Understanding the Influence of Synthetic Data for Text Embedders

Jacob Mitchell Springer[†] **Vaibhav Adlakha**[‡] **Siva Reddy**[‡,§]
Aditi Raghunathan[†] **Marius Mosbach**[‡]
[†] Carnegie Mellon University
[‡] Mila – Quebec AI Institute, McGill University
[§] Canada CIFAR AI Chair
jspringer@cmu.edu marius.mosbach@mila.quebec

Abstract

Recent progress in developing general purpose text embedders has been driven by training on ever-growing corpora of synthetic LLM-generated data. Nonetheless, no publicly available synthetic dataset exists, posing a barrier to studying its role for generalization. To address this issue, we first reproduce and publicly release the synthetic data proposed by Wang et al. (2024) (Mistral-E5). Our synthetic data is high quality and leads to consistent improvements in performance. Next, we critically examine where exactly synthetic data improves model generalization. Our analysis reveals that benefits from synthetic data are sparse and highly localized to individual datasets. Moreover, we observe *trade-offs* between the performance on different categories and data that benefits one task, degrades performance on another. Our findings highlight the limitations of current synthetic data approaches for building general-purpose embedders and challenge the notion that training on synthetic data leads to more robust embedding models across tasks.

jakespringer/open-synthetic-embeddings
🤗 jspringer/open-synthetic-embeddings

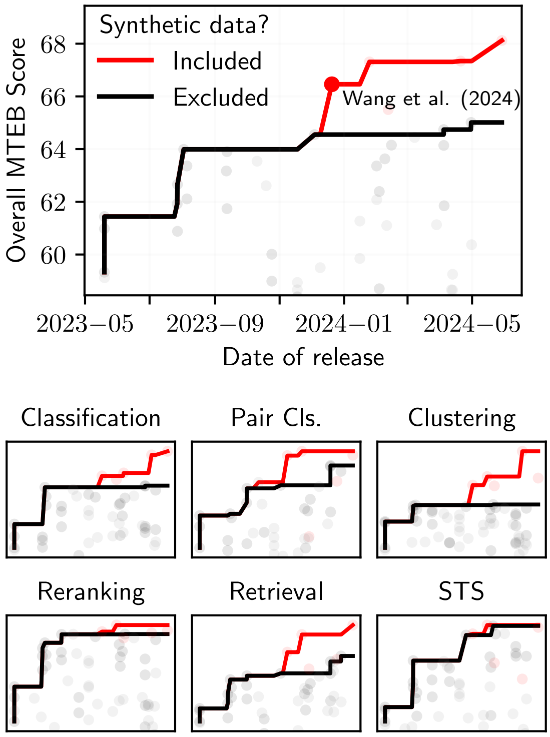


Figure 1: Performance on MTEB across time. Starting with Wang et al. (2024), models trained on synthetic data have led to considerable improvements on the MTEB leaderboard. We exclude more recent models that were trained on in-domain data.

1 Introduction

Mirroring the success of generative LLMs (Ouyang et al., 2022; Jiang et al., 2023; Llama Team, 2024), the NLP community has invested in building general-purpose embedding models—single models capable of producing embeddings for a wide array of embedding tasks, spanning classification, clustering, retrieval, reranking, and text-similarity estimation (Li et al., 2023; Wang et al., 2024; Springer et al., 2024; BehnamGhader et al., 2024; Muennighoff et al., 2024; Lee et al., 2025).

Beyond architectural innovations and progress in base models, much of the recent progress on general-purpose embedders can be attributed to training on synthetic training data—for example, by leveraging GPT-4 to produce synthetic data that expands existing training datasets for embeddings to new tasks (see Figure 1) (Wang et al., 2024). Leveraging synthetic data is based on the premise that LLMs generate more diverse and high-quality data compared to human-annotated datasets, which are often limited in size.

In our work, we critically examine the implicit assumption of this paradigm that training on synthetic data will broadly improve the general-purpose quality of the model. We approach this by training models on different compositions of synthetic data to estimate the influence of differ-

Findings of the Association for Computational Linguistics: ACL 2025, pages 22551–22567
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

ent commonly adopted types of synthetic data on the downstream performance of the model. Strikingly, **we find that training on synthetic examples designed for a particular task can degrade the performance of other tasks, challenging the notion that training on more diverse synthetic data is strictly better.** Moreover, we observe that synthetic data leads to sparse improvement across tasks, showing no statistically significant improvement on a majority of MTEB tasks.

To conduct our analysis, we reproduce the synthetic data from Wang et al. (2024) and release this data publicly. In our reproduction, we compare the effectiveness of LLaMA-3.1-8B and LLaMA-3.1-70B for generating synthetic data, both of which are more cost-efficient than the GPT-family models used by Wang et al. (2024) and are publicly available (Llama Team, 2024). Our results show that synthetic data generated by LLaMA-3.1-8B performs nearly as well as data from the larger 70B model, while costing 5× less than the 70B model and over 50× less than GPT-4o.

In total, our results underscore the need to develop broad and robust methods for generating and training on synthetic data that do not exhibit tradeoffs, and we hope that our public release of synthetic training data will accelerate open research into the development and understanding of general-purpose embedding models.

2 Generating high quality synthetic data

Wang et al. (2024) have demonstrated that the addition of large quantities of LLM generated synthetic data can led to substantial improvements in embedding quality. In fact, synthetic data has been so successful that the current gold standard text embedding benchmark—MTEB (Muennighoff et al., 2023)—is largely dominated by models that train, at least in part, on synthetic data (Lee et al., 2025; Muennighoff et al., 2024; Meng et al., 2024).

However, the synthetic datasets used to train these models are typically generated using proprietary LLMs and remain unavailable to the scientific community, making it difficult to understand their role for the generalization of general-purpose embedding models.

We address this by reproducing and publicly releasing the synthetic data from our replication of Wang et al. (2024). We describe our approach for generating synthetic data as follows.

Dataset generation pipeline. We generate data following the pipeline proposed by Wang et al. which offers an effective method to generate synthetic data that has diverse structure and content. More specifically, we generate data of six different categories based on query and document length: *short-short, long-long, short-long, long-short, bitext, and STS*. Short examples, such as queries in short-long—consist of a few words or a single sentence, while long examples comprise multiple sentences. In short-short, long-long, STS, and bitext pairs, queries and documents are drawn from the same distribution, modeling semantic similarity estimation. In contrast, long-short and short-long pairs involve different distributions, covering tasks such as classification and retrieval. To ensure diversity of content, we generate data in two steps:

1. *Brainstorming:* We generate task descriptions for each of the categories outlined above.

2. *Instance generation:* We generate training examples based on both the output of the brainstorming stage and the associated category. Each training example consists of an instruction, a query, a positive example relevant to the query, and a hard negative example[1] that is only superficially relevant to the query.

Following the pipeline described above, we generate approximately 500k synthetic examples from LLaMA-3.1-8B and LLaMA-3.1-70B, respectively. More details of the generation process along with prompts, examples, and the final composition of synthetic data are in Appendix B.

Training and evaluation setup. We experiment with three different models: Mistral-v0.1-7B, Mistral-v0.2-7B (Jiang et al., 2023), and Qwen2-1.5B (Yang et al., 2024), and compare training without synthetic data to training with synthetic data sourced from either Llama-3.1-70B or Llama-3.1-8B. We follow Wang et al. (2024) and mix our synthetic data with the public E5 dataset, using the version released by Springer et al. (2024).[2]

We evaluate on the Massive Text Embedding Benchmark (MTEB; Muennighoff et al. 2023), which consists of 56 embedding datasets spanning seven different tasks. For a full list of tasks and

[1] In some cases, the hard negative is mined instead of generated. We refer to Appendix B for details.

[2] Springer et al. (2024) is the only work that replicates the dataset curation of Wang et al. (2024) and publicly releases it.

