Benchmarking LLMs on the Semantic Overlap Summarization Task

John Salvador, Naman Bansal, Mousumi Akter, Souvika Sarkar,
Anupam Das, and Shubhra Kanti Karmaker ("Santu")
Big Data Intelligence Lab
Department of Computer Science and Software Engineering
College of Engineering, Auburn University
{jms0256, nbansal, mza0170, szs0239, sks0086}@auburn.edu
anupam.das@ncsu.edu

Abstract

This paper presents a comprehensive evaluation of Large Language Models (LLMs) on the Semantic Overlap (SOS) Task, assessing their ability to extract overlapping information from multiple narratives. Utilizing established metrics like ROUGE, BERTscore, and Sem-F1, we compare the quality of LLM-generated summaries against reference summaries, providing insights into their effectiveness in capturing semantically overlapping information. We also introduce a novel dataset for the SOS task, facilitating robust experimentation and benchmarking. Through experimentation, we analyze the strengths and limitations of various LLMs, offering valuable insights into their capabilities in capturing overlapping information. The code and datasets used to conduct this study are available at https://github.com/jmsalvador2395/llm_eval

1 Introduction

In this paper we conduct a comprehensive evaluation of LLMs on the Semantic Overlap Summarization (SOS) Task (Bansal et al., 2022b). Inspired by (Karmaker Santu et al., 2018), this task focuses on extracting overlapping information multiple narratives. Leveraging established metrics such as ROUGE (Lin, 2004), BERTscore (Zhang et al., 2020), and Sem-F1 (Bansal et al., 2022a), we assess the quality of generated summaries against reference summaries, providing insights into the effectiveness of these models in capturing semantically overlapping information.

Moreover, we introduce a novel dataset to serve as an additional benchmark for the SOS task, enriching the landscape for semantic overlap assessment. As a part of our evaluation framework, we devise a set of prompts utilizing the TeLER taxonomy (Karmaker Santu and Feng, 2023), which outlines categories of prompting methods for instruction-tuned LLMs.

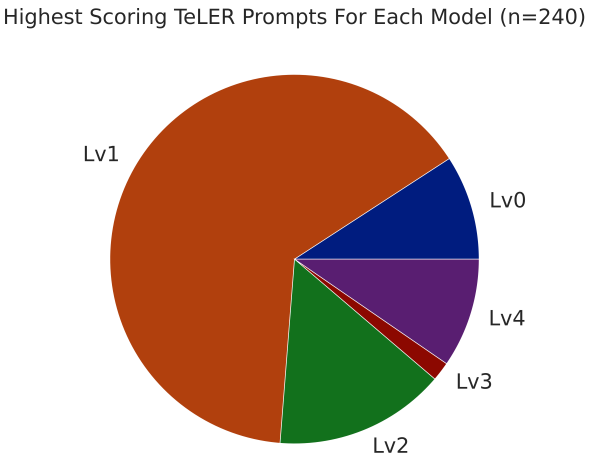

Figure 1: Count of top prompts by level for each combination of model and metric. Level 1 prompts generally outperform the other TeLER levels for these datasets.

2 Datasets

In this study we evaluate on two datasets for the SOS task: the AllSides dataset which was previously presented by (Bansal et al., 2022b) and our newly introduced PrivacyPolicy dataset.

2.1 AllSides

The AllSides dataset is collected from AllSides.com, a third-party online news forum known for presenting news and information from various political perspectives. To build the dataset, the authors crawled news articles from AllSides.com, focusing on stories covering 2,925 events. These articles were sourced from media outlets affiliated with both "Left" and "Right" political leanings, such as the New York Times and Fox News, respectively. Each news story includes a factual description labeled as "Theme" by AllSides, which serves as a neutral point of view for readers to reference.

2.2 PrivacyPolicy

The PrivacyPolicy dataset is an additional evaluation set containing 158 thoroughly validated sam-

