The Power of LLM-Generated Synthetic Data for Stance Detection in Online Political Discussions

Stefan Sylvius Wagner
Department of Computer Science
Heinrich Heine University Düsseldorf
stefan.wagner@hhu.de

Maike Behrendt
Department of Computer Science
Heinrich Heine University Düsseldorf
maike.behrendt@hhu.de

Marc Ziegele
Department of Social Sciences
Heinrich Heine University Düsseldorf
marc.ziegele@hhu.de

Stefan Harmeling
Department of Computer Science
Technical University Dortmund
stefan.harmeling@tu-dortmund.de

Abstract

Stance detection holds great potential to improve online political discussions by being deployed in discussion platforms for purposes such as content moderation, topic summarization or to facilitate more balanced discussions. Transformer-based models are typically employed directly for stance detection, requiring vast amounts of data. However, the wide variety of debate topics in online political discussions makes data collection particularly challenging. LLMs have revived stance detection, but their online deployment in online political discussions faces challenges like inconsistent outputs, biases, and vulnerability to adversarial attacks. We show how LLM-generated synthetic data can improve stance detection for online political discussions by using reliable traditional stance detection models for online deployment, while leveraging the text generation capabilities of LLMs for synthetic data generation in a secure offline environment. To achieve this, (i) we generate synthetic data for specific debate questions by prompting a Mistral-7B model and show that fine-tuning with the generated synthetic data can substantially improve the performance of stance detection, while remaining interpretable and aligned with real world data. (ii) Using the synthetic data as a reference, we can improve performance even further by identifying the most informative samples in an unlabelled dataset, i.e., those samples which the stance detection model is most uncertain about and can benefit from the most. By fine-tuning with both synthetic data and the most informative samples, we surpass the performance of the baseline model that is trained on true labels, while labelling considerably less data.

1 Introduction

With the recent advent of powerful generative Large Language Models (LLMs) such as ChatGPT, Llama [Touvron et al., 2023] and Mistral [Jiang et al., 2023], new ways of performing stance detection have opened up via zero-shot or chain-of-thought prompting. This is especially important in the area of online political discussion where topics are complex and labelled data is hard to come by. At the same time, an ever important use case in online political discussions is being able to use stance detection for an ongoing discussion to, e.g., suggest suitable comments for engagement between participants [Küçük and Can, 2020, Behrendt et al., 2024]. In the case of LLMs, while strong at analysing complex topics and at open-ended text generation, explicit classification can be inconsistent [Cruickshank and Xian Ng, 2023], they are prone to biases [Ziems et al., 2023] and open to adversarial attacks [Greshake et al., 2023]. More traditional stance detection models based on, e.g.,

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

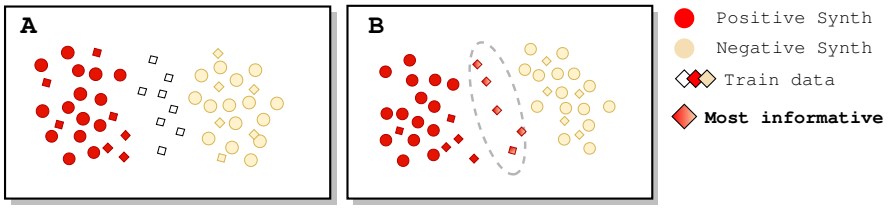

Figure 1: **We investigate the use of LLM-generated synthetic data for stance detection in online political discussions.** (A) We generate synthetic data ●○ for specific questions using a Mistral-7B model. The synthetic data is then used to fine-tune the stance detection model. We show that fine-tuning with synthetic data improves the performance of the model, since the synthetic data is roughly faithful to the real data's ◆◇ underlying distribution. However, some real world samples ◇ cannot be captured by the synthetic data. (B) We therefore use the synthetic data to identify the most informative samples ◆ in the unlabelled real data pool, which are better off labelled by human experts. Combining the synthetic data with the manually labelled most informative samples improves the performance of the model even further.

BERT [Devlin et al., 2019] are task-specific and therefore consistent in their output, however it is known that they need large amounts of labelled data [Mehrafarin et al., 2022, Vamvas and Sennrich, 2020] to perform well.

In this work, we combine both traditional stance dectection and LLMs to get the best of both worlds. For stance dectection, we use BERT as a lightweight stance detection model that produces fast and consistent output given the data it has been fine-tuned on. To address the issue of needing large amounts of data, we propose to generate synthetic data with an LLM to augment the stance detection model for fine-tuning. This allows us to leverage LLMs in an offline setting to enhance classical stance detection models, which are better suited and safer for use in an online setting.

We illustrate our method in Figure 1: We view stance detection as a binary classification problem (*favor* or against). **(Q1)** We first analyse whether fine-tuning the BERT model with synthetic data improves stance detection and demonstrate that this approach is superior to using zero-shot Mistral-7B. **(Q2)** Our second question analyses the generated synthetic data and how well it aligns with the real training data. We visualise the T-SNE projected embeddings of the stance detection model and find that the synthetic data aligns well with the real data, indicating that the LLM is able to generate comments for both stances and introducing minimal further bias. **(Q3)** Finally, the synthetic data allows us to identify unlabelled real data samples that improve the model even further through active learning. Due to the canonical nature of the synthetic data, we are able to extract real word samples for human labelling that are difficult (ambiguous) for the model to classify. We do this, by determining the k-nearest synthetic neighbours of the real data. The stance detection model is fine-tuned jointly with these samples and the synthetic data, where we surpass the baseline model even when it is fine-tuned on all true labels, while labelling considerably less data manually.

2 Method

Political discussions are typically centered around questions $q \in \mathcal{Q}$ (sometimes also called issues or targets). For stance detection, we usually have for each of these questions q a set of labelled data $\mathcal{D}^{(q)} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{I}$ where $x^{(i)} \in \mathcal{X}$ is a statement (or comment) and $y^{(i)}$ is the stance of the statement, with $y^{(i)} \in \{0, 1\} = \mathcal{Y}$. Note, that we use the notation $\mathcal{D}^{(q)}$ for labelled and for unlabelled datasets (then the labels are ignored). We view the stance detection model as a binary classification function $f : \mathcal{Q} \times \mathcal{X} \to \mathcal{Y}$, where we included the question as input to provide context. The stance detection model such as BERT [Devlin et al., 2019] is *fine-tuned* by minimizing the cross-entropy loss between the predicted labels $\hat{y}^{(i)} = f(q, x^{(i)})$ and the actual labels $y^{(i)}$.

2.1 Generating Synthetic Data for Stance Detection

To generate synthetic samples, we employ a quantized version of the Mistral-7B-instruct-v0.1 model to generate comments on a specific question q, using the following prompt:

```
A user in a discussion forum is debating other users about the following question:
[q] The person is in favor about the topic in question.  What would the person
write?  Write from the person's first person perspective.
```

where "`[q]`" must be replaced with the question q. Similarly, to generate a negative sample, we replace "`is in favor`" with "`is not in favor`". As in the X-Stance dataset [Vamvas and Sennrich, 2020], we assign the two labels 0 and 1. We denote the question-specific synthetic dataset as:

$$\mathcal{D}^{(q)}_{\text{synth}} = \left\{(x^{(m)}_{\text{synth}}, 1)\right\}_{m=1}^{M/2} \cup \left\{(x^{(m)}_{\text{synth}}, 0)\right\}_{m=1+M/2}^{M} \tag{1}$$

where half of the M synthetic data samples have *positive* labels, i.e., are comments in *favor* for the posed question, while the other half is *against*. Since the dataset is in German, we translate the questions q with a "NLLB-300M" [NLLB Team et al., 2022] translation model. The English answers from the Mistral-7B model are then translated back to German using the translation model. Overall, the generated dataset $\mathcal{D}^{(q)}_{\text{synth}}$ will be used in two ways: (i) to augment the existing dataset $\mathcal{D}^{(q)}$ in order to increase the amount of training data, and (ii) to detect the most informative samples in the unlabelled data pool, which is explained next.

2.2 Getting the Most Informative Samples: Synthetic Query By Comittee (SQBC)

To identify the ambiguous (most informative) samples as described in **(Q3)** we take from two active learning methods: Query by Comittee (QBC) [Seung et al., 1992] and Contrastive Active Learning (CAL, Margatina et al. [2021]). Instead of using QBC's ensemble of experts and the KL-divergence based information score in CAL, we directly use the synthetic data and its labels to identify ambigous samples using k nearest neighbors. The most informative samples are then the data points with the most indecisive scores. SQBC consists of three steps:

(1) Generate the embeddings. Given some embedding function $g : \mathcal{Q} \times \mathcal{X} \to \mathcal{E}$, we generate embeddings for the unlabelled data, $E = \left\{e^{(i)}\right\}_{i=1}^{I} = \left\{g(q, x^{(i)})\right\}_{i=1}^{I}$ and for the labelled synthetic data $E_{\text{synth}} = \left\{e^{(m)}_{\text{synth}}\right\}_{m=1}^{M} = \left\{g(q, x^{(m)}_{\text{synth}})\right\}_{m=1}^{M}$. Note that q is the question for which we generate the synthetic data and for which we want to detect the most informative samples. If obvious from the context, we often omit the superscript (q).

(2) Using the synthetic nearest neighbours as oracles to score the unlabelled data. For the i-th unlabelled embedding $e^{(i)}$ let $\text{NN}(i)$ be the set of indices of the k nearest neighbours (wrt. to the embeddings using the cosine similarity) among the labelled embeddings E_{synth}. The score for each unlabelled data point counts the number of labels $y^{(m)}_{\text{synth}} = 1$ among the nearest neighbours, i.e.,

$$s(i) = \sum_{m \in \text{NN}(i)} y^{(m)}_{\text{synth}} \in \{0, \dots, k\}. \tag{2}$$

For our experiments, we choose $k = M/2$ which worked well across all experiments (other values for k are possible, but did not lead to significantly better results).

(3) Choosing the most informative samples. The scores take values between 0 and k. For 0, the synthetic nearest neighbours all have labels $y^{(m)}_{\text{synth}} = 0$, for value k, all have labels $y^{(m)}_{\text{synth}} = 1$. The *most informative* samples have a score around $k/2$. We thus adjust the range of the scores so that values in the middle range have the smallest scores (close to 0). We do this by subtracting $k/2$ from the score and taking the absolute value,

$$s'(i) = |s(i) - k/2|. \tag{3}$$

The J most informative samples $\mathcal{D}^{(q)}_{\text{MInf}} \subset \mathcal{D}^{(q)}$ among the unlabelled samples are the J samples with the smallest scores. In the experiments we vary J to study the impact of manually labelled most informative samples. Finally, the most informative samples are labelled by a human expert.