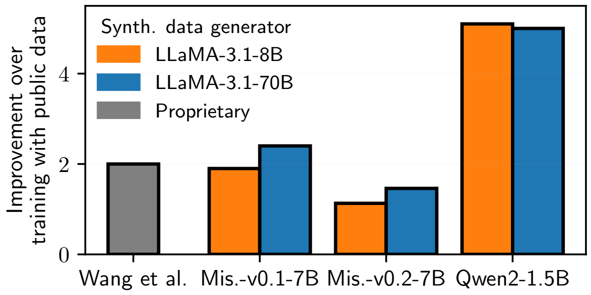


Figure 2: Improvement in MTEB score when adding synthetic data to the training mixture. Across all settings, our results are consistent with Wang et al. (2024), showing that training on synthetic data leads to higher MTEB performance.

Base model	Pub. data	L-8B	L-70B
Mistral-v0.1-7B	63.1	65.0	65.5
Mistral-v0.2-7B	65.4	66.6	**66.9**
Qwen2-1.5B	58.3	63.4	63.3
Wang et al.	64.6	66.6*	
Chen et al.		66.5*	

Table 1: MTEB scores for different base models and synthetic data generators. In all cases, adding synthetic data to the training mixture leads to improved performance. Note that Wang et al. and Chen et al. use the GPT model family for at least part of their synthetic data generation. See Appendix 3 for more detailed evaluations.

datasets along with the details of our evaluation, see Appendix D.

Synthetic data leads to improvements on MTEB. We compare our synthetic data to Wang et al. (2024), and Chen et al. (2024)—a method involving generating synthetic data from smaller language models that are fine-tuned to produce high quality synthetic data.

As shown in Table 1 and Figure 2, our synthetic data replicates the overall relative improvement of Wang et al. (2024) as well as Chen et al. (2024). We note that we do not match the exact scores of Wang et al. (2024), likely due to differences in the hyperparameters and general setup used for training. However, in all cases our synthetic data substantially improves performance. In fact, our Mistral-v0.2-7B model trained on LLaMA-3.1-70B synthetic data outperforms the scores reported by Wang et al. (2024), and Chen et al. (2024) reaching an absolute score of 66.9. It is noteworthy that synthetic data especially benefits Qwen2-1.5B (+5.1 points compared to +2.4 points for Mistral-v0.1-7B).

3 Investigating task influence

Having replicated the findings of Wang et al. (2024), i.e, that synthetic data improves MTEB performance on average, we now turn to studying to what extent the different task types of the synthetic data influences performance of each category of MTEB. More specifically, does each type of synthetic data broadly improve performance, or are improvements localized to particular tasks? Even more importantly, are there inherent performance *trade-offs* between different tasks based upon the exact synthetic data composition? Having answers to these questions is crucial for our understanding on how synthetic data impacts the overall generalization of text embedding models.

Estimating data influence. To address the questions posed above, we measure how each synthetic subset affects model performance with an *influence function*, which estimates the typical improvement that a specific subset of the training data contributes to the final performance. We consider specifically four of the synthetic data categories: short-short, short-long, long-long, and long-short. We train models with all possible $2^4 = 16$ combinations of this synthetic data, with the addition of a non-synthetic base dataset $\mathcal{D}$. To estimate the influence of a specific category $\mathcal{S}$, we split the 16 models into two groups: one group $\mathcal{P}_i^+$ of the eight models with training data that includes $\mathcal{S}$, and another group $\mathcal{P}_i^-$ of the eight remaining models that exclude $\mathcal{S}$ from the training data. We measure the influence of $\mathcal{S}$ by computing the difference between the mean performance of each group: $\mathbb{E}_{P\in\mathcal{P}_i^+}[\text{perf}(\mathcal{D}\cup P)] - \mathbb{E}_{P\in\mathcal{P}_i^-}[\text{perf}(\mathcal{D}\cup P)]$. This difference quantifies the improvement we expect to observe by training on each synthetic data category.

In addition to computing influence functions, we use a two-sided t-test to determine whether each synthetic category has a statistically significant (non-zero) contribution to MTEB performance.

Training and evaluation setup. We run experiments with Mistral-v0.2-7B and Qwen2-1.5B, and compare training on synthetic data from Llama-3.1-8B, and Llama-3.1-70B. For training, we adopt the same setup as §2. Evaluation also follows §2 (MTEB) with the caveat that retrieval datasets are replaced with their faster versions. We refer the

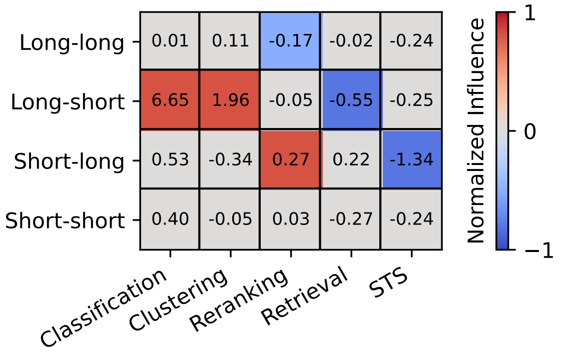


Figure 3: Influence of different training data categories on the MTEB evaluation categories when training Mistral-v0.2-7B on synthetic data from Llama-3.1-70B. Colored cells indicate statistically significant influence ($p < 0.05$). The color indicates the normalized influence: influence is rescaled so that the maximum (absolute) influence has a value of ±1. Additional results in Appendix 3.

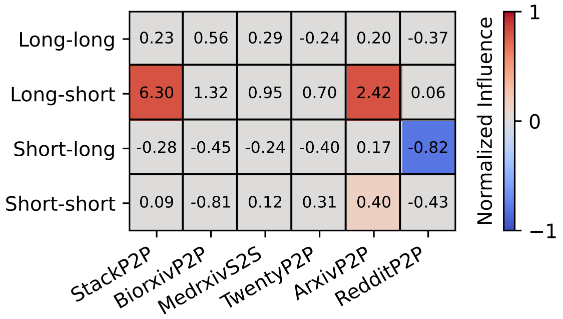


Figure 4: Influence of different training data categories on MTEB clustering tasks when training Mistral-v0.2-7B on synthetic data from Llama-3.1-70B. Colored cells indicate statistically significant influence ($p < 0.05$). The color indicates the normalized influence: influence is rescaled so that the maximum (absolute) influence has a value of ±1. Additional results in Appendix 3.

reader to Appendix C for full details.

Cross-category generalization is sometimes negative. Figure 3 plots the influence of each synthetic data set on each task category for Mistral-v0.2-7B using synthetic data from Llama-3.1-70B. Surprisingly, we find a trade-off between the performance of different categories. For example, training on the synthetic long-short dataset benefits classification and clustering performance by 6.65 and 1.19 points on average. Similarly, short-long improves reranking performance (+0.27 points) but harms sentence similarity (−1.34 points). We observe similar trade-offs when using Qwen2-1.5B as a base model, and when training with LLaMA-3.1-8B synthetic data.

Synthetic data improves performance sparsely. We indicate the statistical significance of each influence estimate in Figure 3. Often, the synthetic data has *no* statistically significant influence on many of the MTEB evaluation categories. In fact, short-short data has no statistically significant improvement for any MTEB category.

Even within a particular evaluation category that improves with synthetic data, the majority of tasks observe no statistically significant improvement with synthetic data. For example, we observe that the long-short dataset, which generally improves the performance of clustering tasks, only improves StackExchangeClusteringP2P and ArxivClusteringP2P with significance (see Figure 4).

4 Related work

General-purpose text encoders. A major challenge when training embedding models is that the notion of similarity these models learn is corpus-driven and hence, models often fail to generalize beyond the similarity definitions they saw during contrastive training (Thakur et al., 2021; Muennighoff et al., 2023; Ravfogel et al., 2024).