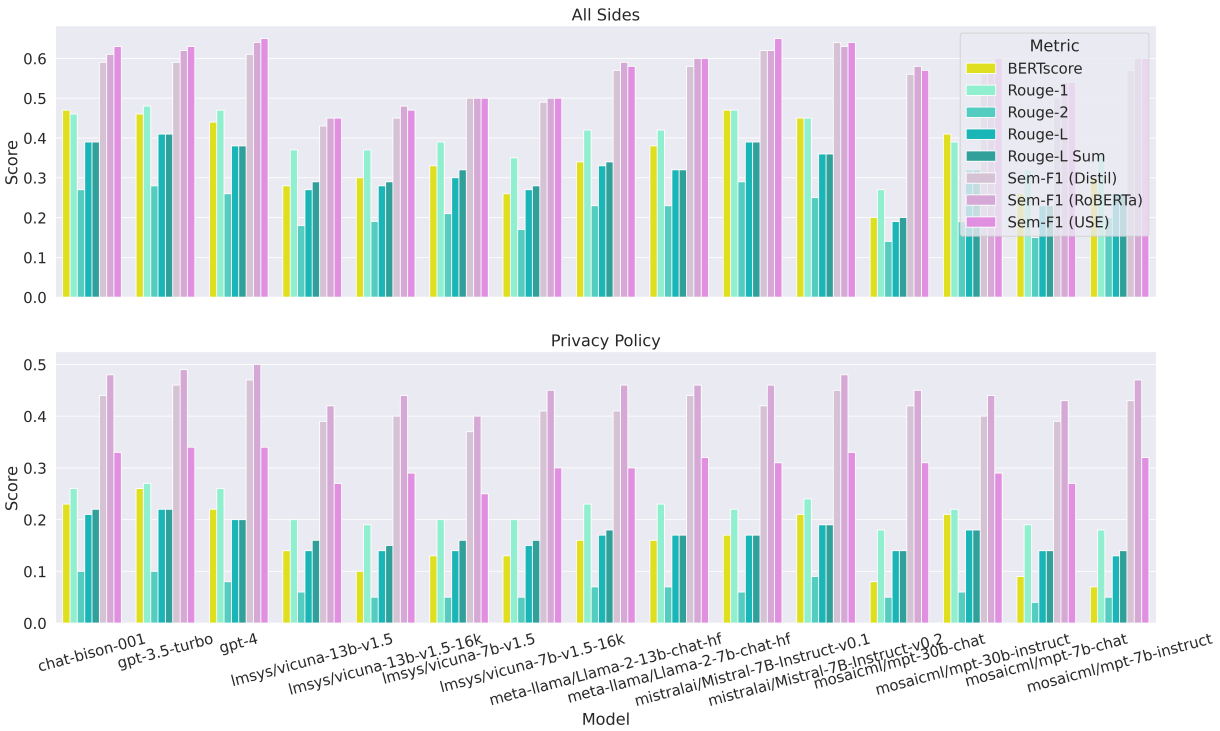

Figure 2: Best scores over each TeLER prompt level for all 15 evaluated LLMs and for each dataset. Yellow shows BERTscore, green shows ROUGE, and pink shows Sem-F1.

ples. These samples were collected by crawling the web for privacy policies posted by various companies such as Amazon or Instagram. For each sample, we pair a subsection of two different policies according to 9 categories: User Access, Edit and Deletion, Other, Data Security, User Choice/Control, Policy Change, Data Retention, International and Specific Audiences, First Party Collection/Use, Third Party Sharing/Collection. Similar to the AllSides dataset, the PrivacyPolicy dataset also features 3 reference summaries per sample.

3 Methodology

We evaluate our datasets using 6 families of LLMs, totalling 15 models. Google PaLM2 (Anil et al., 2023), OpenAI GPT-3.5-turbo and GPT-4 (OpenAI, 2023), MosaicML MPT (Team, 2023), LMSYS Vicuna (Zheng et al., 2023), MistralAI (Jiang et al., 2023), and MetaAI Llama2 (Touvron et al., 2023).

For the commercial LLMs (OpenAI and Google), we are able to use their provided APIs for inference but for open-source LLMs we use the huggingface transformers library (Wolf et al., 2020) to access model weights and evaluate on a server of 4xA4500 20GB GPUs. For additional inference speed we leverage the vLLM library (Kwon et al., 2023).

To generate our summaries, we come up with 5 sets of prompts, each comprising of one of each TeLER level (Karmaker Santu and Feng, 2023) from level 0 to level 4. After collecting our summaries, the set with the highest average scores were kept for final evaluation.

4 Results

See Figure 2 for a comprehensive breakdown of the scores achieved by each LLM for each metric. Here you can see that each model consistently scores lower on the PrivacyPolicy dataset than the AllSides dataset.

Figure 1 shows the spread of best scoring prompt levels for each pair of metric and model totalling 240

5 Conclusion

In this study we provide a comprehensive look into the capability of LLMs for the Semantic Overlap Summarization (SOS) task. To facilitate robust evaluation, we text on a previously created dataset and additionally introduce the PrivacyPolicy dataset. We leverage the TeLER prompting taxonomy to devise the a set of hand-crafted prompts that generate the highest scores we can achieve with pre-trained instruction-tuned LLMs.