3 Experiments

3.1 Datasets

X-Stance dataset. We evaluate on the German dataset of the X-Stance dataset [Vamvas and Sennrich, 2020], which contains $48,600$ annotated comments on many policy-related questions, answered by Swiss election candidates. The comments are labelled either as being in *favor* (positive) or *against* (negative) the posed question. The dataset is split in training and testing questions, i.e,. a question in the training dataset does not appear in the test dataset. Furthermore, for each question q from the training data, there are several annotated comments, which form the dataset $\mathcal{D}^{(q)}_{\text{train}}$. Similarly, for each question q in the test data, the set of annotated comments is written as $\mathcal{D}^{(q)}_{\text{test}}$. To refer to the whole training dataset we write $\mathcal{D}_{\text{train}} = \cup_{q\in\mathcal{Q}}\mathcal{D}^{(q)}_{\text{train}}$. At test-time, we fine-tune all stance detection models for each question separately allowing for better performance since the data distributions can vary greatly between questions. This is also being a common scenario for downstream tasks in online political discussions. To limit computation time, we selected 10 questions from the test dataset to evaluate our method, which best reflect the variability of the data (see Appendix H.1).

Synthetic dataset. For synthetic data-augmentation and active learning based on SQBC (see Section 2.2) we generate synthetic datasets of varying sizes $M = \{200, 500, 1000\}$ for each of the 10 questions. The synthetic data follows the same structure as the data from the X-stance dataset, where for a specific question q we have M comments and M labels. Each set contains $M/2$ positive labels and $M/2$ negative labels, i.e., the synthetic data is balanced. We show samples of the synthetic data in Appendix I.

3.2 Experimental Setup

General setup. For all experiments, we start with a pre-trained BERT base model and adapt to the stance detection task by fine-tuning on the X-Stance training dataset $\mathcal{D}_{\text{train}}$ (all questions). We call this the **Baseline** since it is the vanilla BERT-based stance detection (e.g., Vamvas and Sennrich [2020]).

We evaluate our fine-tuning with synthetic data and SQBC according to the questions proposed in Section 1: **(Q1)**: we analyse the effect of fine-tuning the BERT model with synthetic data and compare it to the BERT model that was only fine-tuned on D_{train}. **(Q2)**: we analyse the synthetic data by projecting the CLS embeddings of the BERT model with T-SNE and visualise its alignment with the real data. Furthermore, we also visualise the embeddings of the most informative samples selected by the active learning methods. **(Q3)**: we combine fine-tuning the synthetic data and the most informative samples.

Baselines. We fine-tune each method separately on each of the 10 questions of $\mathcal{D}^{(q)}_{\text{test}}$:

Baseline: this is the default model only trained on $\mathcal{D}_{\text{train}}$, (e.g., Vamvas and Sennrich [2020]).
True Labels: we fine-tune **Baseline** on the true labels of $\mathcal{D}^{(q)}_{\text{test}}$.
Random, CAL: we use the active learning approaches to get the most informative samples $\mathcal{D}^{(q)}_{\text{MInf}}$.

Our methods.

Baseline+Synth: we fine-tune the **Baseline** on the synthetic data $\mathcal{D}^{(q)}_{\text{synth}}$.

True Labels+Synth: we fine-tune **True Labels** additionally on the synthetic data $\mathcal{D}^{(q)}_{\text{synth}}$.

SQBC+Synth: we apply our active learning approach to get the most informative samples $\mathcal{D}^{(q)}_{\text{MInf}}$.

For further experimental details we refer to Appendix G.

3.3 Results

In Figure 2, we show that fine-tuning with synthetic data only **(Q1)**, improves the stance detection model (see *No Act. Learning*). For $M = 1000$ the performance almost reaches the **True Labels** model. In Appendix C, we also analyse stance detection with zero-shot and fine-tuning approaches on Mistral-7B. This proved challenging, since the model frequently produced inconsistent outputs or

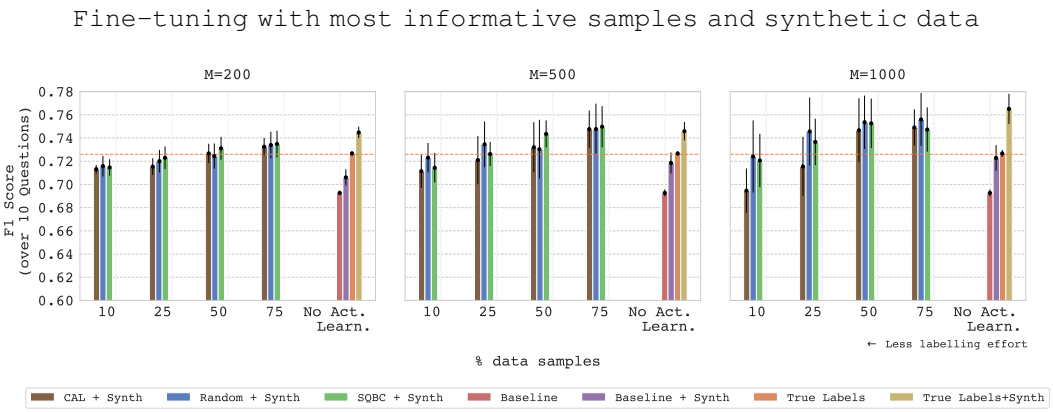

Figure 2: **Fine-tuning with synthetic data improves stance detection, while combining most informative samples and synthetic data surpasses the baseline model fine-tuned with all true labels using less manually labelled data:** The reason for the performance increase can be attributed to two phenomena: (i) the synthetic data helps the model learn the underlying distribution. (ii) The most informative samples improve the model where the synthetic data distribution is not expressive enough.

failed to predict a stance altogether. Our findings suggest that utilizing the LLM for open-ended text generation is more effective than trying to constrain it to produce a specific output, when using the rather complex X-Stance dataset.

We compare the T-SNE projected embeddings of the synthetic and real data **(Q2)** in Figure 3(A) (more visualisations in Appendix E). The synthetic data aligns well with the real world dataset, interpolating between the real samples (see 3(B)). Interestingly, the synthetic comments serve as a reference distribution since both classes are well separated. This allows us to use the synthetic data to identify ambiguous samples that are the most informative for the model. As a sanity check we also manually inspect some of the generated comments in Appendix I and see that the generated comments for both classes are generally of high quality, validating the capabilities of LLMs to produce open-ended text.

Finally **(Q3)**, we show the results of combining the most informative samples and synthetic data in Figure 2. Combining both, we oupterform **True Labels** while using *only* 25% of the labelled data. We compare the selection strategy of the methods in Figure 3(B): Due to the k-nearest neighbours objective of **SQBC**, the model selects samples that are in between the two classes, which proves superior to **CAL** and to **Random** for smaller synthetic data sizes. **CAL** performs the worst across the board: it assumes that similar embeddings that have different outputs are ambiguous, which makes it prone to outliers in the real data, e.g., when the stance detection model misclassifies a sample. Therefore, **CAL** often selects samples from only one class which worsens performance. Interestingly for $M = 1000$, **Random** outperforms both active learning methods **SQBC** and **CAL**. **Random** selects similar samples to **SQBC**, but also uniformly samples from outliers from both classes, extending the decision boundary of the model. We argue this is especially effective for larger synthetic dataset sizes where the synthetic data smoothens the decision boundary and thus mitigates the high variance introduced by the most informative samples. Thus, the model remains robust while extending the decision boundary. However, with severe outliers present, **Random** could select these and worsen performance. This would not happen with **SQBC** due to its k-nearest neighbour objective.

Observations and limitations. We observe in Figure 2, that the standard deviation increases with larger synthetic dataset size. Furthermore, there is an even sharper increase in the standard deviation when training with both the most informative samples and synthetic data. One reason for the former is due to increased variability introduced by a larger synthetic dataset. For the latter, when training with both the most informative samples and synthetic data, the synthetic dataset size relative to the real training dataset size seems to render the model more sensitive to different splits of the real training data (note that for every seed we also train on a different real training data split). We argue that the real data seems to be crucial for model generalisation, i.e., what decision boundary is learned.

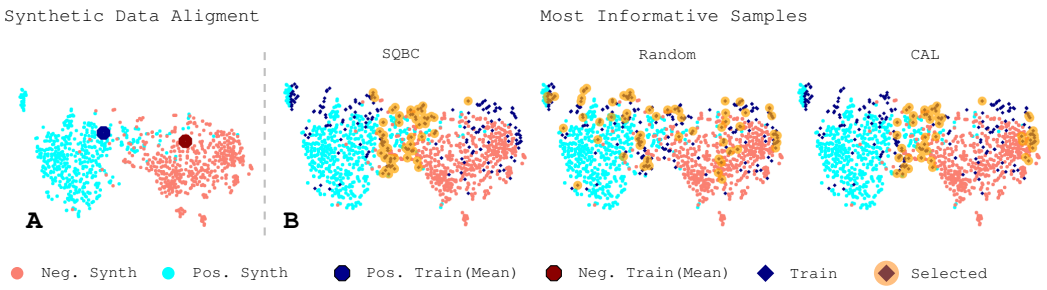

Figure 3: **Analysing the synthetic data (M=1000):** The synthetic data aligns well with the real data, which is crucial for improve stance detection performance and to check for potential biases introduced by the synthetic data. SQBC selects the samples that are in between the two classes, i.e, that are the most ambiguous and informative for the model.

This can also be seen in the T-SNE visualisations in Figure 3(B), where the synthetic data is well separated, while the most informative samples are mostly in between, thus affecting the model's decision boundary.

One limitation of our approach is that we fine-tune a separate model for each question. While this leads to good results, a common approach is to fine-tune a single (and thus more general) model for several questions (like pre-training **Baseline**). However, visualising the synthetic data in Figure 3 and Appendix E, we observe that the underlying data distribution differs (sometimes greatly) for each question, which strongly suggests that each question benefits from fine-tuning a different model. This also aligns well with the per topic setting of online (political) discussions, considering that lightweight stance detection models can be fine-tuned in less than a minute even with a synthetic dataset size of $M = 1000$ on a reasonable GPU (NVIDIA A100). In any case, for both our single-question and a multiple-question model the process of generating and fine-tuning with synthetic data remains the same. Future research, could extend our work to fine-tuning more general models on synthetic data and the most informative samples.