Driven by benchmarks such as BEIR (Thakur et al., 2021) and MTEB (Muennighoff et al., 2023), the community has shifted its focus on building text embedding methods that generalize to multiple tasks and domains. While prior BERT-based approaches relied on complex multi-stage pipelines to achieve this goal (Li et al., 2023; Xiao et al., 2023, *inter alia*), recent approaches which are based on decoder-only LLMs have shown superior performance (Wang et al., 2024; BehnamGhader et al., 2024; Springer et al., 2024; Muennighoff et al., 2024; Lee et al., 2025, *inter alia*). Leveraging advances in instruction-following capabilities of LLMs, these models achieve improved generalization to novel tasks and domains by using natural language instruction combined with multi-task contrastive learning.

Synthetic data for text embeddings. Synthetic data has been previously used in the information retrieval literature to generate pseudo queries or hypothetical documents (Nogueira et al., 2019; Dai et al., 2023; Wang et al., 2023, *inter alia*). Recently, Wang et al. (2024) extended this paradigm to general-purpose text embedding methods, by using LLMs to generate high-quality training data for diverse embedding tasks. Since then, synthetic data generation has become a widely adopted strategy for improving text embedding models, particularly for models competing on the MTEB leaderboard.

5 Conclusion

In this work, we investigate how training on synthetic data influences downstream text embedding tasks. Our experiments confirm previous findings that synthetic data can significantly boost overall performance on MTEB, but offer a more nuanced perspective: we find that training on certain synthetic data categories can exhibit trade-offs in the task-specific performances. For example, while some synthetic data categories improve classification or clustering, they may degrade retrieval performance. Our results highlight that the addition of synthetic data does not always strictly improve text embedding models. Instead, its benefits are nuanced, requiring refined generation and training strategies to balance trade-offs.

Beyond our analysis, we contribute a high quality reproduction of the synthetic data of Wang et al. (2024) which we release publicly, along with code for our reproduction. By releasing our data and code, we aim to support further research into optimizing synthetic training for general-purpose text embeddings.

Limitations

Role of the base dataset. In our experiments, we follow previous work (Wang et al., 2024; Muennighoff et al., 2024) and combine synthetic data with an existing mixture of publicly available datasets (Springer et al., 2024). There might be some non-trivial interactions between our synthetic dataset and the existing data which we did not control for in our setup. We hope that releasing our synthetic data will encourage future work to explore potential interactions between the base dataset and the synthetic data added to it.

Synthetic data generated from more capable models. We cannot rule out the possibility that synthetic data generated from more capable LLMs might lead to different conclusions from ours. However, as with the role of the base dataset discussed above, we hope that by releasing our data to the community, we make it easy for future work to compare to our data and investigate potential properties of synthetic data that are responsible for (larger) improvements in generalization.

Acknowledgments

Marius Mosbach is supported by the Mila P2v5 grant and the Mila-Samsung grant. Siva Reddy is supported by the Canada CIFAR AI Chairs program and the NSERC Discovery Grant program.

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE2140739. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation.

We gratefully acknowledge support from ServiceNow, Apple, NSF and the AI2050 program at Schmidt Sciences (Grant #G2264481).

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. 2024. Little giants: Synthesizing high-quality embedding data at scale. *arXiv preprint arXiv:2410.18634*.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations*.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Veysel Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Suppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan

Dey, Dipam Vasani, Pranjal A Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Mariya Hendriksen, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri K, Maksimova Anna, Silvan Wehrli, Maria Tikhonova, Henil Shalin Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Validad Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. MMTEB: Massive multilingual text embedding benchmark. In *The Thirteenth International Conference on Learning Representations*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint*.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral: Enhance text retrieval with transfer learning. Salesforce AI Research Blog.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

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. *arXiv preprint*.

Shauli Ravfogel, Valentina Pyatkin, Amir DN Cohen, Avshalom Manevich, and Yoav Goldberg. 2024. Description-based text similarity. *First Conference on Language Modeling*.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2024. Repetition improves language model embeddings. *arXiv preprint*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423, Singapore. Association for Computational Linguistics.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-Pack: Packaged resources to advance general chinese embedding. *arXiv preprint*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

A Background

Let $\mathcal{V}$ be a finite vocabulary of tokens. A **sequence** $s \in \mathcal{V}^*$ is a finite concatenation of tokens $\mathbf{s} = (s_1, \dots, s_{|s|})$ where each $s_i \in \mathcal{V}$. Our goal is to train an **embedder** $\phi : \mathcal{V}^* \rightarrow \mathcal{R}^d$, parameterized by $\theta \in \mathcal{R}^p$, which maps sequences to embeddings. The ideal embedder accurately estimates the similarity between any pair of examples $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{V}^*$, parameterized by a metric of similarity between the embeddings. We follow common practice in the embedding literature and use cosine similarity as to estimate the similarity between examples $\hat{f}(\mathbf{s}_1, \mathbf{s}_2) = \cos(\phi(\mathbf{s}_1), \phi(\mathbf{s}_2))$ (Li et al., 2023; Wang et al., 2024; Springer et al., 2024; BehnamGhader et al., 2024; Muennighoff et al., 2024, *inter alia*).

Contrastive learning. We train a general-purpose embedding model via contrastive learning on a dataset $\mathcal{D} = \{\langle s_i, s^+{}_i, s^-{}_i \rangle\}_{i=1}^n$. Below we state the loss for a single instance $\langle s_i, s^+{}_i, s^-{}_i \rangle$:

$$\mathcal{L}(\langle s_i, s^+{}_i, s^-{}_i \rangle) \tag{1}$$
$$= -\log \frac{\exp(\hat{f}(s_i, s^+{}_i))}{\exp(\hat{f}(s_i, s^+{}_i)) + \exp(\hat{f}(s_i, s^-{}_i))} \quad .$$

For simplicity, we stick to a single negative example above but note that in practice, we typically consider multiple negative examples per instance. In that case, the second term in the denominator becomes $\sum_j \exp(s(s_i, s^-{}_{i,j}))$, where we take a sum over all negative examples for a given input s_i.

B Synthetic data generation

We follow the synthetic data generation pipeline of Wang et al. (2024). We follow their prompt template for both brainstorming and instance generation. The composition of synthetic data for both LLaMA-3.1-8B and LLaMA-3.1-70B across different categories in detailed in Table 2. We also provide an example of short-short category sample in Table 4, short-long in Table 5, long-short in Table 6, and long-long in Table 7.

C Training details

Wang et al. (2022) were among the first to demonstrate that a powerful decoder-only LLM can be transformed into a high-quality text encoder. To obtain text embedding, they appended an [EOS] token to each input and constructed an embedding form its last layer representation. They fine-tuned on sentence-pair data using a contrastive learning objective (see Appendix A) to ensure that the resulting text representations effectively capture the semantic content of the input text. Wang et al. trained Mistral-7B (`Mistral-7B-v0.1`, Jiang et al. 2023) on 1.8 million sentence pairs. They perform parameter-efficient fine-tuning via LoRA (Hu et al., 2022), using a batch size of 2048. The entire fine-tuning process takes roughly 18 hours on 32 V100 GPUs.

Our training procedure largely follows Wang et al., but we make minor modifications inspired by subsequent work. First, we use a more recent base model from the same model family (`Mistral-7B-Instruct-v0.2`). Next, following BehnamGhader et al. (2024) and Muennighoff et al. (2024), we enable bidirectional connections within the model architecture and employ mean pooling over the token embeddings instead of relying on the final [EOS] token representation.

For the public portion of Wang et al.'s training data, we use the replication provided by Springer et al. (2024). This data consists of about 1.5M samples. For the synthetic portion, we generate about ~500,000 samples following the methodology described in §2.