References

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. (arXiv:2305.10403). ArXiv:2305.10403 [cs].

Naman Bansal, Mousumi Akter, and Shubhra Kanti Karmaker Santu. 2022a. Sem-f1: an automatic way for semantic evaluation of multi-narrative overlap summaries at scale. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 780–792, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Naman Bansal, Mousumi Akter, and Shubhra Kanti Karmaker Santu. 2022b. Semantic overlap summarization among multiple alternative narratives: An exploratory study. In *Proceedings of the 29th International Conference on Computational Linguistics*, page 6195–6207, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. er: A general taxonomy of llm prompts for benchmarking complex tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 14197–14203, Singapore. Association for Computational Linguistics.

Shubhra Kanti Karmaker Santu, Chase Geigle, Duncan Ferguson, William Cope, Mary Kalantzis, Duane Searsmith, and Chengxiang Zhai. 2018. Sofsat: Towards a setlike operator based framework for semantic analysis of text. *ACM SIGKDD Explorations Newsletter*, 20(2):21–30.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. (arXiv:2309.06180). ArXiv:2309.06180 [cs].

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. (arXiv:2303.08774). ArXiv:2303.08774 [cs].

MosaicML NLP Team. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2024-01-30.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. (arXiv:2307.09288). ArXiv:2307.09288 [cs].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von

Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. (arXiv:1910.03771). ArXiv:1910.03771 [cs].

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. (arXiv:1904.09675). ArXiv:1904.09675 [cs].

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. (arXiv:2306.05685). ArXiv:2306.05685 [cs].

A Extra Figures

A.1 All Data

A.2 Metric Correlation

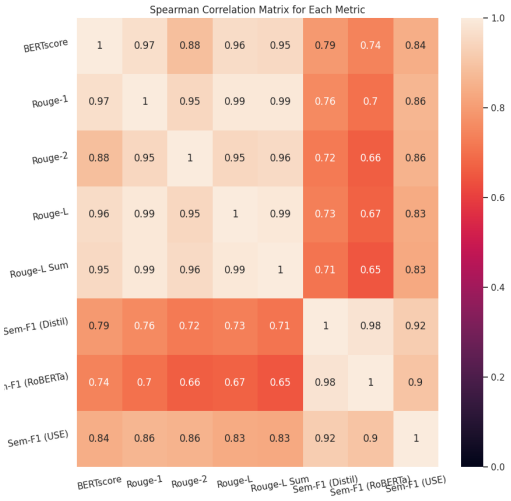

Figure 3: Correlation scores between all evaluation metrics.