Another concern are biases that could be potentially introduced through the synthetic data. We addressed this in Section 3 by comparing the distributions of the synthetic data and real world data. Similarly to the above, analysing potential biases that could be introduced to the stance detection model through the synthetic data is easier in a single-question setting. In a multiple-question setting data from other topics could introduce biases into the model that are harder to detect. We argue this could be true for both real and synthetic data. Future work, could study how to use our active learning approach to detect certain type of comments that are of specific interest such as low quality comments while using the synthetic data as reference distribution. In any case, we consider an analysis of the synthetic data is required when using different LLMs. While our generated synthetic data worked well for the online political discussion setting, we cannot make a general assessment on the quality of synthetic data for future models.

4 Conclusion

In this work, we presented how to improve stance detection models for online political discussions utilizing LLM-generated synthetic data: (i) we showed that fine-tuning with synthetic data related to the question improves the performance of the stance detection model. (ii) We attribute this to the LLM-generated synthetic data aligning well with the real data for the given question. (iii) Fine-tuning with synthetic data can be further enhanced by adding the most informative samples, outperforming the method that uses all true labels. Our findings demonstrate the potential of synthetic data to improve the efficiency and effectiveness of stance detection in online political discussions.

References

Abeer ALDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102597. URL https://www.sciencedirect.com/science/article/pii/S0306457321000960.

Maike Behrendt, Stefan Sylvius Wagner, and Stefan Harmeling. Supporting online discussions: Integrating ai into the adhocracy+ participation platform to enhance deliberation, 2024. URL https://arxiv.org/abs/2409.07780.

Michael Burnham. Stance Detection: A Practical Guide to Classifying Political Beliefs in Text. *arXiv e-prints*, art. arXiv:2305.01723, May 2023. doi: 10.48550/arXiv.2305.01723.

Alan Darmasaputra Chowanda, Albert Richard Sanyoto, Derwin Suhartono, and Criscentia Jessica Setiadi. Automatic debate text summarization in online debate forum. *Procedia Computer Science*, 116:11–19, 2017. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2017.10.003. URL https://www.sciencedirect.com/science/article/pii/S1877050917320409. Discovery and innovation of computer science technology in artificial intelligence era: The 2nd International Conference on Computer Science and Computational Intelligence (ICCSCI 2017).

Iain J. Cruickshank and Lynnette Hui Xian Ng. Prompting and Fine-Tuning Open-Sourced Large Language Models for Stance Classification. *arXiv e-prints*, art. arXiv:2309.13734, September 2023. doi: 10.48550/arXiv.2309.13734.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, AISec '23, page 79–90, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702600. doi: 10.1145/3605764.3623985. URL https://doi.org/10.1145/3605764.3623985.

İlker Gül, Rémi Lebret, and Karl Aberer. Stance Detection on Social Media with Fine-Tuned Large Language Models. *arXiv e-prints*, art. arXiv:2404.12171, April 2024. doi: 10.48550/arXiv.2404.12171.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. A survey on stance detection for mis- and disinformation identification. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.94. URL https://aclanthology.org/2022.findings-naacl.94.

Alex Holub, Pietro Perona, and Michael C Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv e-prints*, art. arXiv:2106.09685, June 2021. doi: 10.48550/arXiv.2106.09685.

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. Mistral 7B. *arXiv e-prints*, art. arXiv:2310.06825, October 2023. doi: 10.48550/arXiv.2310.06825.

Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Comput. Surv.*, 53(1), feb 2020. ISSN 0360-0300. doi: 10.1145/3369026. URL https://doi.org/10.1145/3369026.

Kostiantyn Kucher, Carita Paradis, Magnus Sahlgren, and Andreas Kerren. Active learning and visual analytics for stance classification with alva. *ACM Trans. Interact. Intell. Syst.*, 7(3), oct 2017. ISSN 2160-6455. doi: 10.1145/3132169. URL https://doi.org/10.1145/3132169.

Chang Li and Dan Goldwasser. Encoding social information with graph convolutional networks forPolitical perspective detection in news media. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1247. URL https://aclanthology.org/P19-1247.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=MmBjKmHIND.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nick Beauchamp, and Lu Wang. POLITICS: Pretraining with Same-story Article Comparison for Ideology Prediction and Stance Detection. *arXiv e-prints*, art. arXiv:2205.00619, May 2022. doi: 10.48550/arXiv.2205.00619.

Ghazaleh Mahmoudi, Babak Behkamkia, and Sauleh Eetemadi. Zero-shot stance detection using contextual data generation with LLMs. In *AAAI-2024 Workshop on Public Sector LLMs: Algorithmic and Sociotechnical Design*, 2024. URL https://openreview.net/forum?id=n9yozEWDGO.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. Active Learning by Acquiring Contrastive Examples. *arXiv e-prints*, art. arXiv:2109.03764, September 2021. doi: 10.48550/arXiv.2109.03764.

Houman Mehrafarin, Sara Rajaee, and Mohammad Taher Pilehvar. On the importance of data size in probing fine-tuned models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 228–238, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.20. URL https://aclanthology.org/2022.findings-acl.20.

Blake Miller, Fridolin Linder, and Walter R. Mebane. Active learning approaches for labeling text: Review and assessment of the performance of active learning approaches. *Political Analysis*, 28 (4):532–551, 2020. doi: 10.1017/pan.2020.4.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks. *arXiv e-prints*, art. arXiv:2304.13861, April 2023. doi: 10.48550/arXiv.2304.13861.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv e-prints*, art. arXiv:2207.04672, July 2022. doi: 10.48550/arXiv.2207.04672.

Julia Romberg and Tobias Escher. Automated topic categorisation of citizens' contributions: Reducing manual labelling efforts through active learning. In *Electronic Government: 21st IFIP WG 8.5 International Conference, EGOV 2022, Linköping, Sweden, September 6–8, 2022, Proceedings*, page 369–385, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-15085-2. doi: 10.1007/978-3-031-15086-9_24. URL https://doi.org/10.1007/978-3-031-15086-9_24.

Julia Romberg and Tobias Escher. Making sense of citizens' input through artificial intelligence: A review of methods for computational text analysis to support the evaluation of contributions in public participation. *Digit. Gov.: Res. Pract.*, jun 2023. doi: 10.1145/3603254. URL https://doi.org/10.1145/3603254. Just Accepted.

Klaus Schmidt, Andreas Niekler, Cathleen Kantner, and Manuel Burghardt. Classifying speech acts in political communication: A transformer-based approach with weak supervision and active learning. In *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pages 739–748, 2023. doi: 10.15439/2023F3485.

H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 287–294, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. doi: 10.1145/130385.130417. URL https://doi.org/10.1145/130385.130417.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. doi: https://doi.org/10.48550/arXiv.2302.13971.

Jannis Vamvas and Rico Sennrich. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland, jun 2020. URL http://ceur-ws.org/Vol-2624/paper9.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. Generating faithful synthetic data with large language models: A case study in computational social science. *ArXiv*, abs/2305.15041, 2023. URL https://api.semanticscholar.org/CorpusID:258866005.

Wenqian Zhang, Shangbin Feng, Zilong Chen, Zhenyu Lei, Jundong Li, and Minnan Luo. KCD: Knowledge Walks and Textual Cues Enhanced Political Perspective Detection in News Media. *arXiv e-prints*, art. arXiv:2204.04046, April 2022. doi: 10.48550/arXiv.2204.04046.

Marc Ziegele, Timo Breiner, and Oliver Quiring. What Creates Interactivity in Online News Discussions? An Exploratory Analysis of Discussion Factors in User Comments on News Items. *Journal of Communication*, 64(6):1111–1138, 10 2014. ISSN 0021-9916. doi: 10.1111/jcom.12123. URL https://doi.org/10.1111/jcom.12123.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can Large Language Models Transform Computational Social Science? *arXiv e-prints*, art. arXiv:2305.03514, April 2023. doi: 10.48550/arXiv.2305.03514.

A Broader Impact

Our work has the potential to improve the efficiency and effectiveness of stance detection in online political discussions, especially in scenarios where online deployment is required. This presents different challenges and risks compared to performing stance detection after a discussion has ended. This can be beneficial for social media platforms, news outlets, and political campaigns, where the detection of stance during a discussion can help improve the discussion quality. Our methods attempts to increase transparency and security, since we can analyse the generated synthetic data before using it to enhance the online model. Similarly, the synthetic data also allows for an analysis regarding the nature of the real data. Stance detection is prone to also be used to for bad purposes, e.g., to manipulate the public opinion. Our work attempts to provide a middle ground between improving the stance detection performance and understanding the possible generated biases of introducing synthetic data from LLMs, whereas the behaviour of an LLM is much less predictable and open to misuse.

B Background

Stance Detection for Online Political Discussions. Stance detection, a sub-task of sentiment analysis [Romberg and Escher, 2023] and opinion mining [ALDayel and Magdy, 2021], aims to automatically identify an author's stance (*favor*, *against*, or *neutral*) towards a discussed issue or target. In online political discussions, this involves determining if the contribution in question is *for* or *against* a topic like tax increases. Stance detection has been identified as an important task for improving discussion summarization [Chowanda et al., 2017], detecting misinformation [Hardalov et al., 2022], and evaluating opinion distributions in online political discussion and participation processes [Romberg and Escher, 2023]. Stance detection is also used in recommender systems and discussion platforms [Küçük and Can, 2020]. Still, due to its dependency on context, stance detection is a highly challenging task. Identifying stance requires understanding both the question and the contributor's position, complicated by users often deviating from the original question and discussing multiple topics in the same thread [Ziegele et al., 2014], leading to little usable training data. Some works in stance detection use graph convolutional networks to learn more out of the present data [Zhang et al., 2022, Li and Goldwasser, 2019]. Recently, fine-tuning transformer-based models [Vaswani et al., 2017, Liu et al., 2022] to solve stance detection is a common practice, but training these models requires a large amount of annotated data, which for the large variety of questions in online political discussions is unfeasible to acquire. We therefore in our work, show how to improve stance detection for online political discussions with synthetic data.

