Benchmarking LLMs on the Semantic Overlap Summarization Task

Anonymous ACL submission

Abstract

Semantic Overlap Summarization (SOS) is a constrained multi-document summarization task, where the constraint is to capture the common/overlapping information between two alternative narratives. In this work, we perform a benchmarking study of popular Large Language Models (LLMs) exclusively on the SOS task. Additionally, we introduce the PrivacyPolicyPairs (3P) dataset to expand the space of SOS benchmarks in terms of quantity and variety. This dataset provides 135 high-quality SOS data samples sourced from privacy policy documents. We then use a standard prompting taxonomy called TELeR to create and evaluate $905,216$ distinct LLM-generated summaries over two SOS datasets from different domains, and we further conduct human evaluation on a subset of 540 samples. We conclude the paper by analyzing models' performances and the reliability of automatic evaluation[1].

1 Introduction

In the field of Natural Language Processing (NLP), Large Language Models (LLMs) have proven themselves to be the most capable text generation models in a variety of tasks and fields (Bubeck et al., 2023; Dai et al., 2022; Du et al., 2022; Smith et al., 2022; Schäfer et al., 2024; School, 2023; Thirunavukarasu et al., 2023). One task where LLMs are understudied is Semantic Overlap Summarization (SOS) (Bansal et al., 2022b; Karmaker Santu et al., 2018), where the goal is to summarize the common/overlapping information between two alternative narratives conveying similar information. Applications for this task include isolating facts from opinions in news articles, information aggregation for legal documents, extracting common issues related to a product reported in online reviews, etc. In this paper, we conduct a comprehensive benchmarking study on how LLMs perform on the SOS task using 16 popular models.

As LLMs' performance can widely vary with prompt variations (Rodriguez et al., 2023; Reynolds and McDonell, 2021), we use a standard prompting taxonomy, TELeR (Santu and Feng, 2023), to devise a comprehensive set of prompts with different degrees of detail before invoking LLMs to perform the SOS task. Our evaluation includes two different alternative narrative-pairs datasets. The first dataset is the previously introduced *AllSides* dataset released by Bansal et al. (2022b), and the second dataset is our original contribution, which was built with extensive human annotation effort, which we name as the *PrivacyPolicyPairs* (3P) dataset.

We report ROUGE, BERTscore, and SEM-F_1 on the Allsides and 3P datasets for each combination of LLMs and prompt style, totaling 905,216 distinct samples. We further collected human annotations on a subset of 540 samples to truly gauge the capabilities of LLMs in capturing overlapping information from multiple narratives. Finally, we analyze LLMs' performances and the reliability of automatic evaluation via correlation analysis against human annotations.

2 The Benchmark Datasets

2.1 The AllSides Data

The AllSides dataset is the first to be introduced for the SOS task. To build this dataset, Bansal et al. (2022b) crawled news articles from AllSides.com to create 2,788 sample training set and 137 sample test set. Each sample contains 2 source documents of left and right-leaning sources and is accompanied by a reference summary. The test set includes an additional 3 human-annotated summaries for more robust evaluation.

[1] The code and datasets used to conduct this study are available at `https://anonymous.4open.science/r/llm_eval-E16D`

2.2 The *PrivacyPolicyPairs* (3P) Data

For a more diverse evaluation, we introduce the *PrivacyPolicyPairs* (3P) dataset, focusing on the SOS task for a different domain and containing 135 human-annotated samples. Each sample comprises 2 source documents (two different privacy policy narratives), the category of passage, 3 reference summaries, company names, and word counts (example figure in the appendix). Our (3P) dataset is built on the OPP-115 Corpus introduced by Wilson et al. (2016), which comprises 115 privacy policies (267K words) spanning 15 sectors (Arts, Shopping, News, etc.). The policy data of the OPP-115 corpus are also tagged with the following categories:

- First Party Collection/Use
- Third Party Sharing/Collection
- User Choice/Control
- User Access, Edit, & Deletion
- Data Retention
- Data Security
- Policy Change
- Do Not Track
- International & Specific Audiences
- Other

These categories are associated with text spans in each document that denote where the labels were relevant. Our motivation behind introducing a new dataset for SOS evaluation is to 1) extend the amount of available testing data from just 137 samples from the AllSides evaluation set to 272 total evaluation samples with a combined total of 953 human annotations and 2) provide data from a domain different from the AllSides data.

Constructing the 3P Dataset: To build the 3P dataset, we set out to create pairs of passages from the original OPP-115 corpus. To ensure a degree of overlap, we first grouped each document into the 15 sectors that were originally assigned by Wilson et al. (2016) (Arts, Shopping, Business, News, etc.). Then, within each sector, we paired different passages according to their category labels (First Party Collection, Data Retention, etc.). This process resulted in 6110 passage pairs across all sectors.

Out of the 15 sectors, we focused on *eCommerce*, *Technology*, and *Food and Drink*. We then recruited three volunteer annotators from the department and instructed them to write a summary of common information present in each document pair. The exact instructions can be found in Appendix A.6. After the initial round of annotation, the annotators came together, discussed the differences in each of their summaries, and revised their original summaries accordingly. After revising and removing samples with no overlap, we yielded 3 annotations per passage pair for a total of 405 annotations for 135 high-quality samples.

3 Methodology

3.1 Evaluated Large Language Models

We choose to test our datasets using 7 families of instruction-tuned LLMs, totaling 16 models which are listed in Table 1. OpenAI and Google provide their own unique APIs but for open source LLMs, we used the transformers library (Wolf et al., 2020) to access model weights and run inference on a server with 4 Nvidia A4500 20GB GPUs. For additional speedup, we utilized the vLLM library (Kwon et al., 2023).