We train the models with LoRA $r = 16$ and $\alpha = 16$ using a batch size of 2048. We use a maximum sequence length of 512 tokens for fair comparison to previous approaches. We use the AdamW optimizer with a learning rate of $4e-4$, linear learning rate warm-up for the first 100 steps, and weight decay with 0.1 coefficient afterwards. We train all models for one epoch. Training Mistral-7B on public + synthetic data (~2M samples) takes about 16 hours on 8 H100 GPUs. We will release the dataset, pre-trained models and the training code upon publication.

D MTEB evaluation details

Text embeddings have been widely used in various NLP tasks, however, traditionally the evaluation of text embeddings has been limited to a small set of datasets from a single task such as semantic textual similarity or text retrieval (Karpukhin et al., 2020; Wang et al., 2021), making it difficult to estimate generalization of the proposed methods.

To address this issue, Muennighoff et al. (2023) proposed MTEB – Massive Text Embedding Benchmark – a single comprehensive evaluation suite that spans a total of 56 datasets across 7 dis-

Model	short-short	short-long	long-long	long-short	bitext	STS	Total
LLaMA-3.1-8B	19,769	146,717	17,344	106,577	88,228	99,612	478,247
LLaMA-3.1-70B	19,932	153,934	19,236	108,487	89,611	99,791	490,991

Table 2: Composition of synthetic data from LLaMA-3.1-8B and LLaMA-3.1-70B across different categories.

tinct tasks (retrieval (15), reranking (4), classification (12), clustering (11), pair classification (3), semantic textual similarity (STS, 10), and summarization (1). The individual dataset sizes vary widely: STS datasets range from 1K to 20K pairs, classification datasets range from 500-5000 samples, and retrieval datasets such as MS MARCO (Bajaj et al., 2018) include thousands of queries with a 6M document corpus.

In MTEB, every task is reformulated as an embedding task where the only requirement is that the model produces a vector (embedding) for each text input. For example, in classification, MTEB uses the embeddings as fixed features and trains a lightweight linear classifier (typically logistic regression) on top. The performance of the classifier is treated as a proxy for the quality of the embeddings. In clustering, embeddings are fed into standard clustering algorithms (like mini-batch k-means) to group similar texts. Retrieval, reranking and STS follow standard evaluation protocol is which embeddings of pairs of text are compared using cosine similarity.

The unified casting of diverse tasks into an embedding framework, simplicity of use, open-source evaluation code[3], and a public leaderboard[4] has led to widespread use of this benchmark within the NLP community, making it the de facto standard for evaluating text embedding models.

Faster version of MTEB. One drawback of MTEB is that evaluating a model is very computationally intensive. This is largely due to the retrieval task category, where each dataset has separate corpus containing millions of documents. To address this, Enevoldsen et al. (2025) developed smaller versions of the retrieval datasets contained in MTEB by carefully selecting candidate documents for each query in the dataset. They showed that keeping only 250 documents per query, selected via hard-negative mining, maintains the absolute scores and model ranking compared to evaluating on the original datasets. While we use original datasets for evaluation in §2 for a fair comparison with Wang et al. (2022), our analysis in §3 uses the faster version of the retrieval datasets.

E Additional evaluations

E.1 Full MTEB evaluations

We evaluate each model that we consider in the main paper on MTEB in Table 1. In this section, we expand these results by plotting the average scores for each of the categories in Table 3. We find that our synthetic data improves clustering performance most substantially.

E.2 Omitted influence plots

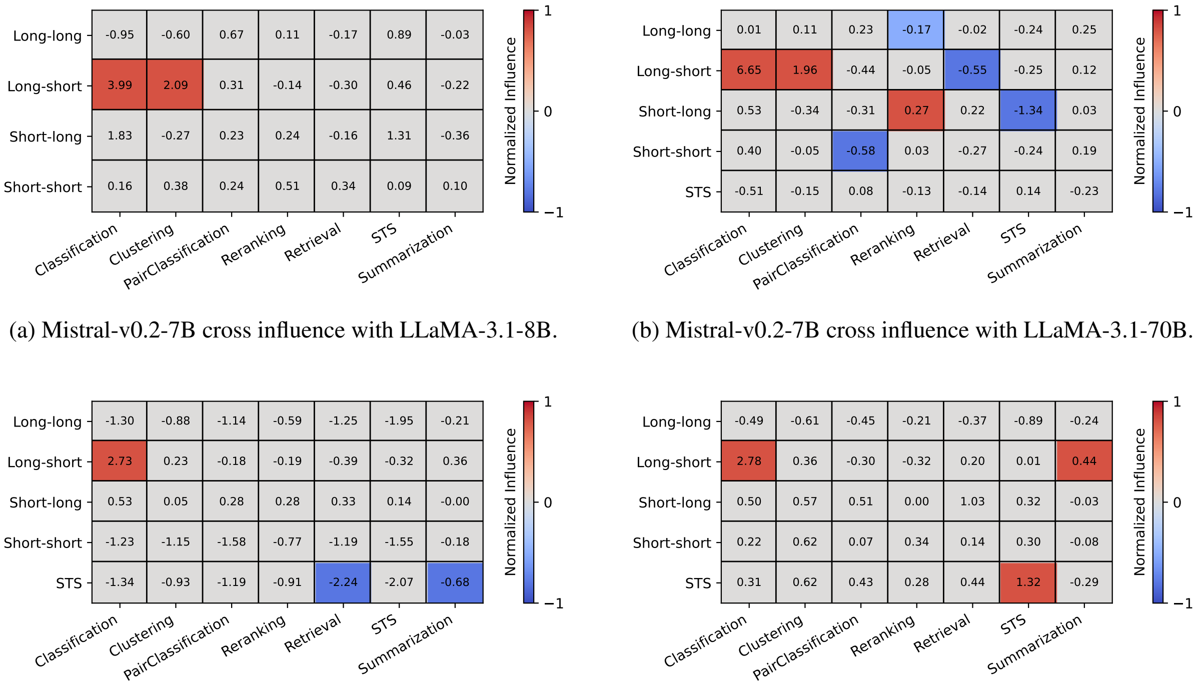
We extend Figures 3 and 4 with the results from all four settings: Mistral-v0.2-7B and Qwen2-1.5B base models, trained on LLaMA-3.1-8B or LLaMA-3.1-70B synthetic data. We plot the results in Figures 5, 6, 7, 8, and 9. In general, we observe similar trends from the main paper.

[3] https://github.com/embeddings-benchmark/mteb

[4] https://huggingface.co/spaces/mteb/leaderboard

		Cls.	**Clust.**	**P. Clas.**	**Rera.**	**Retr.**	**STS**	**Sum.**	**Mean**
# of datasets →		12	11	3	4	15	10	1	56
Qwen2-1.5B	Public data only	70.2	42.3	83.1	51.6	48.2	76.2	31.4	58.3
Qwen2-1.5B	LLaMA-3.1-8B	76.6	47.0	86.2	55.5	54.1	80.2	30.8	63.4
Qwen2-1.5B	LLaMA-3.1-70B	76.5	47.2	86.9	55.0	54.4	79.3	29.7	63.3
Mis.-v0.1-7B	Public data only	75.6	44.3	86.3	57.5	54.7	81.4	30.0	63.1
Mis.-v0.1-7B	LLaMA-3.1-8B	77.6	47.3	87.3	58.2	56.5	83.1	30.8	65.0
Mis.-v0.1-7B	LLaMA-3.1-70B	77.9	49.1	87.4	57.6	57.1	82.9	30.1	65.5
Mis.-v0.2-7B	Public data only	76.7	46.8	87.6	59.0	58.7	83.7	30.1	65.4
Mis.-v0.2-7B	LLaMA-3.1-8B	78.7	49.8	87.9	58.5	59.1	83.7	30.7	66.6
Mis.-v0.2-7B	LLaMA-3.1-70B	78.3	50.5	88.2	60.0	58.2	85.7	31.3	66.9

Table 3: Full MTEB evaluations for each model.