Active Learning. The aim of *active learning* is to minimize the effort of labelling data, while simultaneously maximizing the model's performance. This is achieved by selecting a *query strategy* that chooses the most interesting samples from a set of unlabelled data points, which we refer to as *most informative* samples. These samples are then passed to, e.g., a human annotator for labelling. There exist many different query strategies such as Query By Comittee (QBC, [Seung et al., 1992]), Minimum Expected Entropy (MEE, Holub et al. [2008] or Contrastive Active Learning (CAL, Margatina et al. [2021]). By actively choosing samples and asking for the correct labelling, the model is able to learn from few labelled data points, which is advantageous especially when annotated datasets are not available. Within the domain of political text analysis, many different tasks lack large amounts of annotated data. It has been already shown in the past that these tasks can benefit from the active learning: e.g., stance detection [Kucher et al., 2017], topic modeling [Romberg and Escher, 2022], speech act classification [Schmidt et al., 2023] or toxic comment classification [Miller et al., 2020]. In this work, we examine how LLM-generated synthetic data can be used instead of real labelled data to select the most informative samples to be manually labelled.

Using LLM-generated synthetic data for fine tuning. Recent work has shown that synthetic data generated from LLMs can be used to improve the performance of a model on downstream tasks. Møller et al. [2023] showed that synthetic data can be used to improve the performance of a model on downstream classification tasks by comparing the performance of a model finetuned on LLM-generated data to crowd annotated data. In many cases the model finetuned on LLM-generated data outperforms the model finetuned on crowd annotated data. Mahmoudi et al. [2024] study the use of synthetic data for data augmentation in stance-detection. The authors use GPT-3 to generate synthetic data for a specific topic with mixed results due to the inability of GPT-3 to generate good

data samples. In our work,we use a newer LLM model, Mistral-7B, which generates better synthetic data samples and show that we can generate synthetic data that matches the real data distribution. Veselovsky et al. [2023] analyse in which ways synthetic data is best generated for tasks like sarcasm detection and sentiment analysis. The authors reach the conclusion that grounding the prompts to generate the synthetic data to real samples helps improve the quality of the synthetic data. Similarly, Li et al. [2023] argue that subjectivity in the classification task determines whether synthetic data can be used effectively.

Using LLMs directly for stance detection. It has been shown that LLMs can be used directly for stance detection such as [Cruickshank and Xian Ng, 2023], [Burnham, 2023] [Ziems et al., 2023]. However, the general conclusion of these studies is that while LLMs are competitive with other transformer models such as BERT, especially for edge cases, they exhibit replication issues. [Burnham, 2023] also discuss the posibility of pre-training models on more specific data to improve the generalisation capability of the model. Ziems et al. [2023] highlight the potential biases that can emerged in open ended generation tasks and classification performance varies depending on how representative the training data is. We therefore focus on using LLMs to generate synthetic data to solve key challenges in stance detection such as the lack of available data for specific topics and labelling large amounts of data, rather than using LLMs directly for the task.

C Using LLMs directly for X-Stance

Recently LLMs have been tested on stance detection datasets due to their superior text analysis capabilities. Cruickshank and Xian Ng [2023] and Gül et al. [2024] have shown promising results using zero-shot stance detection and fine-tuning various LLMs on common stance detection datasets such as SemEval-2016 and P-Stance. The X-Stance dataset differs from the SemEval-2016 and P-Stance datasets in the amount of different topics and questions which are related to swiss policy making. The questions are specific and usually represent a very niche issue(see Appendix I). In contrast SemEval-2016 and P-Stance mostly focus on general tweets regarding US politicians. We adopt the prompt and fine-tuning scenario (fine-tuning over 4 epochs with LoRA [Hu et al., 2021]) as in Gül et al. [2024] and use our Mistral-7B model for both zero-shot stance detection and fine-tuned stance detection.

F1 Score (average over 10 Questions)	
Fine-tuned LLM	0.182
Zero-shot LLM	0.419
Baseline	0.693
Baseline+Synth (M=1000)	0.723
SQBC+Synth (M=1000)	0.754

Table 1: **LLM-based stance detection vs BERT-based stance detection:** We compare the Mistral-7B performance to the our BERT stance detection models. We see that zero-shot stance detection barely reaches the pre-trained baselines' performance. Fine-tuning the LLM also proved difficult where even after 10 fine-tuning epochs the performance of the fine-tuned model worsened, most likely needing more resource intensive fine-tuning. Our findings for X-Stance suggest that while LLMs are good at producing open-ended text, they struggle when being prompted to give a specific stance, often refusing to answer the question outright or not understanding the context of the comment and question.

Table 1 shows that zero-shot stance detection barely reaches the performance of the pre-trained BERT baseline. While this is not surprising since the BERT model has been pre-trained on stance detection it also shows that on complex datasets such as X-Stance zero-shot detection proves difficult with smaller models such as Mistral-7B. Suprisingly, fine-tuning the Mistral-7B model with $\mathcal{D}_{\text{train}}$ worsened performance even further. We tried various hyperparameters and even trained for up to 10 epochs, more than the 4 used in Gül et al. [2024].

We attribute the poor performance to the varied nature of the X-Stance dataset. While further fine-tuning may likely improve performance, training for 10 epochs on an Nvidia A100 already took 12 hours. The biggest issue while using zero-shot stance detection with the Mistral-7B model was that it

would not give a consistent output or would often refuse to predict stance. Our findings suggest that using the LLM for open-ended text generation proves more effective rather than forcing it to give a specific output.

D Synthetic Data Alignment

We perform an ablation on fine-tuning with synthetic data in order to determine whether the improved performance comes from the increased dataset size or due to the content of the generated comments. For this, we perform 3 different fine-tuning runs where the comments are shuffled, that is the posed questions and the synthetic data are misaligned. We show that fine-tuning is only effective when the synthetic data is aligned with the posed questions.

F1 Score (average over 10 Questions)			
	M=200	M=500	M=1000
Baseline	0.693	0.693	0.693
Baseline+Synth	0.711	0.717	0.723
Baseline+Synth (Misaligned)	0.699	0.704	0.694

Table 2: **Topic alignment is crucial for the synthetic data to be effective:** To determine whether improvement with synthetic data is due to the dataset size or the synthetic data itself, we augment the stance detection model with misaligned synthetic data. That is, the synthetic data does not align with the question given to the stance detection moel. We show that while increasing synthetic data set size does improve performance, it is also important that the synthetic data aligns with the posed question.

E Visualizations

E.1 Visualizing the synthetic data

We visualize the synthetic data together with the real world data for $M = 1000$ and $M = 200$ in Figures 4 and 5. We plot the data points of the synthetic data in blue and red for the positive and negative samples, respectively. The means of the real world data are plotted as a regular polygon with 8 sides. We observe that the synthetic data extends the real world data, which we consider a factor as to why fine-tuning with synthetic data is effective in online political discussions. Also, the larger the synthetic dataset size, the more the synthetic data matches the distribution of the real world data since for $M = 200$ (see Figure 5) the mean are not as well aligned with the synthetic data. Furthermore, the positive and negative samples are well separated, which we attribute to having pre-trained the BERT-model on $\mathcal{D}_{\text{train}}$ of the X-Stance dataset, giving the prior knowledge about the stance detection task.

E.2 Visualizing the query strategies of the active learning methods

We visualize the selected samples of **SQBC**, **CAL** and Random query strategies for $M = 1000$ and $M = 200$ in Figures 6 and 7. We plot the selected samples of the unlabelled data in green. The positive and negative synthetic data samples are plotted in blue and red, respectively. The selected samples are highlighted in orange. We observe that **SQBC** selects the unlabelled samples that are mostly in between the two classes of the synthetic data. This is the expected behaviour since we select the samples where the classification score is ambiguous. For **Random**, the range of selected samples is broad: some similar samples between the two classes like **SQBC** are selected, but also within class samples that are not covered by the synthetic data set. This explains why random selection works well with a large synthetic dataset, since it further extends the decision boundary of the model. For the smaller synthetic dataset $M = 200$, the random selection is not as effective, since the selected samples are spread out over the whole data space and not necessarily in between the two classes as with the larger synthetic dataset. Finally, **CAL** selects samples similar to **SQBC**, but mostly tends to select samples from only one class.