LLM Family	Model
Google Gemini (Team et al., 2024)	gemini-1.5-pro-001 (May 2024)
OpenAI (OpenAI, 2023)	gpt-3.5-turbo-0125 (May 2024)
MosaicML MPT (Team, 2023)	mosaicml/mpt-7b-chat (7B) mosaicml/mpt-30b-chat (30B) mosaicml/mpt-7b-instruct (7B) mosaicml/mpt-30b-instruct (30B)
LMSYS Vicuna (Zheng et al., 2023)	lmsys/vicuna-7b-v1.5 (7B) lmsys/vicuna-13b-v1.5 (13B) lmsys/vicuna-7b-v1.5-16k (7B) lmsys/vicuna-13b-v1.5-16k (13B)
MistralAI (Jiang et al., 2023)	mistralai/Mistral-7B-Instruct-v0.1 (7B) mistralai/Mistral-7B-Instruct-v0.2 (7B)
MetaAI Llama2 (Touvron et al., 2023)	meta-llama/Llama-2-7b-chat-hf (7B) meta-llama/Llama-2-13b-chat-hf (13B)
Microsoft Phi-3 (Abdin et al., 2024)	microsoft/Phi-3-mini-4k-instruct (3.8B) microsoft/Phi-3-mini-128k-instruct 3.8B)

Table 1: The list of models evaluated in this paper with parameter counts. We use 7 families of models, 2 of which are closed source, and 5 open source.

3.2 Prompt Design

We prompted LLMs in a zero-shot setting as these methods have gained popularity with the growing capabilities of LLMs (Sarkar et al., 2023, 2022). Specifically, we utilize the guidelines laid out by the TELeR taxonomy due to its use and reference in previous studies (Hadi et al., 2023; Li et al., 2024; Hackl et al., 2023; Eigner and Händler, 2024a,b; Rodrigues et al., 2024). For this study, we used TELeR levels 0 through 4 (5 out of the 7). To ensure comprehensive prompt engineering, we created templates for TELeR levels 0 through 4 and In-Context Learning styled prompts (Brown et al., 2020) (details in appendix A.6). For each template, we then created variations of prompts that follow their respective formats. For example, the group of TELeR L1 prompts is comprised of 8 prompts: 5 general, 3 AllSides-specific, and 3 3P-specific. Then, to construct our final set of prompts, we took all possible combinations of system roles and prompts, creating $56,576$ prompts for each of our 16 models and, thus, creating $905,216$ distinct evaluation samples in total.

3.3 Evaluation

Automatic Evaluation: We conduct automatic evaluation using 11 different metrics. For lexical overlap metrics we use **ROUGE** (Lin, 2004), **BLEU** (Papineni et al., 2002), **METEOR** (Lavie and Agarwal, 2007), **chrF** (Popović, 2015), **Translation Edit Rate** (Snover et al., 2006), and **CIDEr** (Vedantam et al., 2015). For embedding-based metrics we use **BERTscore** (Zhang et al., 2020), **SEM-F1** (Bansal et al., 2022a), **BLEURT** (Sellam et al., 2020), **MoverScore** (Zhao et al., 2019), and **Sentence Mover's Similarity** (Clark et al., 2019). See Appendix A.4 for details of each metric.

Human Evaluation: We recruited 3 human volunteer for annotation purposes. To avoid the burden of having annotators analyze 9 million samples, we reduce the number of evaluation samples by 1) evaluating a subset of data that corresponds to 15 narrative pairs (7 from AllSides and 8 from 3P) out of the 272 test set samples from AllSides and 3P, 2) evaluating only the largest/newest models from each family and 3) evaluating only the summaries that correspond to the best-performing prompts within each TELeR level. This strategy reduced the number of summary evaluations from 9M to 540 samples per annotator. The annotators scored model summaries on a scale of 0-5 based on how well they captured the overlapping information between the two documents given. After individually scoring the summaries, the annotators sat together to resolve disagreements and assign a final score to each sample, giving us 2,160 scores across all samples.

4 Results

Human Evaluation: The average annotation scores provided by humans are shown in Table 3. Out of all model families, gpt-3.5-turbo summaries were most preferred with an average score of 3.53 followed by mpt-30b-chat with 3.39 average. From the different prompt styles we tested, responses generated from TELeR L2 were most preferred with a 3.42 average.

Automatic Evaluation: We report automatic evaluation results for all metrics, all models, and all datasets in Table 2. This table shows the highest scores achieved by each model across the set of all prompts with different TELeR levels. For the AllSides dataset, the best-scoring models vary with the evaluation metric used, with some metrics yielding `phi-3-mini-128k-instruct` as the best, while others favor `gemini-pro`. For the 3P dataset, `gpt-3.5-turbo` consistently scored the best with `gemini-pro` coming in second across most metrics.

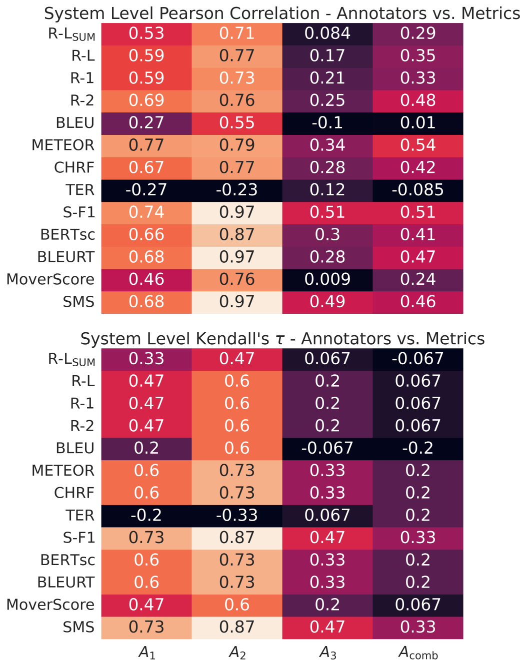

Figure 1: System-level Pearson correlation and Kendall's τ scores between annotator scores and automatic evaluation metrics (higher is better). The "comb" subscript shows the combined score where the annotators sat with each other to settle on a final score for each annotation sample.

