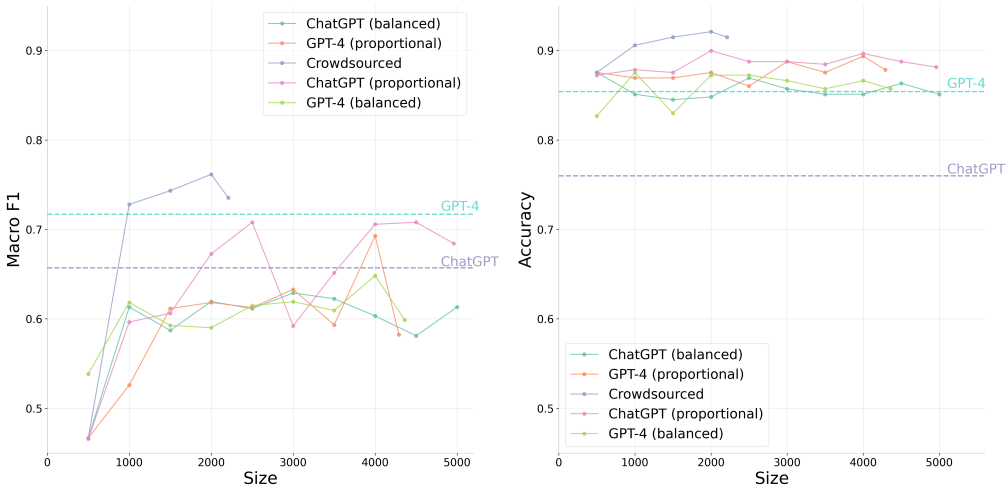

Figure 3: **`Hate Speech.`** Similar to Figure 2, the plot shows the macro F1 score on the test set for different training sample sizes. In this highly unbalanced dataset, we find the Crowdsourced model to be superior to both the zero-shot models and the models trained with synthetic data. The offensive language label is heavily underrepresented and is the reason why we observe the macro F1 drop. Interestingly, balancing the label distribution does not increase the performance, and keeping the label distribution proportional is slightly better. Comparing the two zero-shot approaches, we also find GPT-4 to be superior.

resource settings for three different NLP tasks of varying complexity. In the case of sentiment classification and hate-speech detection, we consistently observed that the synthetic data performed worse than the human-annotated data. However, for the social dimension classification task, using all 5,000 samples, the trained models achieved comparable performances, with a slight advantage to the crowdsourced data. We also observed that the zero-shot approaches obtained either the best or some competitive results. Particularly in the case of sentiment classification where GPT-4 was superior with a notable margin. We hypothesize that the reason why ChatGPT and GPT-4 perform exceptionally well on this task can be twofold. Firstly, the large difference in size between the LLMs compared to the standard LM could potentially allow them to learn more complex patterns and features in the data. Secondly, ChatGPT and GPT-4 may have already been exposed to this kind of data during their training. As the training details and datasets used to fine-tune the LLMs are not publicly available, it is plausible that they have been exposed to a wide range of sentiment classification examples, especially large public ones such as the SemEval tasks.

Irrespective of the selected method, each approach exhibits strengths and weaknesses. First, training models on synthetic data can be beneficial in scenarios where data acquisition and annotation are labor-intensive and expensive. This approach can substantially reduce costs and improve efficiency, particularly for low-resource companies or research units. It could also mitigate harmful side effects of crowdsourcing data (Williamson, 2016). Second, zero-shot approaches require no training and can achieve great performance, making them an attractive option for certain tasks. However, there are several drawbacks to using LLMs-based approaches. For instance, calling OpenAI's API incurs a cost based on the number of tokens produced. Additionally, using third-party services for tasks containing sensitive data may not be possible. Furthermore, as LLMs are autoregressive models that produce text output, there is no guarantee that the output labels will align with the true labels. In the task of social dimensions, ChatGPT produced the labels *appreciation*, *empowerment*, and *apology* despite being prompted to select from only nine different classes in which such example labels are not included. This is, however, ensured by using regular LMs with a dense output layer. Lastly, as these models are under constant development, it is challenging to ensure long-term consistency given the use of third-party services.