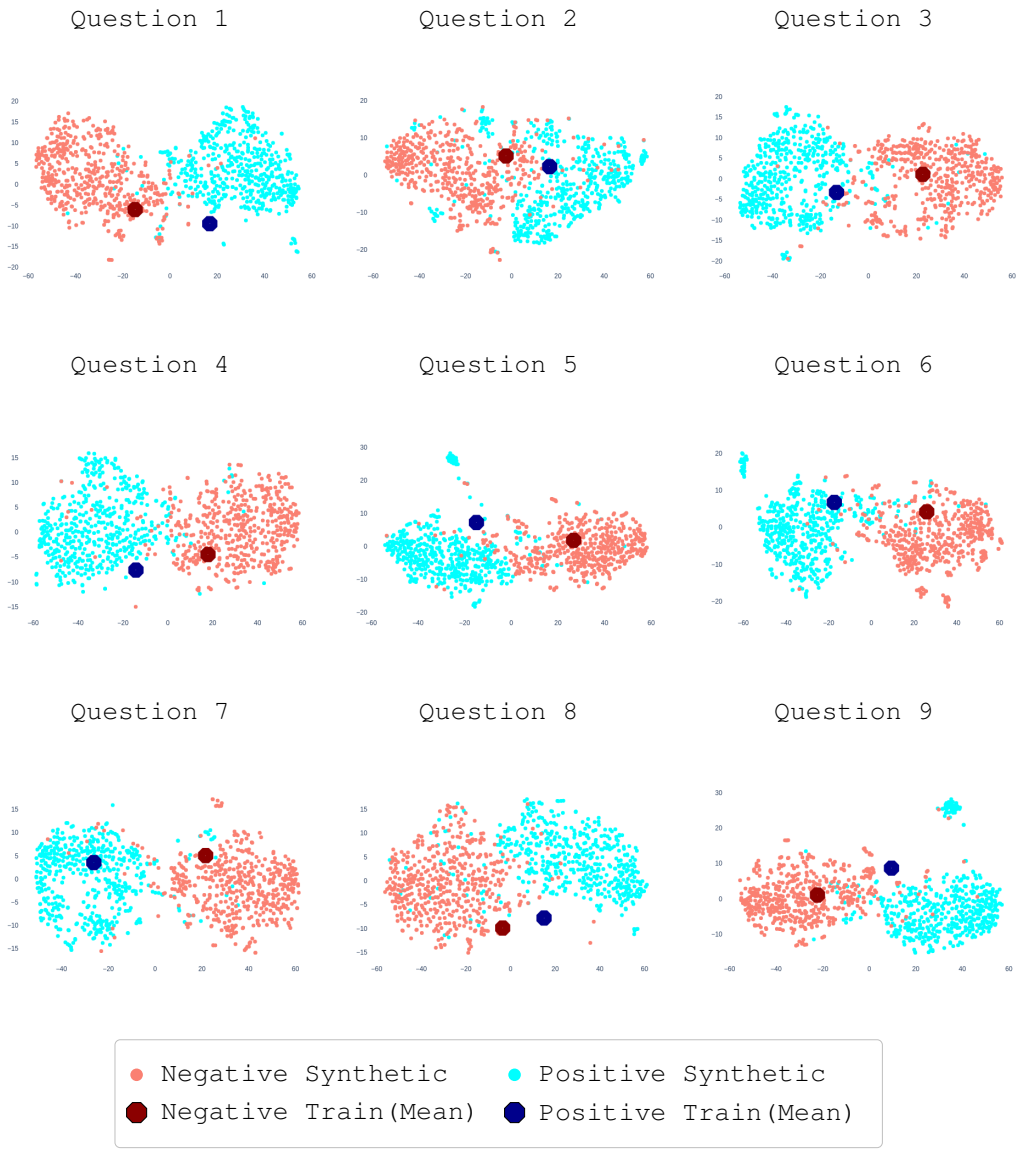

Figure 4: **Visualization of synthetic data with train data means for $M = 1000$ synthetic data.** For a larger synthetic dataset size, the means of the synthetic data are well aligned with the real world data and the positive and negative samples are well separated. The synthetic data thus extends the real world data, which we consider a factor as to why fine-tuning with synthetic data is effective in online political discussions.

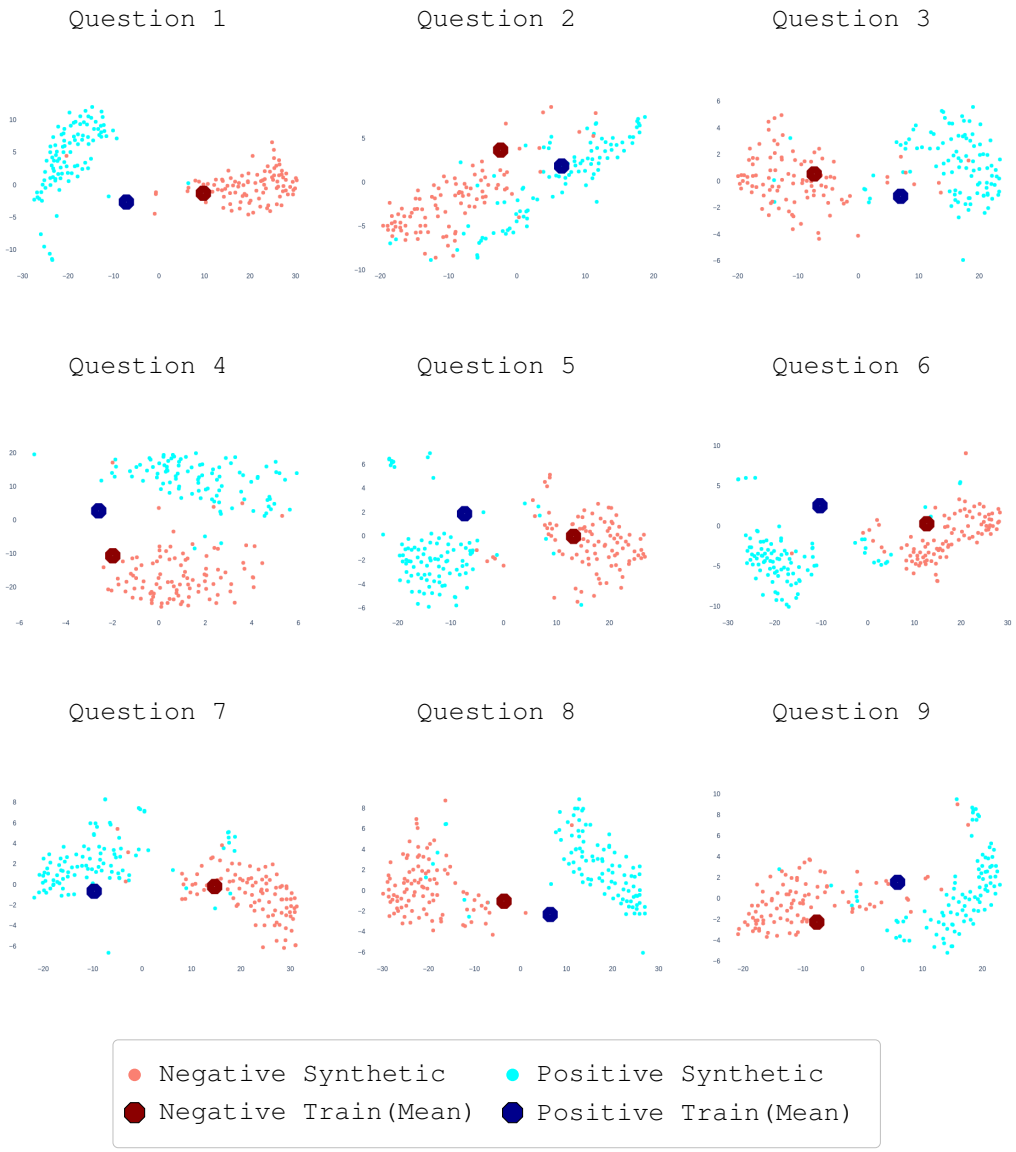

Figure 5: **Visualization of synthetic data with train data means for $M = 200$ synthetic data:** For a smaller synthetic dataset size, the means of the synthetic data are not as well aligned with the real world data as for $M = 1000$. However, the positive and negative samples are still well separated.

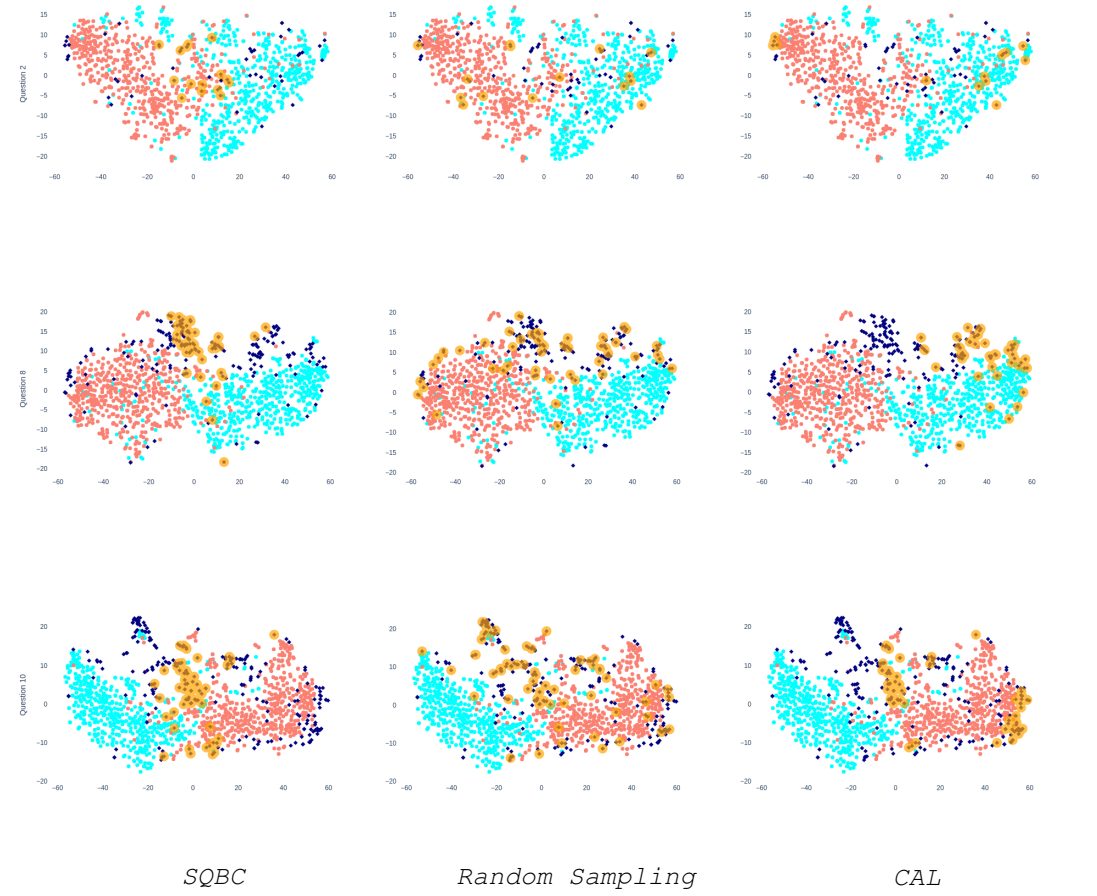

Figure 6: **Visualization of SQBC, Random and CAL query strategies for $M = 1000$ synthetic data: SQBC** selects the unlabelled samples that are mostly in between the two classes of the synthetic data. This is the expected behaviour since we select the samples where the classification score is ambiguous. For random selection, the range of selected samples is broad: some similar samples between the two classes like **SQBC** are selected, but also within class samples that are not covered by the synthetic data set. This explains why random selection works well with a large synthetic dataset, since it further extends the decision boundary of the model. Finally, **CAL** selects samples similar to **SQBC**, but mostly tends to select samples from only one class, resulting in worse performance.