Human Vs. Automatic Evaluation: In Figure 1, we report the System-level Kendall's τ and Pearson's ρ correlation coefficients between all our metrics and our human annotations (Chaganty et al., 2018; Novikova et al., 2017; Peyrard et al., 2017; Bhandari et al., 2020). We show the correlation scores for each individual annotator, but focus on the A_{comb} field, which represents the final score that was agreed upon by all annotators. Interestingly, while Sem-F1 was originally proposed as a specialized metric for the SOS task (Bansal et al., 2022a) and while this is indeed shown to be the case according to the Kendall's τ correlation, we can also see that it is matched by SMS and is also seen being beaten by METEOR in Pearson's ρ.

Key Findings: Our comprehensive benchmarking study provides us with the following interesting

AllSides Dataset													
Model	R-L Sum	R-L	R-1	R-2	BLEU	METEOR	chrF	TER ↓	Sem-F1	BERT score	BLEURT	Mover score	SMS
gemini-pro	0.418 (l1)	0.418 (l1)	**0.499** (l1)	**0.331** (l1)	0.003 (l0)	**0.538** (l1)	**54.634** (l1)	138.21 (l1)	**0.643** (l1)	**0.503** (l1)	**-0.144** (l1)	**0.617** (l1)	**0.617** (l1)
gpt-3.5-turbo	0.421 (l1)	0.421 (l1)	0.494 (l1)	0.300 (l1)	0.003 (icl)	0.528 (l1)	53.151 (l1)	148.21 (l1)	0.641 (l4)	0.490 (l1)	**-0.174** (l1)	**0.616** (l1)	0.612 (l1)
vicuna-13b-v1.5	0.330 (l3)	0.317 (l3)	0.426 (l2)	0.231 (l2)	0.004 (l1)	0.487 (l2)	49.272 (l2)	142.27 (l2)	0.528 (l2)	0.393 (l2)	-0.412 (l3)	0.586 (l3)	0.590 (l1)
vicuna-13b-v1.5-16k	0.326 (l2)	0.296 (l4)	0.410 (l2)	0.236 (l1)	0.003 (l1)	0.462 (l3)	47.970 (l1)	**130.22** (l1)	0.535 (l3)	0.362 (l2)	-0.440 (l3)	0.581 (l4)	0.590 (l1)
vicuna-7b-v1.5	0.355 (l2)	0.333 (l2)	0.446 (l2)	0.255 (l2)	0.004 (l1)	0.497 (l2)	50.817 (l2)	321.37 (l3)	0.549 (l4)	0.405 (l2)	-0.439 (l3)	0.590 (l2)	0.595 (l2)
vicuna-7b-v1.5-16k	0.323 (l2)	0.309 (l2)	0.419 (l2)	0.231 (l2)	0.004 (l1)	0.484 (l2)	48.843 (l2)	308.47 (l3)	0.550 (l3)	0.387 (l2)	-0.407 (l3)	0.582 (l2)	0.586 (l2)
Llama-2-13b-chat-hf	0.372 (l1)	0.357 (l1)	0.442 (l1)	0.257 (l1)	0.002 (l1)	0.495 (l4)	49.459 (l1)	236.98 (l1)	0.563 (l2)	0.369 (l1)	-0.468 (l2)	0.592 (l1)	0.584 (l1)
Llama-2-7b-chat-hf	0.336 (l3)	0.332 (l1)	0.434 (l3)	0.239 (l3)	0.002 (l1)	0.498 (l3)	49.593 (l3)	251.37 (l4)	0.603 (l2)	0.402 (l3)	-0.309 (l1)	0.589 (l1)	0.588 (l3)
Phi-3-mini-128k-instruct	**0.442** (l3)	**0.433** (l3)	**0.507** (l3)	**0.342** (l3)	0.003 (l2)	**0.541** (l1)	**54.296** (l1)	156.74 (l2)	**0.646** (l1)	0.480 (l1)	-0.179 (l1)	**0.616** (l1)	**0.623** (l1)
Phi-3-mini-4k-instruct	0.375 (l1)	0.375 (l1)	0.453 (l1)	0.255 (l1)	0.002 (l1)	0.493 (l1)	49.756 (l1)	198.04 (l1)	0.607 (l3)	0.445 (l1)	-0.188 (l1)	0.600 (l1)	0.588 (l1)
Mistral-7B-Instruct-v0.1	**0.428** (l1)	**0.428** (l1)	0.498 (l1)	0.318 (l1)	0.002 (l1)	0.539 (l1)	53.128 (l1)	190.72 (l1)	0.636 (l3)	**0.494** (l1)	-0.194 (l1)	0.614 (l1)	0.613 (l1)
Mistral-7B-Instruct-v0.2	0.374 (l1)	0.374 (l1)	0.464 (l1)	0.268 (l4)	0.002 (l0)	0.511 (l4)	51.546 (l4)	253.57 (l1)	0.637 (l1)	0.462 (l1)	-0.229 (l1)	0.601 (l1)	0.596 (l1)
mpt-30b-chat	0.340 (l1)	0.338 (l1)	0.419 (l1)	0.252 (l1)	0.001 (l2)	0.476 (l2)	47.994 (l2)	520.50 (l2)	0.596 (l1)	0.374 (l2)	-0.319 (l2)	0.588 (l1)	0.591 (l1)