A.3 Average Score per Level

Model	Dataset	Level	# Samples	Sem-F1 (USE)	Sem-F1 (Distil)	Sem-F1 (RoBERTa)	Rouge-1	Rouge-2	Rouge-L	Rouge-L Sum	BERTscore
lmsys/vicuna-13b-v1.5-16k	all_sides	0	137	0.25	0.3	0.3	0.16	0.07	0.12	0.13	-0.29
lmsys/vicuna-13b-v1.5-16k	all_sides	1	137	0.17	0.23	0.23	0.14	0.07	0.12	0.12	-0.36
lmsys/vicuna-13b-v1.5-16k	all_sides	2	137	0.18	0.24	0.23	0.15	0.07	0.12	0.13	-0.33
lmsys/vicuna-13b-v1.5-16k	all_sides	3	137	0.42	0.41	0.43	0.36	0.17	0.26	0.28	0.28
lmsys/vicuna-13b-v1.5-16k	all_sides	4	137	0.47	0.45	0.48	0.37	0.19	0.28	0.29	0.3
lmsys/vicuna-13b-v1.5	all_sides	0	137	0.24	0.29	0.28	0.17	0.08	0.13	0.14	-0.29
lmsys/vicuna-13b-v1.5	all_sides	1	137	0.2	0.26	0.25	0.16	0.08	0.13	0.13	-0.3
lmsys/vicuna-13b-v1.5	all_sides	2	137	0.42	0.41	0.42	0.36	0.17	0.26	0.28	0.28
lmsys/vicuna-13b-v1.5	all_sides	3	137	0.42	0.41	0.43	0.35	0.16	0.26	0.27	0.27
lmsys/vicuna-13b-v1.5	all_sides	4	137	0.45	0.43	0.45	0.37	0.18	0.27	0.29	0.28
lmsys/vicuna-7b-v1.5-16k	all_sides	0	137	0.38	0.39	0.4	0.24	0.12	0.17	0.19	-0.04
lmsys/vicuna-7b-v1.5-16k	all_sides	1	137	0.5	0.49	0.5	0.34	0.17	0.26	0.26	0.24
lmsys/vicuna-7b-v1.5-16k	all_sides	2	137	0.43	0.42	0.44	0.35	0.17	0.27	0.28	0.25
lmsys/vicuna-7b-v1.5-16k	all_sides	3	137	0.38	0.38	0.39	0.34	0.16	0.25	0.27	0.26
lmsys/vicuna-7b-v1.5-16k	all_sides	4	137	0.36	0.35	0.37	0.29	0.14	0.22	0.23	0.2
lmsys/vicuna-7b-v1.5	all_sides	0	137	0.32	0.33	0.34	0.18	0.08	0.13	0.15	-0.16
lmsys/vicuna-7b-v1.5	all_sides	1	137	0.5	0.5	0.5	0.35	0.19	0.26	0.27	0.25
lmsys/vicuna-7b-v1.5	all_sides	2	137	0.44	0.43	0.45	0.39	0.21	0.3	0.32	0.33
lmsys/vicuna-7b-v1.5	all_sides	3	137	0.4	0.39	0.41	0.34	0.17	0.26	0.27	0.28
lmsys/vicuna-7b-v1.5	all_sides	4	137	0.35	0.35	0.37	0.27	0.13	0.2	0.21	0.19
meta-llama/Llama-2-13b-chat-hf	all_sides	0	137	0.58	0.57	0.59	0.26	0.13	0.17	0.2	0.21
meta-llama/Llama-2-13b-chat-hf	all_sides	1	137	0.44	0.44	0.46	0.42	0.23	0.33	0.34	0.34
meta-llama/Llama-2-13b-chat-hf	all_sides	2	137	0.54	0.49	0.52	0.35	0.16	0.25	0.26	0.29
meta-llama/Llama-2-13b-chat-hf	all_sides	3	137	0.46	0.4	0.46	0.32	0.14	0.22	0.25	0.13
meta-llama/Llama-2-13b-chat-hf	all_sides	4	137	0.5	0.46	0.51	0.32	0.14	0.22	0.24	0.2
meta-llama/Llama-2-7b-chat-hf	all_sides	0	137	0.59	0.58	0.6	0.26	0.13	0.17	0.19	0.21
meta-llama/Llama-2-7b-chat-hf	all_sides	1	137	0.6	0.57	0.59	0.42	0.23	0.32	0.32	0.38
meta-llama/Llama-2-7b-chat-hf	all_sides	2	137	0.54	0.48	0.52	0.35	0.16	0.25	0.26	0.29
meta-llama/Llama-2-7b-chat-hf	all_sides	3	137	0.5	0.45	0.49	0.33	0.14	0.24	0.24	0.26
meta-llama/Llama-2-7b-chat-hf	all_sides	4	137	0.49	0.45	0.48	0.33	0.14	0.24	0.24	0.26
mistralai/Mistral-7B-Instruct-v0.1	all_sides	0	137	0.53	0.52	0.54	0.32	0.13	0.21	0.21	0.25
mistralai/Mistral-7B-Instruct-v0.1	all_sides	1	137	0.65	0.62	0.62	0.47	0.29	0.39	0.39	0.47
mistralai/Mistral-7B-Instruct-v0.1	all_sides	2	137	0.53	0.47	0.52	0.36	0.17	0.26	0.26	0.29
mistralai/Mistral-7B-Instruct-v0.1	all_sides	3	137	0.52	0.46	0.5	0.35	0.17	0.26	0.27	0.27
mistralai/Mistral-7B-Instruct-v0.1	all_sides	4	137	0.52	0.47	0.51	0.37	0.18	0.27	0.27	0.29