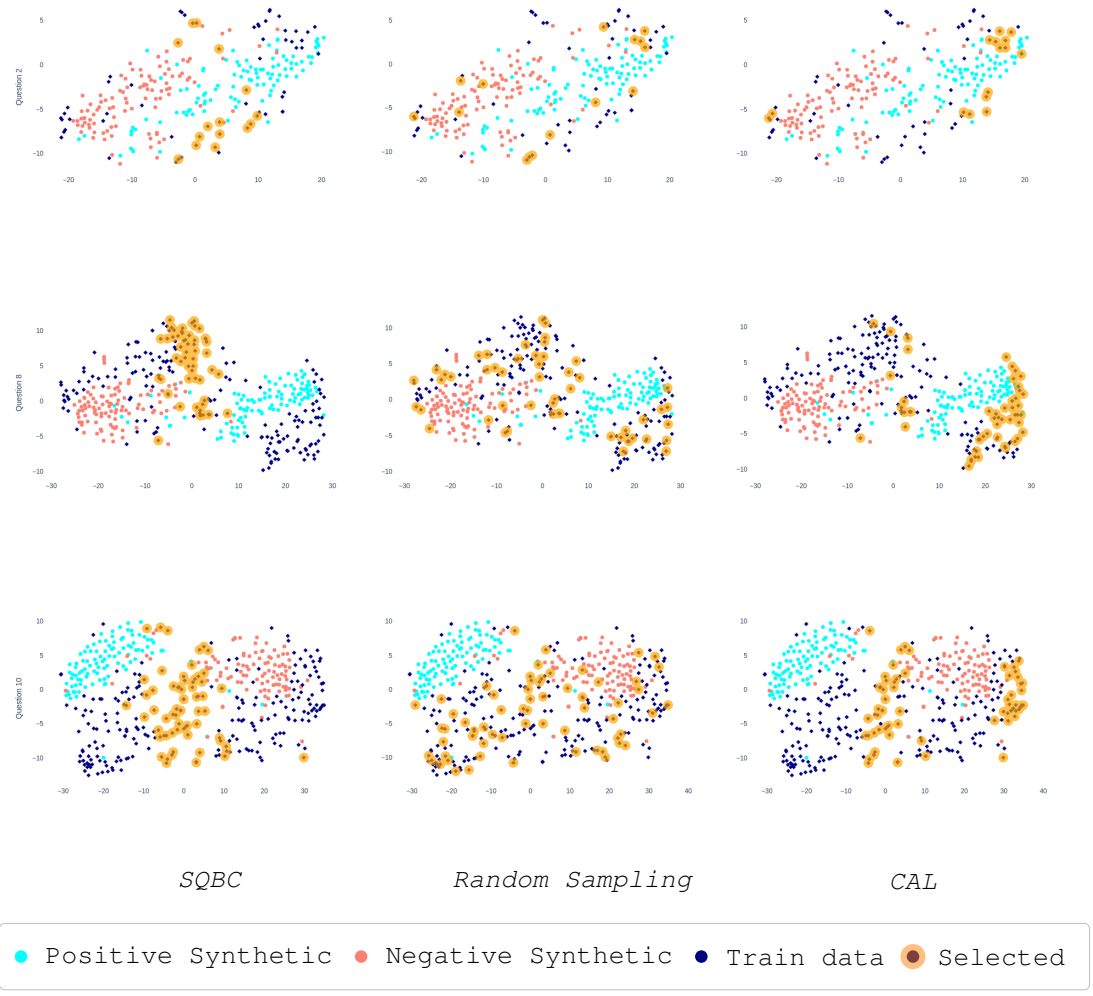

Figure 7: **Visualization of SQBC, Random and CAL query strategies for $M = 200$ synthetic data:** For a smaller synthetic dataset size, **SQBC** is still able to select the unlabelled samples that are mostly in between the two classes of the synthetic data. For **Random** we see that the selected samples are a bit further away from the synthetic data distribution, which is why we argue it does not perform as well as with the larger synthetic dataset.

F Detailed Results

Fine-tuning with synthetic data				
	M=0	M=200	M=500	M=1000
Baseline + Synth	0.693	0.712	0.718	0.723
True Labels + Synth	0.727	0.745	0.746	0.770

Table 3: Results of using synthetic data for fine-tuning.

Fine-tuning with most informative samples selected with synthetic data												
	M=200				M=500				M=1000			
	10%	25%	50%	75%	10%	25%	50%	75%	10%	25%	50%	75%
CAL	0.693	0.697	0.705	0.714	0.692	0.694	0.707	0.718	0.692	0.696	0.708	0.720
Random	0.693	0.696	0.705	0.719	0.693	0.695	0.706	0.715	0.692	0.695	0.706	0.715
SQBC	0.693	0.697	0.709	0.722	0.692	0.700	0.711	0.722	0.692	0.698	0.712	0.721

Table 4: Results of only training with most informative samples.

Fine-tuning with most informative samples and synthetic data												
	M=200				M=500				M=1000			
	10%	25%	50%	75%	10%	25%	50%	75%	10%	25%	50%	75%
CAL+Synth	0.713	0.715	0.727	0.732	0.711	0.721	0.732	0.748	0.695	0.715	0.747	0.749
Random+Synth	0.716	0.720	0.724	0.734	0.723	0.735	0.730	0.748	0.724	0.746	0.754	0.756
SQBC+Synth	0.715	0.723	0.731	0.735	0.714	0.726	0.744	0.750	0.721	0.737	0.753	0.747

Table 5: Tabular version of Figure 2

G Additional Experimental Details

G.1 Evaluation.

For fine-tuning and testing we evaluate the given model separately on 10 chosen questions from the test dataset of X-Stance for all experiments. For each question q we split $\mathcal{D}_{\text{test}}^{(q)}$ into a $60/40$ train/test split (repeated with 5 different seeds to get error bars) and use the train split for fine-tuning to the given question and the test split for evaluation. Our main results report the average F1 score over 10 selected questions from the test dataset evaluated on the comments from the test split. The error bars represent the average standard deviation over the 10 questions for 5 runs with different seeds. More detailed results per question are shown in Appendix J.

G.2 Compute and Runtime

We conduct our experiments on a single NVIDIA A100 80GB GPU and a 32 core CPU. With this setup, for Mistral-7B, the generation of synthetic data takes approximately 3 hours per question for a synthetic dataset size of $M = 1000$. Fine-tuning the BERT model with the synthetic data takes less than a minute. For the active learning methods, the selection of the most informative samples takes less than a minute. Hence the largest computational effort is the generation of the synthetic data.

G.3 Translation of the X-Stance dataset for synthetic data generation

In Figure 8 we show the pipeline for translating the X-Stance dataset for synthetic data generation. We start with a question q from the X-Stance test dataset and translate the question to English with a NLLB-330M model [NLLB Team et al., 2022]. Then we let the Mistral-7B model generate synthetic data, i.e., comments for the translated question. The generated comments are then translated back to German to be used for fine-tuning the model in our experiments.

G.4 Overview of used datasets

In Table 6 we show an overview of the datasets used in our experiments for the different methods we evaluate.

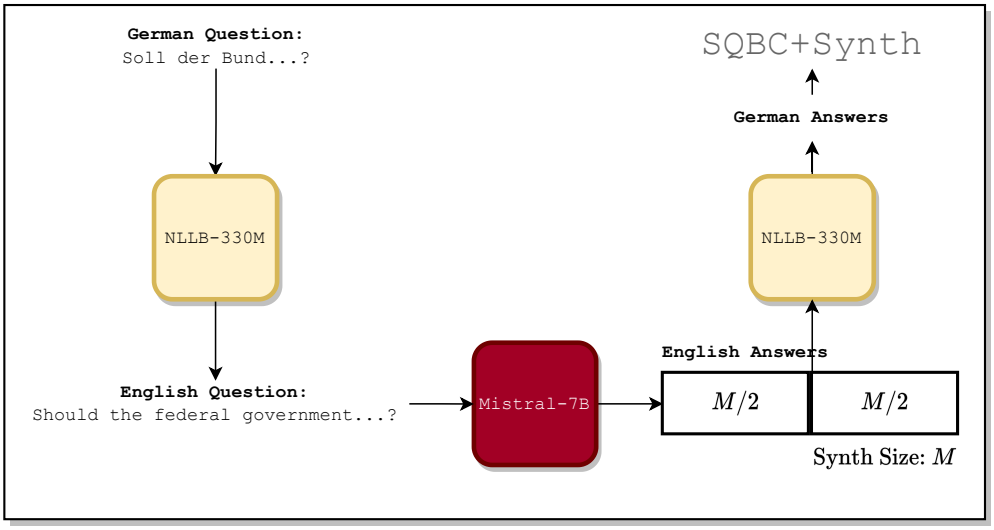

Figure 8: **Overview of the pipeline for active learning with synthetic data:** We start with a question q from the X-Stance test dataset and translate the question to English with a NLLB-330M model [NLLB Team et al., 2022]. Then we let the Mistral-7B model generate synthetic data, i.e., comments for the translated question. The generated comments are then translated back to German to be used for fine-tuning the model in our experiments.

Config \ Datasets	Manual labels D_{MInf}	True Labels $\mathcal{D}_{\text{t}}$	Synth Aug D_{synth}
Baseline			
Baseline + Synth			✓
True Labels		✓	
True Labels + Synth		✓	✓
SQBC	✓		
SQBC + Synth	✓		✓
CAL	✓		
CAL + Synth	✓		✓
Random	✓(randomly selected)		
Random + Synth	✓(randomly selected)		✓

Table 6: Synth: Synthetic Data, Aug: Augmentation. We compare different variants of active learning with synthetic data.

H Dataset

X-Stance is a multilingual stance detection dataset, including comments in German $(48,600)$, French $(17,200)$ and Italian $(1,400)$ on political questions, answered by election candidates from Switzerland. The data has been extracted from smartvote[1], a Swiss voting advice platform. For the task of cross-topic stance detection the data is split into a training set, including questions on 10 political topics, and a test set with questions on two topics that have been held out, namely *healthcare* and *political system*.

H.1 Chosen questions and their distribution

We present the 10 chosen questions for our experiments in Table 7. We show the original questions in German and their corresponding English translations by the translation model. Furthermore, we also

[1] `https://www.smartvote.ch/`

show the (60 / 40) train/test split for each question in Figure 9. We chose 10 questions that reflect the overall distribution of $\mathcal{D}_{\text{test}}^{(q)}$. We choose questions with small amount of comments, unbalanced comments and also balanced comments. Furthermore, for 5 of the questions the majority class is *favor* and for the other 5 the majority class is *against*.