mpt-30b-instruct	0.345 (l1)	0.345 (l1)	0.427 (l1)	0.237 (l2)	**0.010** (l3)	0.445 (l2)	46.618 (l2)	**112.52** (l2)	0.602 (l2)	0.435 (l1)	-0.309 (l1)	0.593 (l1)	0.588 (l2)
mpt-7b-chat	0.267 (l4)	0.263 (l3)	0.356 (l3)	0.206 (l4)	0.003 (icl)	0.434 (l4)	43.745 (l4)	327.89 (l2)	0.578 (l4)	0.304 (l3)	-0.378 (l3)	0.593 (l2)	0.585 (l4)
mpt-7b-instruct	0.278 (l1)	0.277 (l1)	0.370 (l1)	0.195 (l1)	**0.006** (l4)	0.422 (l1)	44.214 (l1)	134.32 (l4)	0.585 (l2)	0.316 (l3)	-0.378 (l3)	0.571 (l1)	0.586 (l2)
PrivacyPolicyPairs (3P) Dataset													
gemini-pro	**0.244** (l4)	**0.243** (l4)	**0.314** (l4)	**0.118** (l1)	0.003 (icl)	**0.347** (l4)	**41.843** (l4)	**150.77** (l1)	**0.528** (l4)	**0.308** (l1)	**-0.198** (l2)	**0.561** (l4)	**0.545** (l4)
gpt-3.5-turbo	**0.262** (l1)	**0.262** (l1)	**0.324** (l1)	**0.117** (l1)	0.003 (l1)	**0.355** (l1)	**41.186** (l2)	171.67 (l1)	**0.535** (l4)	**0.329** (l1)	**-0.156** (l1)	**0.567** (l1)	**0.546** (l1)
vicuna-13b-v1.5	0.196 (l2)	0.180 (l2)	0.250 (l2)	0.088 (l2)	0.002 (l2)	0.339 (l2)	37.375 (l2)	322.60 (l2)	0.445 (l3)	0.205 (l2)	-0.463 (l4)	0.552 (l3)	0.533 (l2)
vicuna-13b-v1.5-16k	0.184 (l2)	0.171 (l2)	0.239 (l2)	0.077 (l2)	0.003 (l1)	0.318 (l2)	36.181 (l2)	164.16 (l1)	0.471 (l0)	0.189 (l2)	-0.423 (l4)	0.546 (l2)	0.529 (l2)
vicuna-7b-v1.5	0.175 (l2)	0.165 (l2)	0.227 (l2)	0.071 (l2)	0.005 (l1)	0.308 (l2)	35.699 (l2)	460.12 (l1)	0.441 (l4)	0.177 (l2)	-0.501 (l1)	0.543 (l1)	0.527 (l1)
vicuna-7b-v1.5-16k	0.188 (l1)	0.186 (l1)	0.247 (l1)	0.069 (l2)	0.003 (l1)	0.303 (l3)	36.652 (l1)	375.69 (l1)	0.497 (l3)	0.204 (l1)	-0.404 (l4)	0.553 (l3)	0.533 (l3)
Llama-2-13b-chat-hf	0.207 (l1)	0.196 (l1)	0.266 (l1)	0.083 (l1)	0.001 (l1)	0.305 (l1)	38.272 (l1)	340.60 (l1)	0.466 (l3)	0.184 (l1)	-0.500 (l4)	0.545 (l1)	0.531 (l1)
Llama-2-7b-chat-hf	0.199 (l1)	0.197 (l1)	0.258 (l1)	0.079 (l1)	0.001 (l1)	0.300 (l4)	37.899 (l1)	361.54 (l1)	0.495 (l1)	0.214 (l1)	-0.383 (l1)	0.547 (l1)	0.529 (l1)
Phi-3-mini-128k-instruct	0.218 (l3)	0.217 (l3)	0.282 (l3)	0.083 (l1)	0.003 (l4)	0.308 (l1)	37.816 (l1)	187.90 (l4)	0.497 (l1)	0.276 (l1)	-0.205 (l1)	0.554 (l1)	0.533 (l1)
Phi-3-mini-4k-instruct	0.215 (l1)	0.215 (l1)	0.278 (l1)	0.083 (l1)	0.002 (l1)	0.321 (l1)	38.572 (l1)	259.86 (l1)	0.503 (l1)	0.251 (l1)	-0.345 (l1)	0.551 (l1)	0.529 (l1)
Mistral-7B-Instruct-v0.1	0.214 (l1)	0.213 (l1)	0.275 (l1)	0.083 (l1)	0.002 (l1)	0.330 (l1)	37.823 (l4)	238.45 (l1)	0.517 (l1)	0.249 (l1)	-0.362 (l2)	0.549 (l1)	0.535 (l1)
Mistral-7B-Instruct-v0.2	0.234 (l1)	0.233 (l1)	0.298 (l1)	0.106 (l1)	0.002 (l1)	0.340 (l4)	39.959 (l1)	247.36 (l1)	0.523 (l1)	0.279 (l1)	-0.291 (l1)	0.558 (l1)	0.540 (l1)
mpt-30b-chat	0.192 (l1)	0.190 (l1)	0.247 (l1)	0.075 (l1)	0.002 (l1)	0.312 (l2)	35.142 (l2)	385.01 (l2)	0.507 (l2)	0.200 (l2)	-0.347 (l2)	0.655 (icl)	0.534 (l2)
mpt-30b-instruct	0.213 (l1)	0.210 (l1)	0.267 (l1)	0.084 (l1)	**0.014** (l1)	0.297 (l2)	35.520 (l1)	**131.85** (l1)	0.487 (l2)	0.268 (l1)	-0.361 (l1)	0.667 (icl)	0.538 (l1)
mpt-7b-chat	0.177 (l2)	0.175 (l2)	0.233 (l2)	0.066 (l1)	0.003 (l0)	0.270 (l1)	33.066 (l2)	352.14 (l2)	0.479 (l2)	0.159 (l2)	-0.464 (l3)	0.651 (icl)	0.530 (l1)
mpt-7b-instruct	0.166 (l1)	0.162 (l1)	0.215 (l1)	0.075 (l1)	**0.006** (l4)	0.270 (l2)	33.105 (l1)	152.96 (l4)	0.469 (l1)	0.127 (l1)	-0.561 (l1)	0.654 (icl)	0.529 (l1)