mistralai/Mistral-7B-Instruct-v0.2	all_sides	0	137	0.48	0.44	0.47	0.18	0.07	0.11	0.13	0.06
mistralai/Mistral-7B-Instruct-v0.2	all_sides	1	137	0.64	0.64	0.63	0.45	0.25	0.36	0.36	0.45
mistralai/Mistral-7B-Instruct-v0.2	all_sides	2	137	0.53	0.48	0.53	0.37	0.18	0.28	0.28	0.33
mistralai/Mistral-7B-Instruct-v0.2	all_sides	3	137	0.52	0.48	0.54	0.38	0.19	0.29	0.29	0.32
mistralai/Mistral-7B-Instruct-v0.2	all_sides	4	137	0.54	0.49	0.54	0.38	0.19	0.29	0.29	0.32
mosaicml/mpt-7b-chat	all_sides	0	137	0.41	0.41	0.44	0.25	0.11	0.17	0.18	0.14
mosaicml/mpt-7b-chat	all_sides	1	137	0.54	0.5	0.54	0.32	0.15	0.23	0.23	0.26
mosaicml/mpt-7b-chat	all_sides	2	137	0.5	0.45	0.5	0.32	0.14	0.23	0.23	0.25
mosaicml/mpt-7b-chat	all_sides	3	137	0.47	0.44	0.49	0.29	0.12	0.21	0.21	0.22
mosaicml/mpt-7b-chat	all_sides	4	137	0.5	0.46	0.5	0.31	0.14	0.23	0.23	0.23
mosaicml/mpt-7b-instruct	all_sides	0	137	0.59	0.57	0.6	0.28	0.2	0.23	0.24	0.25
mosaicml/mpt-7b-instruct	all_sides	1	137	0.6	0.56	0.59	0.36	0.18	0.26	0.26	0.3
mosaicml/mpt-7b-instruct	all_sides	2	137	0.52	0.46	0.5	0.28	0.12	0.2	0.2	0.18
mosaicml/mpt-7b-instruct	all_sides	3	137	0.49	0.44	0.48	0.27	0.11	0.18	0.19	0.14
mosaicml/mpt-7b-instruct	all_sides	4	137	0.53	0.5	0.53	0.28	0.15	0.2	0.22	0.17
mosaicml/mpt-30b-chat	all_sides	0	137	0.57	0.56	0.58	0.27	0.14	0.18	0.19	0.2
mosaicml/mpt-30b-chat	all_sides	1	137	0.52	0.51	0.54	0.25	0.14	0.19	0.2	0.17
mosaicml/mpt-30b-chat	all_sides	2	137	0.46	0.44	0.47	0.22	0.09	0.15	0.16	0.11
mosaicml/mpt-30b-chat	all_sides	3	137	0.44	0.42	0.47	0.2	0.08	0.14	0.15	0.07
mosaicml/mpt-30b-chat	all_sides	4	137	0.45	0.44	0.48	0.23	0.1	0.16	0.16	0.09
mosaicml/mpt-30b-instruct	all_sides	0	137	0.59	0.56	0.58	0.35	0.19	0.26	0.27	0.31
mosaicml/mpt-30b-instruct	all_sides	1	137	0.6	0.56	0.57	0.39	0.19	0.32	0.32	0.41
mosaicml/mpt-30b-instruct	all_sides	2	137	0.44	0.43	0.43	0.29	0.13	0.24	0.24	0.22
mosaicml/mpt-30b-instruct	all_sides	3	137	0.48	0.47	0.47	0.33	0.14	0.28	0.27	0.27
mosaicml/mpt-30b-instruct	all_sides	4	137	0.47	0.44	0.47	0.33	0.14	0.25	0.25	0.26
chat-bison-001	all_sides	0	137	0.55	0.52	0.54	0.22	0.09	0.15	0.17	0.15
chat-bison-001	all_sides	1	137	0.63	0.59	0.61	0.46	0.27	0.39	0.39	0.47
chat-bison-001	all_sides	2	137	0.5	0.46	0.49	0.35	0.16	0.26	0.26	0.28
chat-bison-001	all_sides	3	137	0.49	0.43	0.49	0.33	0.15	0.24	0.25	0.21
chat-bison-001	all_sides	4	137	0.55	0.5	0.54	0.37	0.18	0.28	0.28	0.3
gpt-3.5-turbo	all_sides	0	137	0.53	0.52	0.54	0.32	0.13	0.21	0.21	0.26
gpt-3.5-turbo	all_sides	1	137	0.63	0.59	0.62	0.48	0.28	0.41	0.41	0.46
gpt-3.5-turbo	all_sides	2	137	0.51	0.47	0.51	0.36	0.18	0.28	0.28	0.3
gpt-3.5-turbo	all_sides	3	137	0.51	0.47	0.51	0.37	0.18	0.27	0.28	0.28
gpt-3.5-turbo	all_sides	4	137	0.51	0.47	0.52	0.36	0.18	0.26	0.27	0.29
gpt-4	all_sides	0	137	0.6	0.57	0.6	0.37	0.17	0.25	0.25	0.32
gpt-4	all_sides	1	137	0.65	0.61	0.64	0.47	0.26	0.38	0.38	0.44
gpt-4	all_sides	2	137	0.58	0.51	0.56	0.41	0.2	0.29	0.29	0.33
gpt-4	all_sides	3	137	0.58	0.5	0.55	0.4	0.2	0.29	0.29	0.33
gpt-4	all_sides	4	137	0.6	0.53	0.58	0.4	0.2	0.3	0.3	0.35