There are additional limitations we see in our

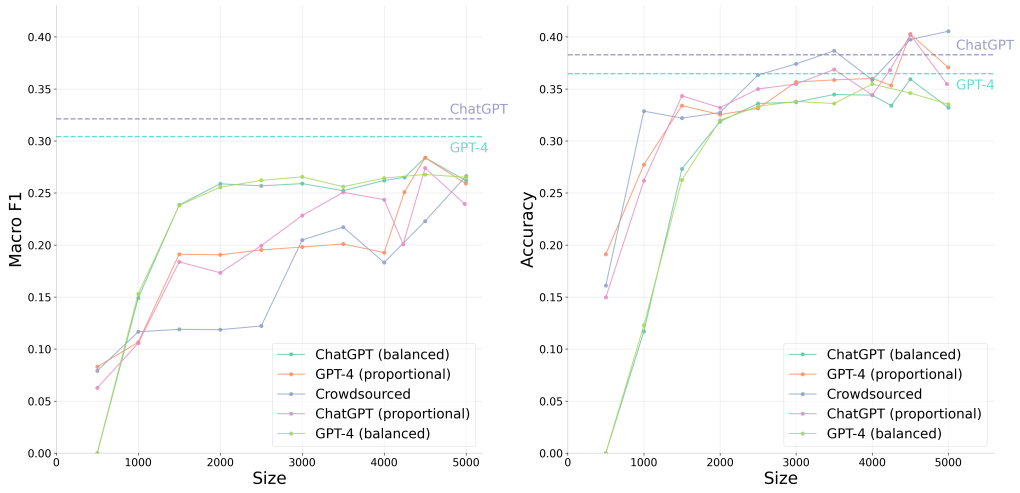

Figure 4: **`Social dimensions.`** The F1-scores indicate ChatGPT (balanced) and GPT-4 (balanced) consistently outperform the other trained models from $1,000$ to $3,000$ training samples. Notably, all models perform comparably when using the full set of $5,000$ samples, except for ChatGPT (proportional), which exhibits slightly lower performance with an F1 score of 0.24. Both zero-shot approaches outperform any trained model, with ChatGPT achieving the highest F1 score of 0.32. In the case of accuracy, using all $5,000$ samples from the mechanical Turk data yields the best overall score of 0.41. ChatGPT performs slightly better than GPT-4 (balanced), highlighting the difficulty of the task.

approach. The simple prompts we use were solely based on empirical best practices from various sources[3] at the time of development. Prompt engineering is a rapidly growing field in LLM research, and we did not optimize the prompts for data augmentation or zero-shot classification.

In the augmentation phase of the balanced data, we selected a temperature value of 1 as we sampled from the same text multiple times to achieve a balanced label distribution. This hyperparameter was based on small empirical experiments that qualitatively evaluated the semantic meaning of the generated text compared to the real data.

In the zero-shot classification prompt, we could have provided additional knowledge about the classes, including examples of each, transforming the task into a few-shot classification. Nonetheless, our objective was to design a simple and consistent framework to test the basic capabilities of the LLMs. Despite our simple approach, the results of our study provide valuable insights into the potential of LLMs for generating synthetic data in low-resource settings.

Using human annotators can have certain advantages, such as selecting diverse and demographically versatile annotators who can closely mimic the general perception of the tasks and classes. Constructing such a human-validated dataset additionally often involves thorough and detailed testing and evaluation of the annotators which implies a high level of quality. However, it is entirely possible to introduce this diversity into the prompt as LLMs are able to closely impersonate specific demographics given the right context (Argyle et al., 2022). In our experiments, we found that the human-annotated data consistently outperformed the augmented data in the sentiment and hate-speech tasks. However, in the most difficult task, social dimensions, the synthetic data performed comparably to the real data. We speculate that this might be because the prompt for generating the synthetic data was more detailed, including short descriptions of the social dimensions. This would also suggest that more careful prompt engineering might lead to better performance for easier tasks as well. In general, a thorough investigation is needed in order to understand how the synthetic examples differ from the original in a number of ways including length and style as well as lexical and semantic similarity. This would in turn inform us on how to create better and more diverse prompts for data generation.