Table 2: The best average scores for each metric over each dataset. Higher is better for all but TER which is indicated by ↓. Bold blue indicates the best score for a given metric, while the second best is indicated by bold black. Each score is accompanied by the TELeR level that was used to produce the score.

Model	Score (0-5)	Template	Score (0-5)
gemini-pro	3.37	ICL	3.08
gpt-3.5-turbo	**3.53**	L1	3.38
mpt-30b-chat	3.39	L2	**3.42**
Mistral-7B-Instruct-v0.2	3.38	L3	3.32
Phi-3-mini-128k-instruct	3.37	L4	3.32
vicuna-13b-v1.5-16k	3.32		

Table 3: Average negotiated preference score for each model and prompt template. "ICL" represents the In-Context Learning style prompts, while "Lx" refers to the level of the TELeR prompt.

insights regarding the relationships between models, evaluation metrics, TELeR Levels, and human preferences for the SOS task.

- **Models vs. TELeR Levels:** When comparing models against TELeR prompts in Table 2, we found that while TELeR L1 generally perform the best, some models show preferences towards other styles. For example, all the vicuna models show favor over L2 (64 top scores), with much fewer L1 prompts showing top scores (23).
- **Datasets vs. TELeR Levels:** Based on Table 2, L1 prompts consistently score the highest, counting 106 and 122 for AllSides and 3P, respectively. L2 comes in second place with 49 and 47, suggesting that brevity is preferred in general while designing prompts for the SOS task.
- **Human Preference Vs. TELeR Levels:** Table 3 shows that human annotators showed bias towards TELeR L2 prompts. However, the variance seems to be relatively small across L1 - L4.

5 Conclusion

In this study, we investigated the capability of LLMs for performing the Semantic Overlap Summarization (SOS) task. We evaluated LLMs on an existing dataset and additionally introduced a new dataset called the *PrivacyPolicyPairs* (3P) dataset. To account for the effects of prompt sensitivity, we adopted the TELeR prompting taxonomy to create a diverse set of prompts and found that: 1) Different TELeR levels impact each model and data set differently, suggesting that the degree of details provided in prompts must be studied and reported before making a final conclusion on LLMs' performance; 2) METEOR, SMS, and Sem-F1 are the metrics that correlate the best with human judgments at the system level; and 3) Human annotators tend to prefer summaries generated from TELeR L2, i.e., prompts with moderate details.

6 Limitations

Dataset Size: At only 135 samples, it is not feasible to train a model on just the 3P data alone. Of course the AllSides dataset exists to accompany the

3P dataset but they represent a different category of documents from the 3P dataset which is another barrier to training. However while the size of the new dataset is small, there is a large amount of time and resource that is required to build a dataset of this nature. Firstly, this dataset requires that for each sample, we find two documents that share an overlapping narrative. Second, each sample is annotated manually by 3 people which for this dataset results in 405 annotations. That is without considering the other annotations where no overlap was found. Third, there have been several instances where disagreements need to be resolved which requires further discussion among annotators. Despite these limitations it is worth noting that this work effectively doubles the amount of samples to evaluate on the SOS task when considering both AllSides data and 3P data combined, taking our initial 137 sample news article test set to a combined 272 sample evaluation set over both news articles and privacy policy documents. In the future, a larger scale effort will be needed to increase the space of data for the SOS task.

Human Annotation: Annotation work is expensive in both time and money. We recruited all our annotators from within our department and saved on money but time cost is unavoidable. To make the process easier for our volunteers we reduced the amount of annotation samples by selecting 15 samples out of all 272 test set samples between AllSides and 3P. We also only evaluated the largest/newest models from each model family and finally, we wonly evaluated summaries that correspond to the best-performing prompts within each TELeR level. It is also important to note that the annotation process was purely for scoring user preference and there is no "right" or "wrong" answers to validate.

Model Finetuning: For this work we did not perform any fine-tuning on the evaluated models. All scores were obtained using the pre-trained weights for each model. This means that it's possible for additional performance to be gained using methods like LoRA (Hu et al., 2021). However the main goal of this study was to benchmark LLMs to set new baselines for the SOS task. In that regard we believe this to be an appropriate setup.

Automatic Evaluation: In this work we show that automatic evaluation cannot yet be trusted for the SOS task. However, reporting automatic evaluation metrics is standard practice so it is important that we take precaution when using these values to draw conclusions.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. (arXiv:2404.14219). ArXiv:2404.14219 [cs].

Sanghwan Bae, Taeuk Kim, Jihoon Kim, and Sanggoo Lee. 2019. Summary level training of sentence rewriting for abstractive summarization. *arXiv preprint arXiv:1909.08752*.

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.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 9347–9359, Online. Association for Computational Linguistics.

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2280–2292.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. (arXiv:2005.14165). ArXiv:2005.14165 [cs].

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.

Arun Tejasvi Chaganty, Stephen Mussman, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evaluation. (arXiv:1807.02202). ArXiv:1807.02202 [cs].

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. *arXiv preprint arXiv:1805.11080*.

Elizabeth Clark, Asli Celikyilmaz, and Noah A. Smith. 2019. Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 2748–2760, Florence, Italy. Association for Computational Linguistics.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2022. Why can gpt learn in-context? language models secretly perform gradient descent as meta optimizers. *arXiv preprint arXiv:2212.10559*.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme Ruiz, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd Van Steenkiste, Gamaleldin Fathy Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Collier, Alexey A. Gritsenko, Vighnesh Birodkar, Cristina Nader Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetic, Dustin Tran, Thomas Kipf, Mario Lucic, Xiaohua Zhai, Daniel Keysers, Jeremiah J. Harmsen, and Neil Houlsby. 2023. Scaling vision transformers to 22 billion parameters. In *Proceedings of the 40th International Conference on Machine Learning*, page 7480–7512. PMLR.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Eva Eigner and Thorsten Händler. 2024a. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*.

Eva Eigner and Thorsten Händler. 2024b. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*.

Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. Is gpt-4 a reliable rater? evaluating consistency in gpt-4's text ratings. In *Frontiers in Education*, volume 8, page 1272229. Frontiers Media SA.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 1:1–26.

Wan-Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. *arXiv preprint arXiv:1805.06266*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Kung-Hsiang Huang, Philippe Laban, Alexander Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2024. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1:*

Long Papers), page 570–593, Mexico City, Mexico. Association for 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 [cs].

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

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 Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, page 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Omer Levy, Ido Dagan, Gabriel Stanovsky, Judith Eckle-Kohler, and Iryna Gurevych. 2016. Modeling extractive sentence intersection via subtree entailment. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, page 2891–2901, Osaka, Japan. The COLING 2016 Organizing Committee.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

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.