Table 1: All data collected on the AllSides dataset for our finalized set of prompts

Model	Dataset	Level	# Samples	Sem-F1 (USE)	Sem-F1 (Distil)	Sem-F1 (RoBERTa)	Rouge-1	Rouge-2	Rouge-L	Rouge-L Sum	BERTscore
lmsys/vicuna-13b-v1.5-16k	privacy_policy	0	137	0.17	0.26	0.27	0.06	0.02	0.05	0.05	-0.38
lmsys/vicuna-13b-v1.5-16k	privacy_policy	1	137	0.15	0.24	0.23	0.1	0.03	0.08	0.08	-0.32
lmsys/vicuna-13b-v1.5-16k	privacy_policy	2	137	0.25	0.37	0.41	0.19	0.05	0.14	0.15	0.1
lmsys/vicuna-13b-v1.5-16k	privacy_policy	3	137	0.23	0.34	0.36	0.14	0.04	0.1	0.1	-0.12
lmsys/vicuna-13b-v1.5-16k	privacy_policy	4	137	0.29	0.4	0.44	0.18	0.05	0.13	0.14	0.08
lmsys/vicuna-13b-v1.5	privacy_policy	0	137	0.2	0.29	0.3	0.07	0.02	0.05	0.06	-0.3
lmsys/vicuna-13b-v1.5	privacy_policy	1	137	0.19	0.29	0.29	0.11	0.04	0.09	0.09	-0.22
lmsys/vicuna-13b-v1.5	privacy_policy	2	137	0.25	0.37	0.41	0.2	0.06	0.14	0.16	0.14
lmsys/vicuna-13b-v1.5	privacy_policy	3	137	0.25	0.38	0.41	0.17	0.05	0.12	0.13	0.08
lmsys/vicuna-13b-v1.5	privacy_policy	4	137	0.27	0.39	0.42	0.18	0.05	0.13	0.14	0.1
lmsys/vicuna-7b-v1.5-16k	privacy_policy	0	137	0.22	0.32	0.33	0.06	0.02	0.04	0.05	-0.27
lmsys/vicuna-7b-v1.5-16k	privacy_policy	1	137	0.27	0.39	0.4	0.2	0.05	0.15	0.15	0.11
lmsys/vicuna-7b-v1.5-16k	privacy_policy	2	137	0.24	0.36	0.39	0.2	0.05	0.14	0.16	0.13
lmsys/vicuna-7b-v1.5-16k	privacy_policy	3	137	0.28	0.4	0.45	0.18	0.05	0.13	0.14	0.11
lmsys/vicuna-7b-v1.5-16k	privacy_policy	4	137	0.3	0.41	0.44	0.18	0.05	0.13	0.13	0.1
lmsys/vicuna-7b-v1.5	privacy_policy	0	137	0.21	0.31	0.31	0.06	0.02	0.04	0.05	-0.26
lmsys/vicuna-7b-v1.5	privacy_policy	1	137	0.25	0.37	0.39	0.17	0.05	0.12	0.13	0.06
lmsys/vicuna-7b-v1.5	privacy_policy	2	137	0.24	0.36	0.39	0.2	0.05	0.14	0.16	0.13
lmsys/vicuna-7b-v1.5	privacy_policy	3	137	0.24	0.37	0.4	0.17	0.04	0.13	0.14	0.09
lmsys/vicuna-7b-v1.5	privacy_policy	4	137	0.24	0.37	0.4	0.18	0.05	0.13	0.14	0.1
meta-llama/Llama-2-13b-chat-hf	privacy_policy	0	137	0.29	0.4	0.44	0.1	0.03	0.07	0.09	-0.07
meta-llama/Llama-2-13b-chat-hf	privacy_policy	1	137	0.29	0.38	0.41	0.23	0.07	0.17	0.18	0.16
meta-llama/Llama-2-13b-chat-hf	privacy_policy	2	137	0.3	0.41	0.46	0.18	0.05	0.13	0.14	0.11
meta-llama/Llama-2-13b-chat-hf	privacy_policy	3	137	0.29	0.41	0.45	0.18	0.04	0.13	0.13	0.1
meta-llama/Llama-2-13b-chat-hf	privacy_policy	4	137	0.28	0.39	0.43	0.16	0.04	0.11	0.12	0.07
meta-llama/Llama-2-7b-chat-hf	privacy_policy	0	137	0.29	0.39	0.43	0.09	0.03	0.06	0.07	-0.08
meta-llama/Llama-2-7b-chat-hf	privacy_policy	1	137	0.32	0.44	0.46	0.23	0.07	0.17	0.17	0.16
meta-llama/Llama-2-7b-chat-hf	privacy_policy	2	137	0.31	0.42	0.46	0.19	0.04	0.13	0.14	0.11
meta-llama/Llama-2-7b-chat-hf	privacy_policy	3	137	0.31	0.41	0.46	0.18	0.04	0.13	0.13	0.11
meta-llama/Llama-2-7b-chat-hf	privacy_policy	4	137	0.3	0.42	0.44	0.18	0.05	0.14	0.14	0.11
mistralai/Mistral-7B-Instruct-v0.1	privacy_policy	0	137	0.3	0.4	0.45	0.14	0.04	0.1	0.1	0.02
mistralai/Mistral-7B-Instruct-v0.1	privacy_policy	1	137	0.28	0.4	0.41	0.22	0.06	0.17	0.17	0.17
mistralai/Mistral-7B-Instruct-v0.1	privacy_policy	2	137	0.31	0.42	0.46	0.2	0.05	0.14	0.14	0.12
mistralai/Mistral-7B-Instruct-v0.1	privacy_policy	3	137	0.3	0.42	0.45	0.19	0.05	0.14	0.14	0.12
mistralai/Mistral-7B-Instruct-v0.1	privacy_policy	4	137	0.31	0.42	0.46	0.2	0.05	0.14	0.14	0.12