(a) Mistral-v0.2-7B cross influence with LLaMA-3.1-8B.

(b) Mistral-v0.2-7B cross influence with LLaMA-3.1-70B.

(c) Qwen2-1.5B cross influence with LLaMA-3.1-8B.

(d) Qwen2-1.5B cross influence with LLaMA-3.1-70B.

Figure 5: Comparison of cross-influence results for different models and synthetic datasets.

Synthetic data category: short-short

Instruction: Identify a famous painting from a brief description.

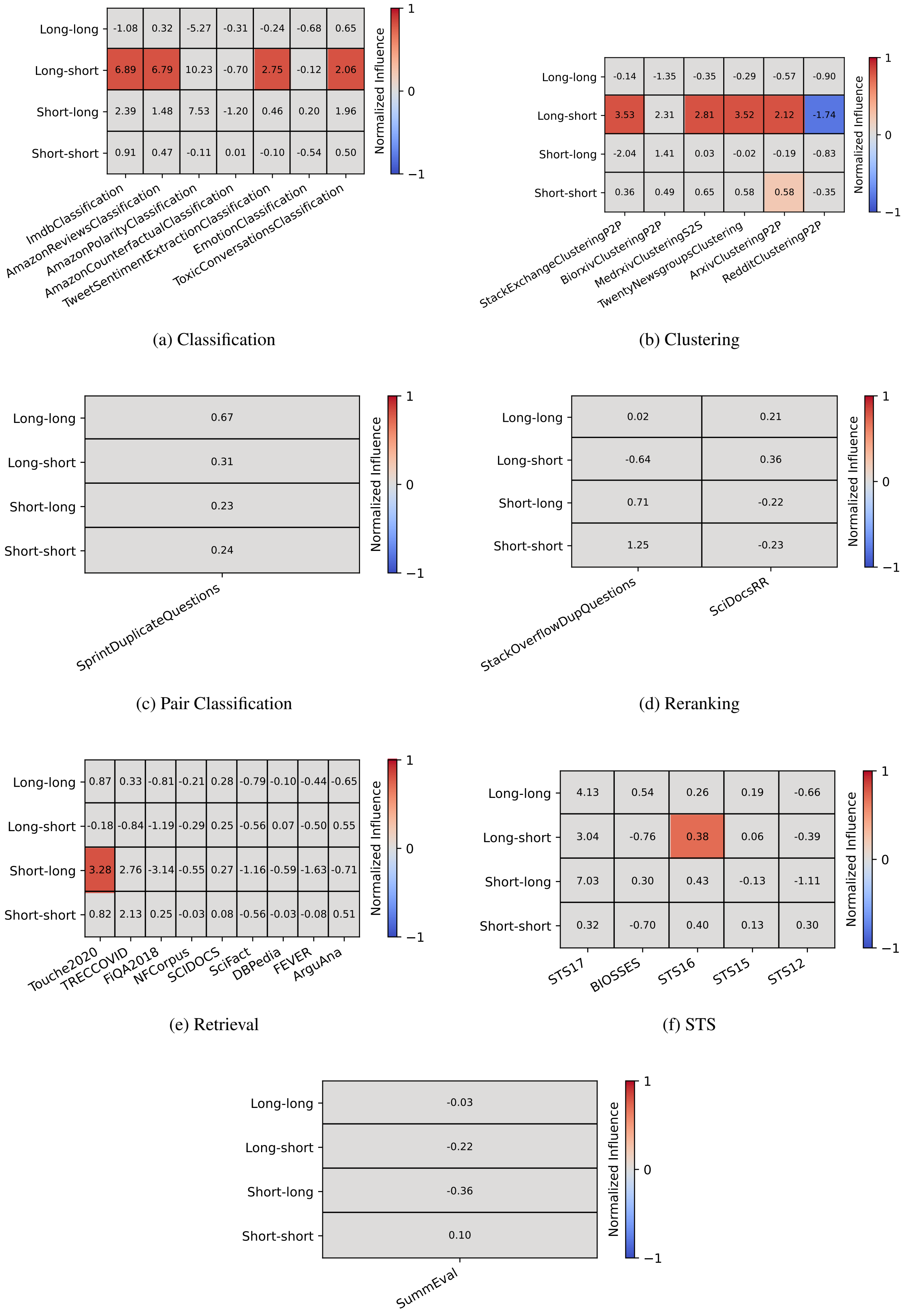
Llama-3.1-70B

Generated data: {
"query": "Painting of a woman with a mysterious smile",
"positive": "The Mona Lisa by Leonardo da Vinci",
"negative": "Claude Monet"
}

Llama-3.1-8B

Generated data: {
"query": "a woman with a clock",
"positive": "Girl with a Pearl Earring by Johannes Vermeer",
"negative": "a toy that runs using a spring"
}

Table 4: Sample from *short-short* subgroup of the synthetic data.



(a) Classification

(b) Clustering

(c) Pair Classification

(d) Reranking

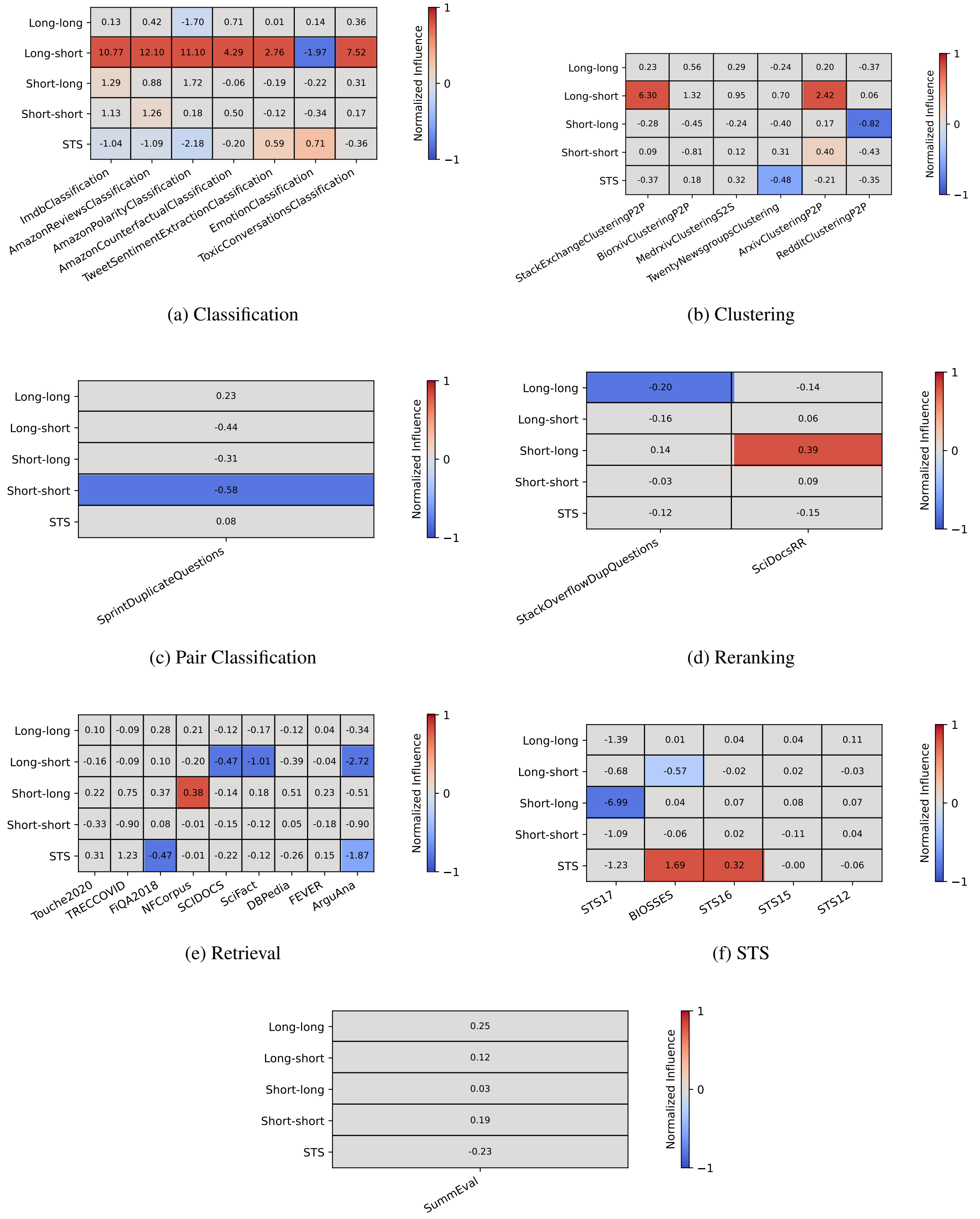
(e) Retrieval

(f) STS

(g) Summarization

Figure 6: Detailed breakdown of Mistral-v0.2-7B influence on various tasks with LLaMA-3.1-8B synthetic data.