Linqing Liu, Yao Lu, Min Yang, Qiang Qu, Jia Zhu, and Hongyan Li. 2017. Generative adversarial network for abstractive text summarization. *arXiv preprint arXiv:1711.09357*.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. On learning to summarize with large language models as references. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 8647–8664, Mexico City, Mexico. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, page 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, page 74–84, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, page 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the*

2021 CHI Conference on Human Factors in Computing Systems, CHI EA '21, page 1–7, New York, NY, USA. Association for Computing Machinery.

Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Dragan Gašević, Geber Ramalho, and Rafael Ferreira Mello. 2024. Assessing the quality of automatic-generated short answers using gpt-4. *Computers and Education: Artificial Intelligence*, 7:100248.

Alberto D. Rodriguez, Katherine R. Dearstyne, and Jane Cleland-Huang. 2023. Prompts matter: Insights and strategies for prompt engineering in automated software traceability. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, page 455–464.

Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. Teler: 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.

Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. 2022. Exploring universal sentence encoders for zero-shot text classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, page 135–147, Online only. Association for Computational Linguistics.

Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. 2023. Zero-shot multi-label topic inference with sentence encoders and llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 16218–16233, Singapore. Association for Computational Linguistics.

Stanford Law School. 2023. Large language models as fiduciaries: A case study toward robustly communicating with artificial intelligence through legal standards.

Max Schäfer, Sarah Nadi, Aryaz Eghbali, and Frank Tip. 2024. An empirical evaluation of using large language models for automated unit test generation. *IEEE Transactions on Software Engineering*, 50(1):85–105.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 7881–7892, Online. Association for Computational Linguistics.

Nan Shao, Zefan Cai, Chonghua Liao, Yanan Zheng, Zhilin Yang, et al. 2023. Compositional task representations for large language models. In *The Eleventh International Conference on Learning Representations*.

Utkarsh Sharma and Jared Kaplan. 2022. Scaling laws from the data manifold dimension. *Journal of Machine Learning Research*, 23(9):1–34.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 4215–4233, Singapore. Association for Computational Linguistics.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Tran, Dani Yogatama, and Donald Metzler. 2023. Scaling laws vs model architectures: How does inductive bias influence scaling? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 12342–12364, Singapore. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. 2021. Scale efficiently: Insights from pretraining and finetuning transformers.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej

Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo-yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Mou-

farek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G. Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, Z. J. Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari,

Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, T. J. Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, M. K. Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirn-

schall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. (arXiv:2312.11805). ArXiv:2312.11805 [cs].

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

Kapil Thadani and Kathleen McKeown. 2011. Towards strict sentence intersection: Decoding and evaluation strategies. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, page 43–53, Portland, Oregon. Association for Computational Linguistics.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(88):1930–1940.

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 [cs].

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 1330–1340, Berlin, Germany. Association for Computational Linguistics.

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 [cs].

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. *arXiv preprint arXiv:1804.07036*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in*

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), page 563–578, Hong Kong, China. Association for Computational Linguistics.

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 [cs].

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. *arXiv preprint arXiv:2004.08795*.

Ming Zhong, Pengfei Liu, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. *arXiv preprint arXiv:1907.03491*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

A Appendix

A.1 Additional Figures

Figure 2 shows Pearson's correlation scores between all metrics on both datasets. The Pearson scores were computed using the SciPy library (Virtanen et al., 2020)

A.2 More on the 3P Dataset

In table 4, we show statistics of the 3P dataset. Figure 5 shows an example of what a sample in the 3P dataset looks like.

3P Dataset Statistics	
# Samples	135
Avg. # Words per Document	331.00
Avg. # Words per Document Pair	662.01
Avg. # Sentences per Document	14.96
Avg. # Sentences per Document Pair	28.99
Avg. # Words per Reference	22.46
Avg. # Sentences per Reference	1.75

Table 4: Dataset statistics for the 3P dataset consisting of 135 document pairs with 3 references each.

A.3 Related Work

Text Summarization: SOS is essentially a summarization task. Over the past two decades, many document summarization approaches have been investigated (Zhong et al., 2019). The two most popular among them are *extractive* approaches (Cao et al., 2018; Narayan et al., 2018; Wu and Hu, 2018; Zhong et al., 2020) and *abstractive* approaches (Bae et al., 2019; Liu et al., 2017; Nallapati et al., 2016). Some researchers have tried combining extractive and abstractive approaches (Chen and Bansal, 2018; Hsu et al., 2018; Zhang et al., 2019).

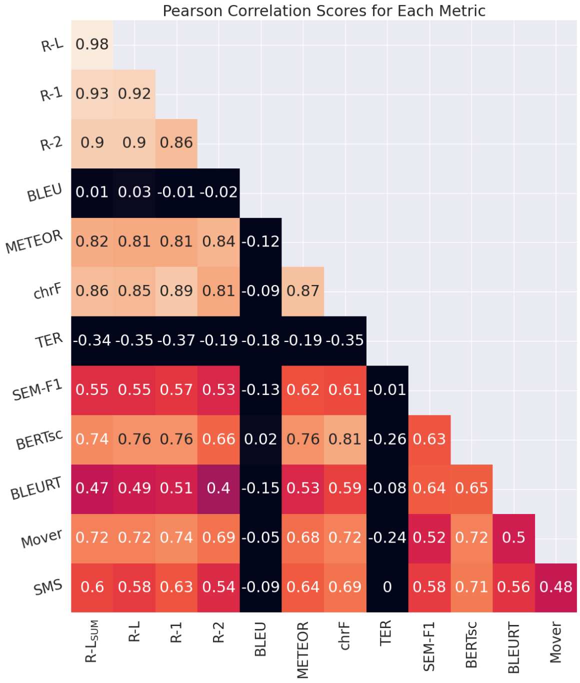

Figure 2: Raw correlation scores between all evaluation metrics.