mistralai/Mistral-7B-Instruct-v0.2	privacy_policy	0	137	0.24	0.37	0.4	0.05	0.01	0.03	0.04	-0.17
mistralai/Mistral-7B-Instruct-v0.2	privacy_policy	1	137	0.33	0.45	0.48	0.24	0.09	0.19	0.19	0.21
mistralai/Mistral-7B-Instruct-v0.2	privacy_policy	2	137	0.31	0.41	0.47	0.21	0.05	0.15	0.15	0.15
mistralai/Mistral-7B-Instruct-v0.2	privacy_policy	3	137	0.31	0.42	0.47	0.21	0.05	0.14	0.15	0.14
mistralai/Mistral-7B-Instruct-v0.2	privacy_policy	4	137	0.31	0.43	0.47	0.21	0.06	0.15	0.15	0.15
mosaicml/mpt-7b-chat	privacy_policy	0	137	0.24	0.36	0.4	0.12	0.03	0.09	0.09	-0.05
mosaicml/mpt-7b-chat	privacy_policy	1	137	0.27	0.39	0.41	0.19	0.04	0.14	0.14	0.09
mosaicml/mpt-7b-chat	privacy_policy	2	137	0.26	0.39	0.43	0.18	0.03	0.12	0.13	0.08
mosaicml/mpt-7b-chat	privacy_policy	3	137	0.27	0.39	0.42	0.17	0.04	0.12	0.13	0.08
mosaicml/mpt-7b-chat	privacy_policy	4	137	0.27	0.39	0.43	0.17	0.04	0.12	0.13	0.06
mosaicml/mpt-7b-instruct	privacy_policy	0	137	0.32	0.43	0.47	0.13	0.04	0.09	0.1	-0.03
mosaicml/mpt-7b-instruct	privacy_policy	1	137	0.28	0.38	0.41	0.18	0.05	0.13	0.14	0.07
mosaicml/mpt-7b-instruct	privacy_policy	2	137	0.27	0.38	0.41	0.15	0.04	0.11	0.12	0.03
mosaicml/mpt-7b-instruct	privacy_policy	3	137	0.27	0.37	0.41	0.16	0.04	0.11	0.13	0.02
mosaicml/mpt-7b-instruct	privacy_policy	4	137	0.27	0.38	0.41	0.14	0.03	0.1	0.11	0
mosaicml/mpt-30b-chat	privacy_policy	0	137	0.29	0.41	0.45	0.12	0.04	0.08	0.09	-0.01
mosaicml/mpt-30b-chat	privacy_policy	1	137	0.31	0.42	0.45	0.18	0.05	0.14	0.14	0.08
mosaicml/mpt-30b-chat	privacy_policy	2	137	0.31	0.41	0.45	0.11	0.03	0.08	0.09	-0.02
mosaicml/mpt-30b-chat	privacy_policy	3	137	0.31	0.41	0.45	0.13	0.04	0.09	0.1	0
mosaicml/mpt-30b-chat	privacy_policy	4	137	0.31	0.42	0.45	0.14	0.04	0.11	0.11	0.02
mosaicml/mpt-30b-instruct	privacy_policy	0	137	0.27	0.38	0.42	0.15	0.05	0.11	0.12	0.03
mosaicml/mpt-30b-instruct	privacy_policy	1	137	0.28	0.38	0.4	0.22	0.06	0.18	0.18	0.21
mosaicml/mpt-30b-instruct	privacy_policy	2	137	0.29	0.4	0.44	0.19	0.05	0.14	0.14	0.11
mosaicml/mpt-30b-instruct	privacy_policy	3	137	0.28	0.4	0.44	0.19	0.05	0.14	0.15	0.11
mosaicml/mpt-30b-instruct	privacy_policy	4	137	0.28	0.39	0.41	0.18	0.04	0.14	0.14	0.08
chat-bison-001	privacy_policy	0	137	0.26	0.36	0.4	0.09	0.02	0.07	0.08	-0.1
chat-bison-001	privacy_policy	1	137	0.33	0.44	0.47	0.26	0.1	0.21	0.22	0.23
chat-bison-001	privacy_policy	2	137	0.32	0.42	0.47	0.22	0.06	0.16	0.16	0.15
chat-bison-001	privacy_policy	3	137	0.33	0.44	0.48	0.2	0.06	0.15	0.16	0.14
chat-bison-001	privacy_policy	4	137	0.33	0.44	0.48	0.22	0.07	0.17	0.17	0.16
gpt-3.5-turbo	privacy_policy	0	137	0.34	0.44	0.48	0.14	0.05	0.1	0.11	0.04
gpt-3.5-turbo	privacy_policy	1	137	0.34	0.46	0.49	0.27	0.1	0.22	0.22	0.26
gpt-3.5-turbo	privacy_policy	2	137	0.33	0.44	0.49	0.21	0.06	0.15	0.15	0.15
gpt-3.5-turbo	privacy_policy	3	137	0.34	0.44	0.49	0.22	0.07	0.16	0.16	0.16
gpt-3.5-turbo	privacy_policy	4	137	0.34	0.45	0.49	0.23	0.07	0.16	0.16	0.17
gpt-4	privacy_policy	0	137	0.31	0.42	0.47	0.13	0.04	0.09	0.1	0.01
gpt-4	privacy_policy	1	137	0.34	0.47	0.49	0.26	0.08	0.2	0.2	0.22
gpt-4	privacy_policy	2	137	0.33	0.45	0.5	0.22	0.06	0.15	0.15	0.17
gpt-4	privacy_policy	3	137	0.33	0.45	0.5	0.21	0.05	0.14	0.14	0.15
gpt-4	privacy_policy	4	137	0.34	0.46	0.5	0.22	0.06	0.15	0.15	0.17