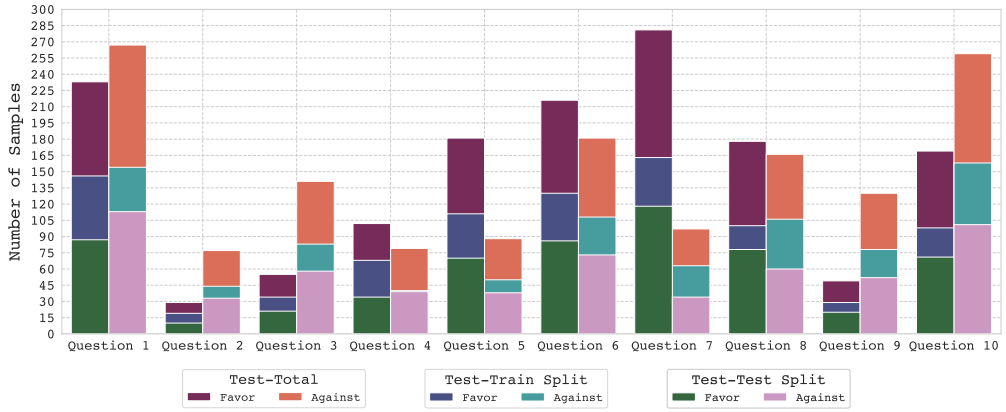

Figure 9: **Distribution of the positive and negative samples for the train and test split of $\mathcal{D}_{\text{test}}^{(q)}$:** We show the distribution of the positive and negative samples of the X-Stance test dataset for the questions Q1-Q10. We also show the 60/40 train/test split for the 10 questions. We chose 10 questions that reflect the overall distribution of $\mathcal{D}_{\text{test}}^{(q)}$. We chose unbalanced, balanced and low sample size questions to evaluate the effectiveness of our approach.

I Synthetic Data

We show the translated questions used for synthetic data generation in Table 7 and some samples of generated comments in 8. We see that the questions are translated correctly and synthetic data can be generated for both favor and against stances.

Question in German	**Question in English**
Sollen sich die Versicherten stärker an den Gesundheitskosten beteiligen (z.B. Erhöhung der Mindestfranchise)	Should insured persons contribute more to health costs (e.g. increase in the minimum deductible)?
Befürworten Sie ein generelles Werbeverbot für Alkohol und Tabak?	Do you support a general ban on advertising alcohol and tobacco?
Soll eine Impfpflicht für Kinder gemäss dem schweizerischen Impfplan eingeführt werden?	Should compulsory vaccination of children be introduced in accordance with the Swiss vaccination schedule?
Soll die Aufenthaltserlaubnis für Migrant/innen aus Nicht-EU/EFTA-Staaten schweizweit an die Erfüllung verbindlicher Integrationsvereinbarungen geknüpft werden?	Should the residence permit for migrants from non-EU/EFTA countries be linked to the fulfilment of binding integration agreements throughout Switzerland?
Soll der Bund erneuerbare Energien stärker fördern?	Should the federal government promote renewable energy more?
Befürworten Sie eine strengere Kontrolle der Lohngleichheit von Frauen und Männern?	Do you support stricter control of gender pay?
Soll der Bund die finanzielle Unterstützung für die berufliche Weiterbildung und Umschulung ausbauen?	Should the Federal Government expand its financial support for continuing vocational training and retraining?
Sollen in der Schweiz vermehrt Spitäler geschlossen werden, um die Kosten im Gesundheitsbereich zu senken?	Should more hospitals be closed in Switzerland to reduce healthcare costs?
Eine eidenössische Volksinitiative verlangt, dass der Bundesrat direkt vom Volk gewählt werden soll. Unterstützen Sie dieses Anliegen?	A sworn federal people's initiative demands that the Federal Council be elected directly by the people. Do you support this concern?
Soll die Einführung der elektronischen Stimmabgabe bei Wahlen und Abstimmungen (E-Voting) weiter vorangetrieben werden?	Should the introduction of electronic voting in elections and voting (e-voting) be further promoted?

Table 7: Chosen questions for stance detection in German and their English translation

Comment in English	Intended Stance
I don't believe that insured persons should contribute more to their own healthcare costs by increasing the minimum deductible. While it may seem like a good idea to increase deductibles as a way to reduce overall healthcare costs, this only shifts the burden onto patients and puts them at risk of facing high out-of-pocket expenses. It's important for insurance companies to find ways to control costs without placing the onus on insured individuals.	AGAINST
Hi, friends! I think it's a great idea for insured person to contribute more to their health costs by increasing the minimum deductible. As someone who believes in taking ownership of my own health and finances, I believe that paying more out-of-pocket can help me better manage my expenses and make more informed decisions about my healthcare. Plus, I think that by contributing more, I can save money in the long run and avoid potential financial strain in other areas of my life.	FAVOR
"In my opinion, it's wrong to link the residence permit for migrants from non-EU/EFTA countries to the fulfillment of binding integration agreements in Switzerland. This policy puts too much emphasis on the ability of migrants to adapt to Swiss culture and language at the expense of their personal freedoms and rights as human beings. It also creates incentives for migrants to cheat or lie in order to gain a residency permit and access to valuable services and resources."	AGAINST
I believe that the residence permit for migrants from non-EU/EFTA countries should be linked to the fulfillment of binding integration agreements throughout Switzerland. This is because integration should not be seen as a one-sided process, but rather as a mutually beneficial relationship between the immigrant and the host country. By linking the residence permit to the successful completion of integration agreements, we incentivize immigrants to actively participate in the integration process and become active members of Swiss society. Additionally, this approach ensures that immigrants do not receive a free ride by simply obtaining a residence permit without contributing to the well-being of Swiss society. Ultimately, a comprehensive integration approach benefits both the immigrant and the host country by creating a more inclusive and harmonious society.	FAVOR
As someone who is against stricter controls on equal pay for women and men, I believe that all individuals should be paid based on their individual skills, performance, and contributions to the company, regardless of their gender. I think it's important to recognize and reward hard work and dedication, regardless of whether someone is male or female. Additionally, I argue that it's already illegal and unethical to discriminate on the basis of gender and that such regulations would only create unnecessary tensions and potential conflicts within the workplace. Instead, we should focus on creating a truly equitable society where everyone has access to equal opportunities for education, job training, and development.	AGAINST
I personally believe that stricter controls on equal pay for women and men are necessary because it's not always fair that men receive higher salaries just because of their gender. In many cases, women have to work harder than men to achieve the same level of performance, but they still earn less money. Stricter controls would help ensure that everyone's hard work is recognized and rewarded fairly, regardless of their gender. Additionally, it's important to promote gender equality and give all individuals an equal opportunity to succeed. Overall, I think stricter controls on equal pay are necessary to help reduce the gender pay gap and create a more equitable society.	FAVOR

Table 8: Sample of comments generated by the LLM used for fine-tuning the stance detection model.

Comment in English	Comment in German
I don't believe that insured persons should contribute more to their own healthcare costs by increasing the minimum deductible. While it may seem like a good idea to increase deductibles as a way to reduce overall healthcare costs, this only shifts the burden onto patients and puts them at risk of facing high out-of-pocket expenses. It's important for insurance companies to find ways to control costs without placing the onus on insured individuals.	Ich glaube nicht, dass Versicherte mehr zu ihren eigenen Gesundheitskosten beitragen sollten, indem sie den Mindestabzug erhöhen. Die Erhöhung der Selbstbehalte als Mittel zur Senkung der gesamten Gesundheitskosten mag zwar als gute Idee erscheinen, verschiebt jedoch die Belastung nur auf die Patienten und gefährdet sie mit hohen Kosten, die sie selbst tragen müssen. Es ist wichtig für Versicherungsgesellschaften, Wege zu finden, Kosten zu kontrollieren, ohne die Versicherten mit der Last zu belasten.
Hi, friends! I think it's a great idea for insured person to contribute more to their health costs by increasing the minimum deductible. As someone who believes in taking ownership of my own health and finances, I believe that paying more out-of-pocket can help me better manage my expenses and make more informed decisions about my healthcare. Plus, I think that by contributing more, I can save money in the long run and avoid potential financial strain in other areas of my life.	Ich finde es eine gute Idee, dass Versicherte mehr zu ihren Gesundheitskosten beitragen, indem sie den Mindestabzug erhöhen. Als jemand, der daran glaubt, die Verantwortung für seine Gesundheit und seine Finanzen zu übernehmen, glaube ich, dass mehr Geld aus eigener Tasche helfen kann, meine Ausgaben besser zu managen und fundierte Entscheidungen über meine Gesundheitsversorgung zu treffen. Außerdem denke ich, dass ich durch einen größeren Beitrag auf lange Sicht Geld sparen und finanzielle Belastungen in anderen Bereichen meines Lebens vermeiden kann.
"In my opinion, it's wrong to link the residence permit for migrants from non-EU/EFTA countries to the fulfillment of binding integration agreements in Switzerland. This policy puts too much emphasis on the ability of migrants to adapt to Swiss culture and language at the expense of their personal freedoms and rights as human beings. It also creates incentives for migrants to cheat or lie in order to gain a residency permit and access to valuable services and resources."	"Meiner Meinung nach ist es falsch, die Aufenthaltserlaubnis für Migranten aus Nicht-EU/EFTA-Ländern mit der Erfüllung verbindlicher Integrationsvereinbarungen in der Schweiz zu verbinden. Diese Politik legt zu viel Wert auf die Fähigkeit der Migranten, sich an die Schweizer Kultur und Sprache anzupassen, auf Kosten ihrer persönlichen Freiheiten und Rechte als Menschen. Es schafft auch Anreize für Migranten, zu betrügen oder zu lügen, um eine Aufenthaltserlaubnis zu erhalten und Zugang zu wertvollen Dienstleistungen und Ressourcen zu erhalten "
I believe that the residence permit for migrants from non-EU/EFTA countries should be linked to the fulfillment of binding integration agreements throughout Switzerland. This is because integration should not be seen as a one-sided process, but rather as a mutually beneficial relationship between the immigrant and the host country. By linking the residence permit to the successful completion of integration agreements, we incentivize immigrants to actively participate in the integration process and become active members of Swiss society. Additionally, this approach ensures that immigrants do not receive a free ride by simply obtaining a residence permit without contributing to the well-being of Swiss society. Ultimately, a comprehensive integration approach benefits both the immigrant and the host country by creating a more inclusive and harmonious society.	Ich glaube, dass die Aufenthaltserlaubnis für Migranten aus Nicht-EU/EFTA-Ländern mit der Erfüllung verbindlicher Integrationsvereinbarungen in der gesamten Schweiz verbunden sein sollte. Die Integration sollte nicht als einseitiger Prozess, sondern als eine gegenseitig vorteilhafte Beziehung zwischen dem Einwanderer und dem Aufnahmeland betrachtet werden. Durch die Verknüpfung der Aufenthaltserlaubnis mit dem erfolgreichen Abschluss von Integrationsvereinbarungen fördern wir die aktive Teilnahme der Einwanderer am Integrationsprozess und die Förderung ihrer Teilnahme an der Schweizer Gesellschaft Darüber hinaus wird durch diese Vorgehensweise sichergestellt, dass Einwanderer nicht einfach eine Aufenthaltserlaubnis erhalten, ohne zum Wohlergehen der Schweizer Gesellschaft beizutragen. Letztlich kommt einem umfassenden Integrationsansatz sowohl der Einwanderer als auch dem Aufnahmeland zugute, da er eine integrativere und harmonischere Gesellschaft schafft

Table 9: Sample of translated comments from comments generated by the LLM used for fine-tuning the stance detection model.