]



(a) Classification

(b) Clustering

(c) Pair Classification

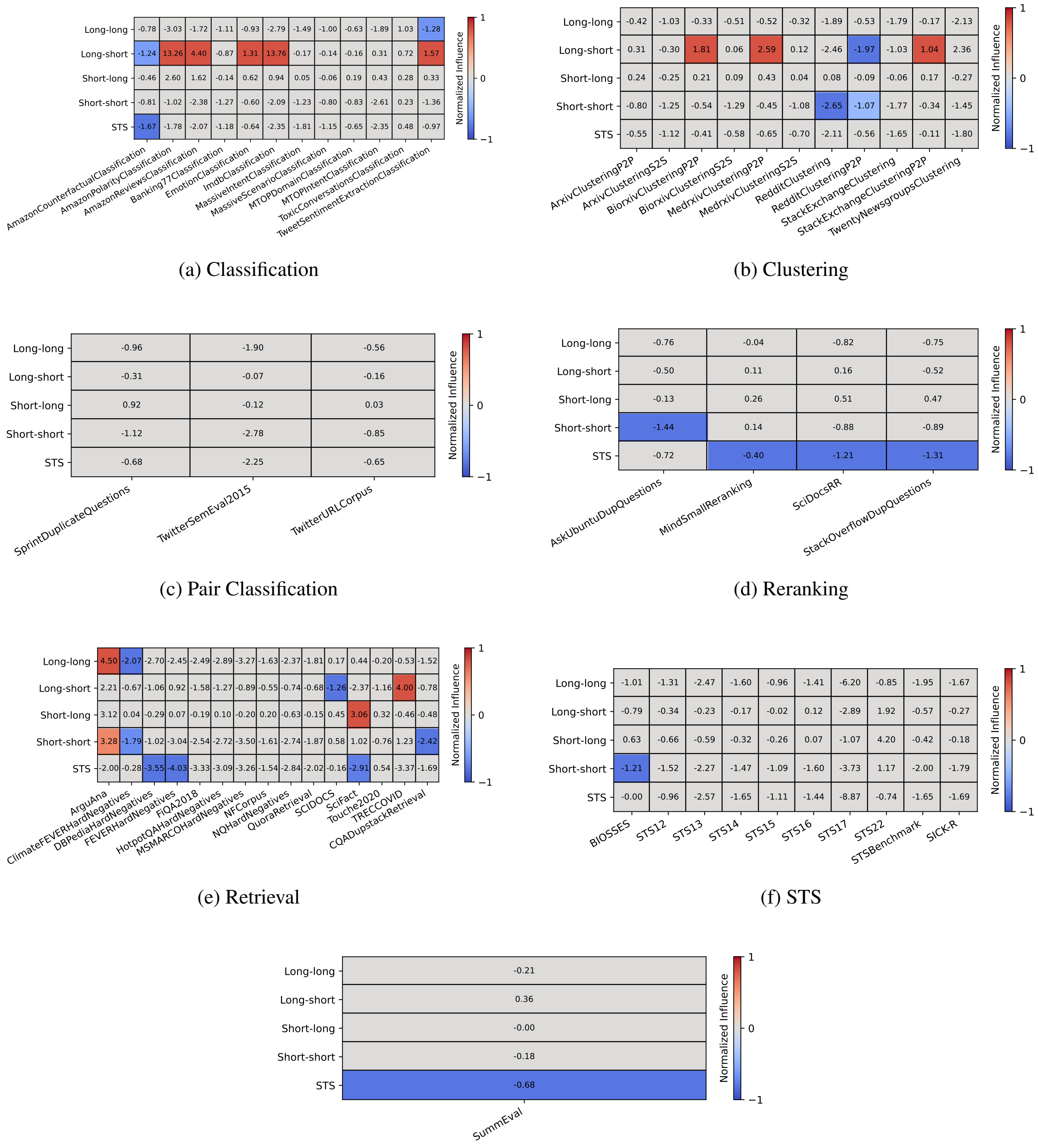
(d) Reranking

(e) Retrieval

(f) STS

(g) Summarization

Figure 7: Detailed breakdown of Mistral-v0.2-7B influence on various tasks with LLaMA-3.1-70B synthetic data.



(a) Classification

	AmazonCounterfactualClassification	AmazonPolarityClassification	AmazonReviewsClassification	Banking77Classification	EmotionClassification	ImdbClassification	MassiveIntentClassification	MassiveScenarioClassification	MTOPDomainClassification	MTOPIntentClassification	ToxicConversationsClassification	TweetSentimentExtractionClassification
Long-long	-0.78	-3.03	-1.72	-1.11	-0.93	-2.79	-1.49	-1.00	-0.63	-1.89	1.03	-1.28
Long-short	-1.24	13.26	4.40	-0.87	1.31	13.76	-0.17	-0.14	-0.16	0.31	0.72	1.57
Short-long	-0.46	2.60	1.62	-0.14	0.62	0.94	0.05	-0.06	0.19	0.43	0.28	0.33
Short-short	-0.81	-1.02	-2.38	-1.27	-0.60	-2.09	-1.23	-0.80	-0.83	-2.61	0.23	-1.36
STS	-1.67	-1.78	-2.07	-1.18	-0.64	-2.35	-1.81	-1.15	-0.65	-2.35	0.48	-0.97

(b) Clustering

	ArxivClusteringP2P	ArxivClusteringS2S	BiorxivClusteringP2P	BiorxivClusteringS2S	MedrxivClusteringP2P	MedrxivClusteringS2S	RedditClustering	RedditClusteringP2P	StackExchangeClustering	StackExchangeClusteringP2P	TwentyNewsgroupsClustering
Long-long	-0.42	-1.03	-0.33	-0.51	-0.52	-0.32	-1.89	-0.53	-1.79	-0.17	-2.13
Long-short	0.31	-0.30	1.81	0.06	2.59	0.12	-2.46	-1.97	-1.03	1.04	2.36
Short-long	0.24	-0.25	0.21	0.09	0.43	0.04	0.08	-0.09	-0.06	0.17	-0.27
Short-short	-0.80	-1.25	-0.54	-1.29	-0.45	-1.08	-2.65	-1.07	-1.77	-0.34	-1.45
STS	-0.55	-1.12	-0.41	-0.58	-0.65	-0.70	-2.11	-0.56	-1.65	-0.11	-1.80

(c) Pair Classification

	SprintDuplicateQuestions	TwitterSemEval2015	TwitterURLCorpus
Long-long	-0.96	-1.90	-0.56
Long-short	-0.31	-0.07	-0.16
Short-long	0.92	-0.12	0.03
Short-short	-1.12	-2.78	-0.85
STS	-0.68	-2.25	-0.65