Table 2: All data collected on the PrivacyPolicy dataset using our finalized set of prompts

Dataset	Lv	BERTscore	Rouge-1	Rouge2	Rouge-L	Rouge-L Sum	Sem-F1 (Distil)	Sem-F1 (RoBERTa)	Sem-F1 (USE)
PrivacyPolicy	0	-0.108	0.101	0.031	0.071	0.080	0.369	0.401	0.263
	1	0.099	**0.204**	**0.063**	**0.157**	**0.160**	0.393	0.413	0.282
	2	**0.111**	0.190	0.049	0.135	0.143	0.401	0.443	0.288
	3	0.086	0.180	0.047	0.129	0.135	0.403	0.443	0.289
	4	0.099	0.185	0.050	0.134	0.138	**0.411**	**0.445**	**0.296**
AllSides	0	0.105	0.255	0.123	0.177	0.190	0.475	0.493	0.481
	1	**0.265**	**0.365**	**0.199**	**0.290**	**0.292**	**0.511**	**0.526**	**0.525**
	2	0.227	0.327	0.154	0.243	0.249	0.443	0.473	0.475
	3	0.239	0.331	0.152	0.243	0.252	0.437	0.474	0.472
	4	0.249	0.332	0.159	0.245	0.251	0.453	0.489	0.486

Table 3: Average scores per metric broken down by level and dataset. highest of each metric and dataset are in bold.