Semantic Overlap Summarization: Semantic Overlap Summarization (SOS) is a task aimed at extracting and condensing shared information between two input documents, D_A and D_B. The output, denoted as D_O, is generated in natural language and only includes information present in both input documents. The task is framed as a constrained multi-seq-to-seq (text generation) task, where brevity is emphasized to minimize the repetition of overlapping content. The output can be extractive summaries, abstractive summaries, or a combination of both (Karmaker Santu et al., 2018). This is similar to the sentence intersection task, where your input is comprised of sentences instead of documents and your output contains only the common information (Levy et al., 2016; Thadani and McKeown, 2011).

To facilitate research in this area, Bansal et al. (2022b) introduced the AllSides dataset for training and evaluation, which we also used for evaluation in this work.

LLMs and Summarization: As the transformer architecture gained popularity, further research

3P Data Sample	
Category: Data Security	
Policy 1: Amazon (410 Words)	**Policy 2**: Lids (312 Words)
Amazon.com knows that you care how information about you is used and shared, and we appreciate your trust that we will do so carefully and sensibly ... We work to protect the security of your information during transmission by using Secure Sockets Layer (SSL) software, which encrypts information you input. We reveal only the last four digits of your credit card numbers when confirming an order. Of course, we transmit the entire credit card number to the appropriate credit card company during order processing. It is important for you to protect against unauthorized access to your password and to your computer. Be sure to sign off when finished using a shared computer. Click here for more information on how to sign off ...	Any personal information that we collect will be stored in secure servers hosted in the U.S. or Canada ... We work to protect the security of your information during transmission by using Thawte Certified Secure Sockets Layer (SSL) software, which encrypts information you input. We reveal only the last four digits of your credit card numbers when confirming an order. Of course, we transmit the entire credit card number to the appropriate credit card company during order processing. Security lies in your hands as well. It is important for you to protect against unauthorized access to your password and to your computer. Be sure to sign off when finished using a shared computer. In the event of unauthorized use of your credit card, you must notify your credit card provider in accordance with its reporting rules and procedures. ...

Reference Summaries		
A_1	A_2	A_3
We work to protect the security of your information during transmission by using Secure Sockets Layer (SSL) software, which encrypts information you input. We reveal only the last four digits of your credit card numbers when confirming an order. Of course, we transmit the entire credit card number to the appropriate credit card company during order processing. It is important for you to protect against unauthorized access to your password and to your computer. Be sure to sign off when finished using a shared computer.	Companies work to protect the security of your information during transmission by using Secure Sockets Layer (SSL) software, which encrypts information you input. They reveal only the last four digits of your credit card numbers when confirming an order. Of course, They transmit the entire credit card number to the appropriate credit card company during order processing. It is important for you to protect against unauthorized access to your password and to your computer. Hence, be sure to sign off when finished using a shared computer.	Even though the entire credit card number is transmitted, only the last 4 digits of the credit card number is visible during confirmation. SSL is used to save info during transmission. Sign off is recommended.

Table 5: A single sample from the 3P dataset. For each sample, you are given the category name, company names, the corresponding policy subsections, the count of words in each policy, and the 3 reference summaries. The highlighted text shows the overlapping information.

showed favorable behavior at scale, allowing the creation of larger and more performant models (Kaplan et al., 2020; Sharma and Kaplan, 2022; Tay et al., 2023, 2021; Dehghani et al., 2023). With the rising prevalence of these large language models, summarization naturally became one of the many areas of NLP that have progressed as a result. LLM performance has been evaluated in tasks such as news summarization (Zhang et al., 2024), multi-document summarization (Huang et al., 2024), and dialogue summarization (??) but there has also been research into using them as annotators or evaluators (Shen et al., 2023; Liu et al., 2024).

Prompt Engineering for LLMs: "Prompt Engineering" is a technique for maximizing the utility of LLMs in various tasks (Zhou et al., 2022). It involves crafting and revising the query or context to elicit the desired response or behavior from LLMs (Brown et al., 2022). Prompt engineering is an iterative process requiring multiple trial and error runs (Shao et al., 2023). In fact, differences in prompts along several key factors can significantly impact the accuracy and performance of LLMs in complex tasks. To address this issue, Santu and Feng (2023) recently proposed the TELeR taxonomy, which can serve as a unified standard for benchmarking LLMs' performances by exploring a wide variety of prompts in a structured manner.

The TELeR Taxonomy: As shown in Figure 3, the TELeR taxonomy introduced by Santu and Feng (2023) categorizes complex task prompts based on four criteria.

1. **Turn**: This refers to the number of turns or shots used while prompting an LLM to accomplish a complex task. In general, prompts can be classified as either single or multi-turn.
2. **Expression**: This refers to the style of expression for interacting with the LLM, such as questioning or instructing.
3. **Level of Details**: This dimension of prompt style deals with the granularity or depth of question or instruction. Prompts with higher levels of detail provide more granular instructions.
4. **Role**: LLMs can provide users with the option of specifying the role of the system. The response of LLM can vary due to changes in role definitions in spite of the fact that the prompt content remains unchanged.

The taxonomy outlines 7 distinct levels starting from level 0 to level 6. With each increase in level comes an increase in complexity of the prompt. In level 0, only data/context is provided with no further instruction. Level 1 extends level 0 by providing single-sentence instruction. Then level 2 extends level 1, and so on, until level 6, where all characteristics of previous levels are provided along with the additional instruction for the LLM to explain its output. For more details on the TELeR taxonomy and its applications, see Santu and Feng (2023). For convenience, we include the outline diagram from the paper in Appendix A.6.

A.4 Evaluation Metrics

SEM-F1 (Bansal et al., 2022a): Semantic F_1 computes the sentence-wise similarity (e.g., cosine similarity between two sentence embeddings) to infer the semantic overlap between a system-generated sentence and a reference sentence from both precision and recall perspectives and then, combine them into the F1 score.

BERTscore (Zhang et al., 2020): An automatic evaluation metric for text generation. Analogously to common metrics, BERTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence.