J Extended Results

We present the extended results for the different synthetic dataset sizes $M = 200$, $M = 500$ and $M = 1000$ in Figures 10, 11 and 12. As in Figure 2, we show the results for the different active learning methods and the different configurations of the synthetic data, while varying the amount of samples that need to be labelled. We compare all methods to **True Labels**, hence the horizontal line corresponds to the performance of the baseline model fine-tuned with the true labels.

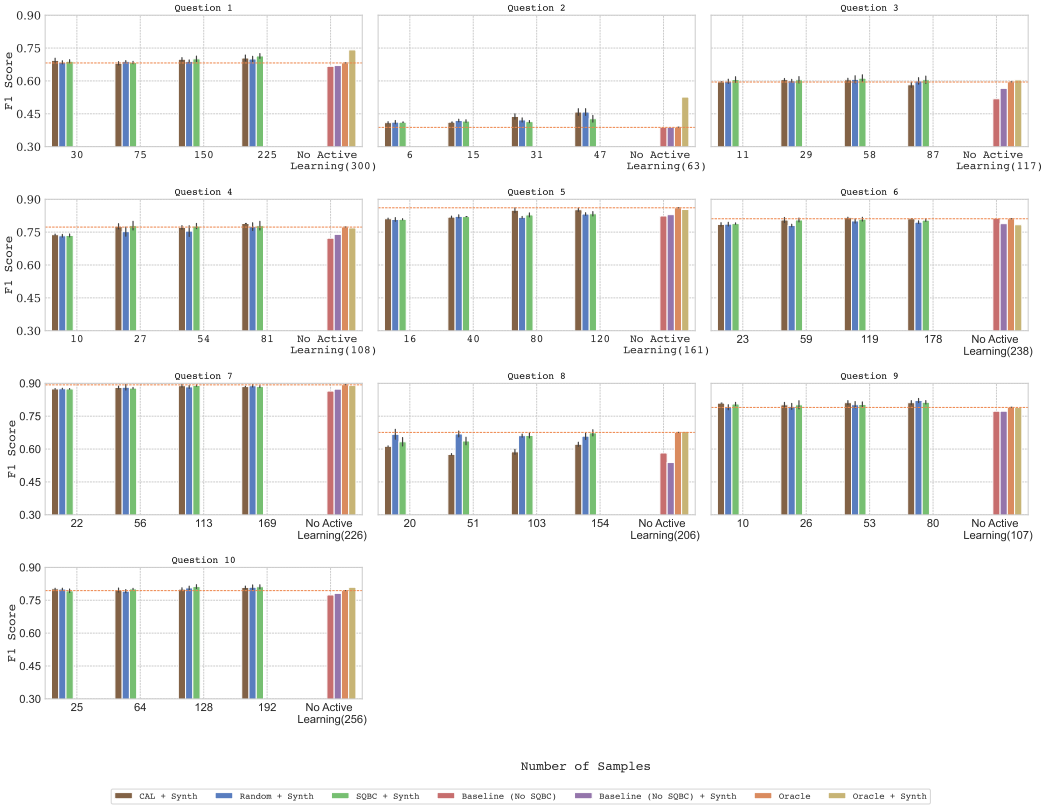

Figure 10: **Extended results of Figure 2 for M=200:**

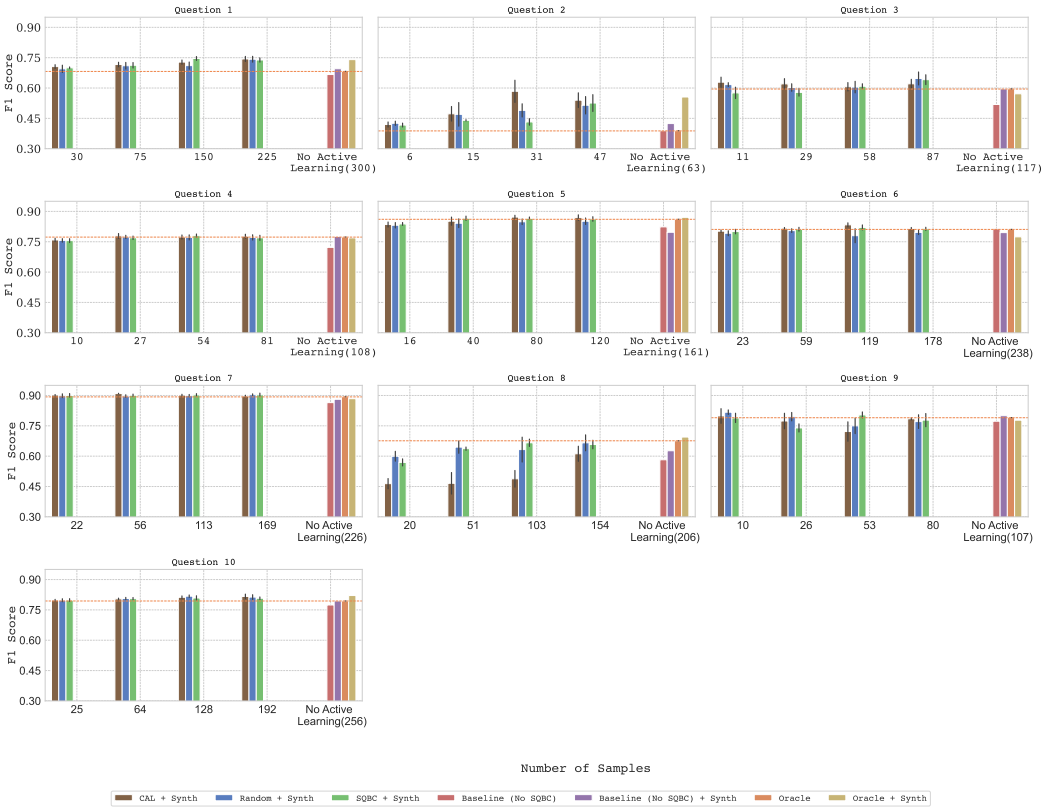

Figure 11: **Extended results of Figure 2 for M=500**

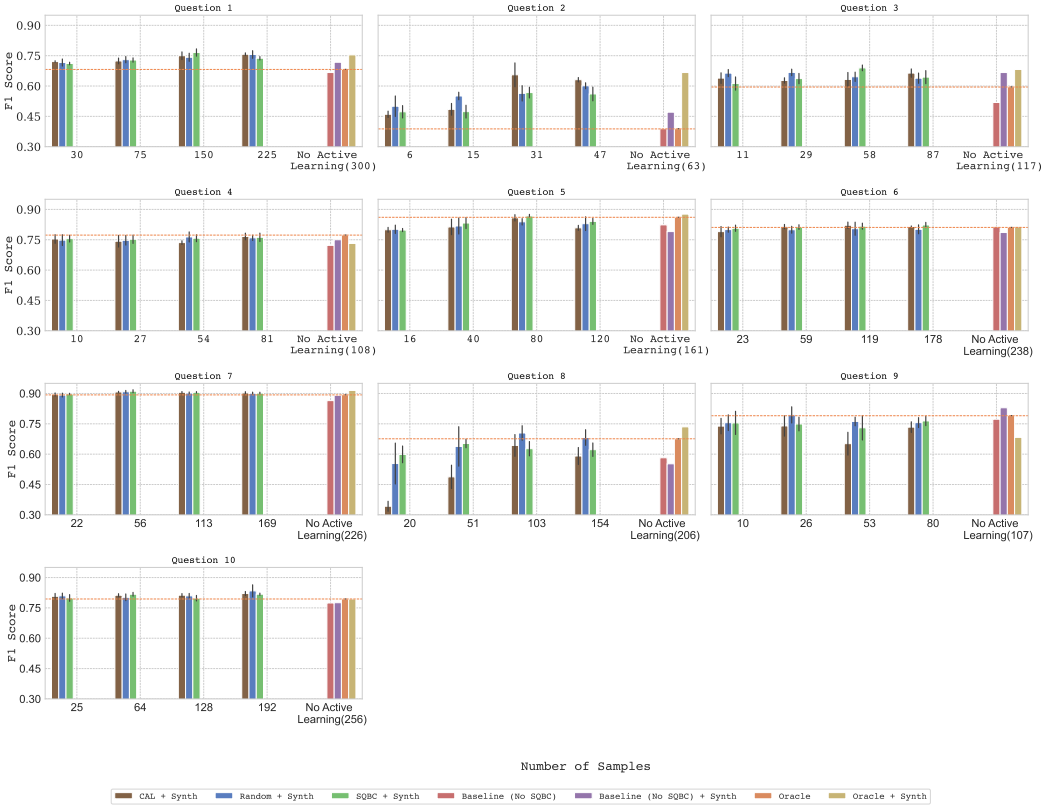

Figure 12: **Extended results of Figure 2 for M=1000**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction reflect the contributions and scope of the paper. We present how to leverage synthetic data for stance detection in online discussions. We are able to show that we can use LLM-generated synthetic data effectively to fine-tune a BERT-model and to also select most informative samples. When combining the two approaches we have a model that on average is better than the model trained with all true labels.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We point out the limitations of our work in the discussion section. Specifically, we discuss that our model has to be fine-tuned for every question specifically, which may be costly. However, given that the distributions of comments between different questions rarely match (which we also see in this paper) this is warranted and requires analysis in future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We have no theoretical results in our paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe our method and experiments in detail in Sections 2 and 3 in the main paper and provide additional details in the supplemental material in Appendix G. We also provide the code and data to reproduce the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the code and data to reproduce the results. We also provide instructions on how to run the code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the training and test details in Section 3.2, while also providing additional details in Section G. Figure 9 shows the train/test split for the data.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the F1-score for the different methods and configurations to account for class imbalance with standard deviation error bars in the figures (standard deviation over 5 seeds).

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the hardware used in the Appendix G. We also state the time needed for the synthetic data generation and the fine-tuning of the models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: We discuss the societal impact of our work in Appendix A.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- samples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We only use packages from the Hugging Face Transformers library and the X-Stance dataset, which are properly credited and the licenses are respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce new assets in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.