(d) Reranking

	AskUbuntuDupQuestions	MindSmallReranking	SciDocsRR	StackOverflowDupQuestions
Long-long	-0.76	-0.04	-0.82	-0.75
Long-short	-0.50	0.11	0.16	-0.52
Short-long	-0.13	0.26	0.51	0.47
Short-short	-1.44	0.14	-0.88	-0.89
STS	-0.72	-0.40	-1.21	-1.31

(e) Retrieval

	ArguAna	ClimateFEVERHardNegatives	DBPediaHardNegatives	FEVERHardNegatives	FiQA2018	HotpotQAHardNegatives	MSMARCOHardNegatives	NFCorpus	NQHardNegatives	QuoraRetrieval	SCIDOCS	SciFact	Touche2020	TRECCOVID	CQADupstackRetrieval
Long-long	4.50	-2.07	-2.70	-2.45	-2.49	-2.89	-3.27	-1.63	-2.37	-1.81	0.17	0.44	-0.20	-0.53	-1.52
Long-short	2.21	-0.67	-1.06	0.92	-1.58	-1.27	-0.89	-0.55	-0.74	-0.68	-1.26	-2.37	-1.16	4.00	-0.78
Short-long	3.12	0.04	-0.29	0.07	-0.19	0.10	-0.20	0.20	-0.63	-0.15	0.45	3.06	0.32	-0.46	-0.48
Short-short	3.28	-1.79	-1.02	-3.04	-2.54	-2.72	-3.50	-1.61	-2.74	-1.87	0.58	1.02	-0.76	1.23	-2.42
STS	-2.00	-0.28	-3.55	-4.03	-3.33	-3.09	-3.26	-1.54	-2.84	-2.02	-0.16	-2.91	0.54	-3.37	-1.69

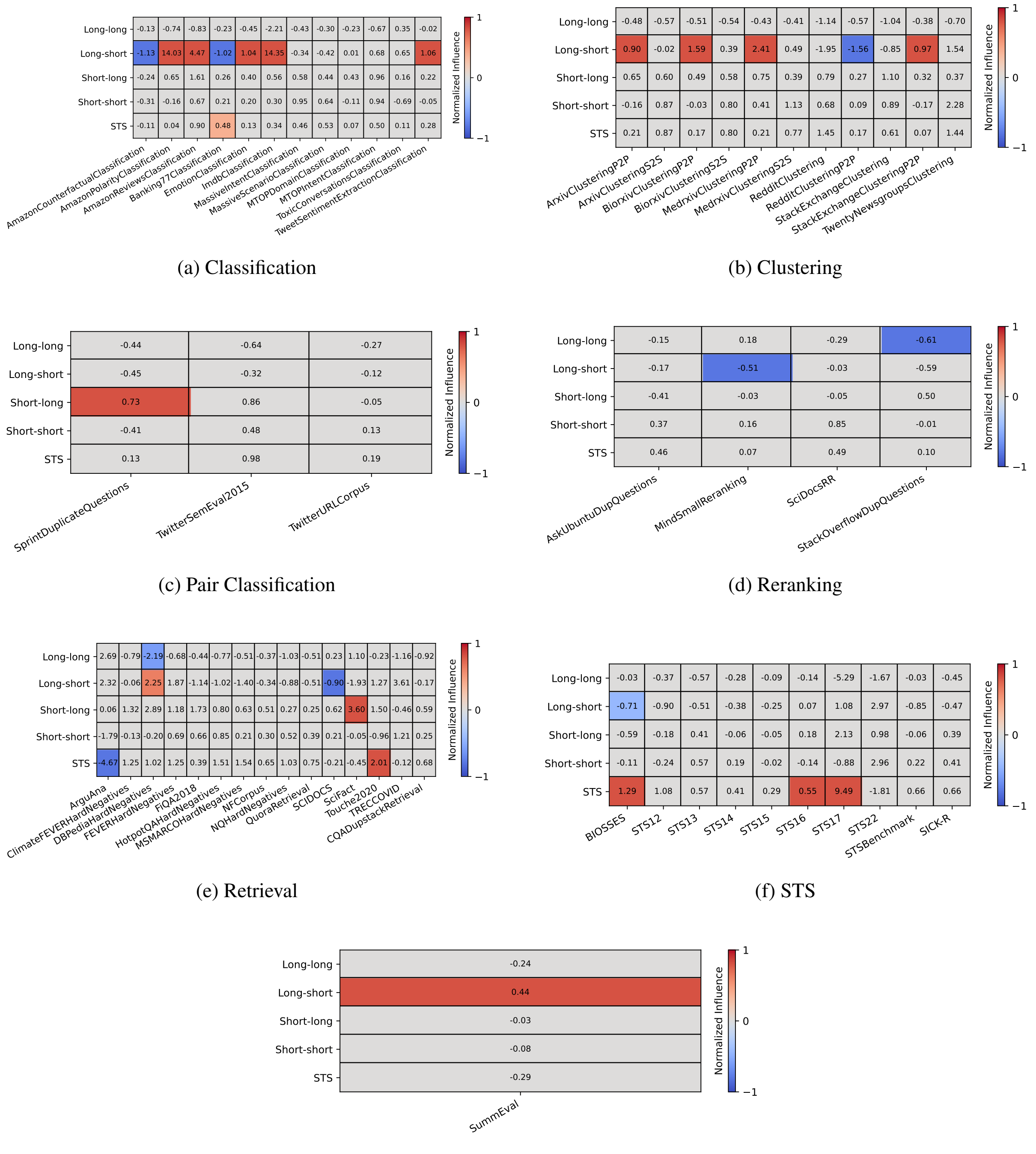
(f) STS

	BIOSSES	STS12	STS13	STS14	STS15	STS16	STS17	STS22	STSBenchmark	SICK-R
Long-long	-1.01	-1.31	-2.47	-1.60	-0.96	-1.41	-6.20	-0.85	-1.95	-1.67
Long-short	-0.79	-0.34	-0.23	-0.17	-0.02	0.12	-2.89	1.92	-0.57	-0.27
Short-long	0.63	-0.66	-0.59	-0.32	-0.26	0.07	-1.07	4.20	-0.42	-0.18
Short-short	-1.21	-1.52	-2.27	-1.47	-1.09	-1.60	-3.73	1.17	-2.00	-1.79
STS	-0.00	-0.96	-2.57	-1.65	-1.11	-1.44	-8.87	-0.74	-1.65	-1.69

(g) Summarization

	SummEval
Long-long	-0.21
Long-short	0.36
Short-long	-0.00
Short-short	-0.18
STS	-0.68

Figure 8: Detailed breakdown of Qwen2-1.5B influence on various tasks with LLaMA-3.1-8B synthetic data.



(a) Classification

(b) Clustering

(c) Pair Classification

(d) Reranking

(e) Retrieval

(f) STS

(g) Summarization

Figure 9: Detailed breakdown of Qwen2-1.5B influence on various tasks with LLaMA-3.1-70B synthetic data.

Synthetic data category: short-long

Instruction: Search for documentaries about the effects of pollution on human health.