OpenAI and other companies and research units

[3] https://www.promptingguide.ai/

that develop LLMs put significant effort into implementing safety protocols and bias regulations (Perez et al., 2022; Ganguli et al., 2022). LLMs are heavily evaluated on safety metrics such as toxicity and bias (Gehman et al., 2020; Nangia et al., 2020). As a result, when directly prompting ChatGPT or GPT-4 to generate hate-speech, the models will rightfully reject the request as it goes against OpenAI's moral and ethical principles. However, we were able to easily bypass this safety protocol for both ChatGPT and GPT-4 by using the following 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."*

This finding highlights the need for continued efforts to ensure that these models do not produce any harmful or biased output. While safety protocols and regulations are in place, further research is necessary to ensure that LLMs produce ethical and safe outputs in all scenarios.

6 Conclusion

We investigate the effectiveness of synthetic data augmentation utilizing GPT-4 and ChatGPT, comparing the outcomes to human-annotated data across three computational social science classification tasks with distinct levels of complexity. To evaluate the data, we train a 110M parameter model and assess GPT-4 and ChatGPT in a zero-shot classification context. We find solid zero-shot performance by both LLMs in all three tasks. We also observe how synthetic data can be useful in low-resource tasks with limited data availability. Furthermore, our results indicate that human-annotated data has predictive power, outperforming synthetic data in two out of the three tasks. We hypothesize that further refinement of our prompts may induce more diverse text output.

Acknowledgments

We acknowledge the support from the Carlsberg Foundation through the COCOONS project (CF21-0432). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript

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].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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.

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. 2023a. Sparks of Artificial General Intelligence: Early experiments with GPT-4. ArXiv:2303.12712 [cs].

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. 2023b. Sparks of artificial general intelligence: Early experiments with gpt-4.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models.

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.

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. (arXiv:2001.08361). ArXiv:2001.08361 [cs, stat].

Nathan Lambert, Louis Castricato, Leandro von Werra, and Alex Havrilla. 2022. Illustrating reinforcement learning from human feedback (rlhf). *Hugging Face Blog*. Https://huggingface.co/blog/rlhf.

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

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].

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

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].

Nanzneen Rajani, Nathan Lambert, and Lewis Tunstall. 2023. Red-teaming large language models. *Hugging Face Blog*. Https://huggingface.co/blog/red-teaming.

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.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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.

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. ArXiv:2212.03533 [cs].

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [cs].

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

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences.

Appendix

A `Sentiment.` GPT-4 classification report on test set

Label	Precision	Recall	F1	Support
negative	0.688	0.847	0.759	3972
neutral	0.779	0.598	0.677	5937
positive	0.646	0.769	0.702	2375
Accuracy			0.712	12284
Macro avg	0.704	0.738	0.713	12284
Weighted avg	0.724	0.712	0.708	12284

Table 2: Classification report from GPT-4 on the test set.

B `Hate-speech.` `intfloat/e5-base` trained on $2,000$ samples classification report on test set

Label	Precision	Recall	F1	Support
NOT	0.923	0.993	0.957	288
OFF	0.895	0.415	0.567	41
Accuracy			0.921	329
Macro avg	0.909	0.704	0.762	329
Weighted avg	0.919	0.921	0.908	329

Table 3: Classification report on the test set from `intfloat/e5-base` trained on $2,000$ samples.

C `Social dimensions.` ChatGPT classification report on test set

Label	Precision	Recall	F1	Support
conflict	0.509	0.642	0.568	321
fun	0.27	0.730	0.394	37
knowledge	0.349	0.540	0.424	163
neutral	0.445	0.192	0.274	570
power	0.056	0.154	0.081	13
respect	0.290	0.155	0.202	129
similarity/identity	0.292	0.339	0.314	56
social support	0.319	0.527	0.397	169
Accuracy			0.383	1497
Macro avg	0.308	0.391	0.321	1497
Weighted avg	0.402	0.383	0.363	1497

Table 4: Classification report from ChatGPT on the test set.

D Prompts

D.1 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 10 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 10
   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 10 new
   semantically similar
   examples in style of a
   social media comment, that
   show the same intent and
   social dimension.

Text: {text}

Answer:
```

D.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:
```