Llama-3.1-70B

Generated data: {
"query": "Documentaries exploring the long-term health effects of microplastic consumption in human populations",
"positive": "The Devil We Know is a 2018 documentary that examines the effects of plastics on human health and the environment. The film focuses on the history of plastics and their impact on human health, including the leaching of chemicals into food and water. ...***truncated for brevity*** .... While the film does not focus exclusively on microplastic consumption, it provides a comprehensive overview of the issue and its effects on human health. The documentary has been praised for its thorough research and engaging storytelling, and has won several awards for its contribution to the public understanding of environmental issues. The film is a valuable resource for anyone looking to understand the impact of plastics on human health, and the ways in which individuals and communities can work to reduce their exposure to these pollutants.",
"negative": "Plastic Ocean is a 2016 documentary that explores the impact of plastic pollution on marine ecosystems. The film follows a team of researchers as they travel the world, documenting the effects of plastic waste on ocean life. ...***truncated for brevity*** .... The film has been praised for its visually stunning footage and its ability to raise awareness about the issue of plastic pollution."
}

Llama-3.1-8B

Generated data: {
"query": "The effects of particulate matter on cardiovascular disease in low-income communities in urban areas with high levels of air pollution",
"positive": "Air pollution is a significant public health concern worldwide, particularly in low-income communities. The World Health Organization (WHO) estimates that 9 out of 10 people worldwide breathe polluted air, resulting in over 7 million premature deaths annually. One of the primary pollutants responsible for this alarming rate is particulate matter (PM). ...***truncated for brevity*** ....To mitigate the effects of PM on cardiovascular disease, several strategies can be implemented. These include reducing emissions from vehicles, industrial sites, and power plants, increasing the use of renewable energy sources, and promoting green infrastructure such as green roofs and urban parks. Additionally, improving air quality monitoring and public education can help raise awareness about the risks associated with PM exposure. In conclusion, the effects of particulate matter on cardiovascular disease in low-income communities in urban areas with high levels of air pollution are a pressing public health concern. It is essential to address this issue through a combination of policy changes, technological innovations, and community-based interventions to reduce PM exposure and protect the health and well-being of vulnerable populations.",
"negative": "The effects of air pollution on human health have been a topic of interest for many years. However, the relationship between particulate matter and cardiovascular disease is not as clear-cut as previously thought. While some studies have found a link between PM exposure and cardiovascular disease, others have found no significant association. A 2019 study published in the Journal of Exposure Science & Environmental Epidemiology found that exposure to PM2.5 was associated with an increased risk of cardiovascular disease in a cohort of older adults. However, this study had several limitations, including a small sample size ...***truncated for brevity*** .... In conclusion, while some studies have found a link between PM exposure and cardiovascular disease, the evidence is not conclusive, and more research is needed to fully understand the relationship between these two factors. It is essential to consider the complexity of the issue and the potential confounding variables when interpreting the results of studies on this topic.
}

Table 5: Sample from *short-long* subgroup of the synthetic data.

Synthetic data category: long-short
Instruction: Classify data breach notifications into high, medium, and low risk categories.
Llama-3.1-70B

Generated data: {
"query": "A security incident has been reported at a prominent e-commerce platform, affecting the sensitive information of over 500,000 customers. According to the company's investigation, a vulnerability in their login system allowed unauthorized access to their database. While the company assures that no financial data was compromised, email addresses, phone numbers, and physical addresses were exposed. The incident occurred due to a misconfigured security update, which left the system open to attacks for a period of three weeks. The company is working closely with law enforcement agencies to identify the perpetrators and has implemented additional security measures to prevent similar incidents in the future. As a precautionary measure, all affected customers have been notified and advised to change their passwords immediately.",
"positive": "high risk",
"negative": "medium risk"
}

Llama-3.1-8B

Generated data: {
"query": "A sophisticated phishing attack targeted a large corporation's IT department, resulting in the unauthorized access to sensitive customer data. The attackers used a combination of social engineering tactics and zero-day exploits to breach the company's network, ultimately gaining access to over 1 million customer records. The breach was discovered after a routine security audit revealed suspicious network activity. The affected data includes names, addresses, and payment card information. The company is taking immediate action to notify affected customers and is cooperating with law enforcement to investigate the incident.",
"positive": "high",
"negative": "medium"
}

Table 6: Sample from *long-short* subgroup of the synthetic data.

Synthetic data category: long-long

Instruction: Given comprehensive insights into the utility of meditation, locate opposing views arguing against its psychological benefits.

Llama-3.1-70B

Generated data: {
"query": "While numerous studies have demonstrated the benefits of meditation for both physical and mental well-being, a growing body of research has begun to challenge the notion that meditation is a universally beneficial practice. Critics argue that meditation can have negative effects on certain individuals, particularly those with a history of trauma or mental health issues. For example, some studies have found that mindfulness meditation can exacerbate symptoms of anxiety and depression in individuals who are prone to rumination, or the tendency to repetitively think negative thoughts. Additionally, meditation can be used as a form of avoidance, allowing individuals to disengage from their problems rather than confronting and resolving them. Furthermore, the emphasis on individualism and self-improvement in Westernized forms of meditation can be alienating for individuals from collectivist cultures, who may view meditation as a selfish pursuit. Moreover, the commercialization of meditation has led to the proliferation of inexperienced and unqualified teachers, who may not be equipped to provide adequate guidance and support to students. In light of these findings, it is essential to approach the practice of meditation with a critical and nuanced perspective, recognizing both its potential benefits and limitations.",
"positive": "The benefits of meditation have been extensively researched and documented, with studies demonstrating its positive impact on both physical and mental health. ...***truncated for brevity*** .... As the scientific evidence supporting the benefits of meditation continues to grow, it is becoming increasingly clear that this ancient practice has the potential to improve overall well-being and quality of life.",
"negative": "A recent study published in the Journal of Clinical Psychology found that participants in an MBSR program exhibited significant reductions in symptoms of anxiety and depression, as well as improved mood and quality of life. The study, which was conducted at a large urban medical center, included 100 participants who were randomly assigned to either an MBSR group or a wait-list control group. ...***truncated for brevity*** .... Participants in the wait-list control group did not receive any mindfulness training, but were given the option to participate in the MBSR program after the study was completed. The results of the study showed that participants in the MBSR group exhibited significant reductions in pain and improved mood, compared to participants in the wait-list control group. The study's authors note that further research is needed to fully understand the mechanisms by which MBSR programs achieve their benefits for individuals with chronic pain, but suggest that the program may be a useful tool for improving pain management outcomes."
}

Llama-3.1-8B

Generated data: {
"query": "The effectiveness of meditation in reducing stress and anxiety has been extensively documented in various scientific studies. Meditation has been shown to decrease the production of stress hormones, such as cortisol, and increase the production of neurotransmitters that promote feelings of relaxation and well-being. Regular meditation practice has also been linked to improved sleep quality, reduced blood pressure, and a stronger immune system. Furthermore, meditation has been used as a tool for managing chronic pain, depression, and anxiety disorders. The benefits of meditation are not limited to physical health; it has also been shown to improve cognitive functioning, including attention, memory, and problem-solving skills. Meditation can be practiced in various forms, including mindfulness meditation, loving-kindness meditation, and transcendental meditation. The scientific community has extensively studied the effects of meditation, and the evidence suggests that it is a valuable tool for promoting overall well-being and reducing stress and anxiety. What are some opposing views that argue against the psychological benefits of meditation?",
"positive": "Meditation has been widely accepted as a valuable tool for promoting mental and physical well-being. Research has shown that regular meditation practice can reduce stress and anxiety, improve sleep quality, and even alleviate symptoms of depression and anxiety disorders....***truncated for brevity*** .... The scientific community has extensively studied the effects of meditation, and the evidence suggests that it is a valuable tool for promoting overall well-being and reducing stress and anxiety. Regular meditation practice has been shown to have a positive impact on both physical and mental health, making it an important component of a healthy lifestyle.",
"negative": "While the benefits of prioritizing mental health are well-documented, there are also potential drawbacks to consider. For example, a study published in the Journal of Behavioral Addictions found that excessive exercise and physical activity can actually exacerbate anxiety and depression in some individuals (Kvam et al., 2016). ...***truncated for brevity*** .... Moreover, some researchers have argued that the emphasis on individual responsibility for mental health can be problematic, as it can lead to stigma and shame for those who are struggling (Slaby, 2002). By taking a more nuanced approach to mental health, one that acknowledges the complexities and limitations of the current understanding, we can work towards creating a more comprehensive and effective approach to mental health support."

}

Table 7: Sample from *long-long* subgroup of the synthetic data.