ROUGE (Lin, 2004): Recall-Oriented Understudy for Gisting Evaluation counts the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans. This metric is mainly used for evaluating text generation.

BLEURT (Sellam et al., 2020): A learned evaluation metric based on BERT that can model human judgments with a few thousand possibly biased training examples. This metric is primarily evaluating machine translation systems.

BLEU (Papineni et al., 2002): Bilingual Evaluation Understudy score is a precision-based metric that evaluates the quality of generated text by measuring n-gram overlap between the generated and reference texts. It is primarily used for machine-translation tasks.

METEOR (Lavie and Agarwal, 2007): An automatic metric for machine translation evaluation that is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations.

chrF (Popović, 2015): character n-gram F-score for automatic evaluation of machine translation output.

MoverScore (Zhao et al., 2019): Built upon a combination of contextualized representations of system and reference texts and a distance between these representations measuring the semantic distance between system outputs and references.

Sentence Mover's Similarity (Clark et al., 2019): Measures the semantic similarity between two texts by computing the minimum cost of transforming one set of sentence embeddings into another using the Earth Mover's Distance (EMD).

CIDEr (Vedantam et al., 2015): Measures the similarity between generated and reference texts by computing TF-IDF-weighted n-gram overlap, emphasizing important and distinctive words. It was originally designed for image captioning

TER (Snover et al., 2006): Measures the number of edits (insertions, deletions, substitutions, and shifts) needed to transform a generated text into a reference text, normalized by the total number of words in the reference. Lower TER scores indicate better translations, as fewer edits are required.

A.5 System Level and Summary Level Correlation

To understand the performance of automatic evaluation metrics in comparison to human evlautions we examine the correlations between the distribution of scores.

Rather than a raw correlation computation between human scores and automatic scores, the system-level and summary-level methods are the commonly used for computing correlation (Chaganty et al., 2018; Novikova et al., 2017; Peyrard et al., 2017; Bhandari et al., 2020).

We use the definition from Liu et al. (2023) to describe these methods. Given m system outputs on each of the n data samples and two different evaluation methods (human evaluations vs automatic evaluations) resulting in two n-row, m-column score matrices X and Y , the summary-level correlation is an average of samplewise correlations:

$$r_{sum}(X, Y) = \frac{\sum_i \mathcal{C}(X_i, Y_i)}{n},$$

where X_i, Y_i are the evaluation results on the i-th data sample and $\mathcal{C}$ is a function calculating a correlation coefficient (*e.g.* , the Pearson correlation coefficient). In contrast, the system-level correlation is calculated on the aggregated system scores:

$$r_{sys}(X, Y) = \mathcal{C}(\bar{X}, \bar{Y}),$$

where $\bar{X}$ and $\bar{Y}$ contain m entries which are the system scores from the two evaluation methods averaged across n data samples, *e.g.* , $\bar{X}_0 = \sum_i X_{i,0}/n$

A.6 Prompt Design

We prompted LLMs in a zero-shot setting with TELeR since zero-shot approaches to NLP tasks have gained popularity with the growing capabilities of LLMs. For example, works from Sarkar et al. (2023, 2022) explore their zero-shot use cases in topic inference and text classification. The taxonomy is best outlined by Figure 3.

For this study, we used TELeR levels 0 through 4 (5 out of the 7). We chose not to prompt using levels 5 and 6 because their use of retrieval augmented prompting does not necessarily apply to the SOS task. This is due to all relevant context being present, *i.e.*, the two source narratives are already provided as part of the prompt. Furthermore, requirement number 5 for level 6 also specifies asking the LLM to explain its own output, which would negatively affect the generated summaries during evaluation. We also experiment with in-context learning prompts (Brown et al., 2020).

In Section 3.2, we discussed having different prompt variations for TELeR levels 0 through 4 and In-Context Learning prompts. The number of variations for each group is shown in Table 6.

Template Group	For PPP	For AllSides	For Both	Total
Systm Role	2	2	6	10
TELeR L0	0	0	1	1
TELeR L1	3	3	5	11
TELeR L2	3	3	3	9
TELeR L3	3	3	2	8
TELeR L4	3	3	2	8
In-Context Learning	0	0	1	1

Table 6: The number of prompts created for each template group. The "For PPP/AllSides columns indicate how many prompts were created for that dataset only. The "For Both" column is for the prompts that could be applied to both datasets. For exact prompt details, refer to Appendix A.6 for exact prompt contents.

For each group, our templates follow these general patterns:

- **TELeR Level 0**: {**Document 1**} {**Document 2**}
- **TELeR Level 1**:
  Document 1: {**Document 1**}
  Document 2: {**Document 2**}
  Summarize the overlapping information between these two documents
- **TELeR Level 2**:
  {**TELeR Level 1 Prompt Text**}
  This information must keep in mind the 5W1H facets of the documents. Do not include any uncommon information.
- **TELeR Level 3**:
  {**TELeR Level 1 Prompt Text**}
  - This information must keep in mind the 5W1H facets of the documents.
  - Do not include uncommon information.
- **TELeR Level 4**:
  {**Level 3 Prompt Text**}.
  Your response will be evaluated against a set of reference summaries. Your score will depend on how semantically similar your response is to the reference.
- **In-context Learning**:
  Document 1: {**Example Document 1**}
  Document 2: {**Example Document 2**}
  Summary: {**Example Summary**}

  Document 1: {**Document1**}
  Document 2: {**Document2**}
  Summary:

The exact prompts are laid out in the following passage.

System Role Variations Our system role templates are made up of 2 AllSides-specific items, 2 3P specific-items and 6 for general purpose. These are written as follows

- **AllSides**
  - you will be given two news articles to read. then you will be given an instruction. follow these instructions as closely as possible
  - you will read two news articles and answer any questions about them
- **3P**
  - you are to read two privacy policies and briefly provide information according to the user's needs
  - you are to read two privacy policies and provide concise answers to the user
- **Both**
  - you are to read several documents and briefly provide information according to the user's needs
  - you are to read several documents and provide concise answers to the user
  - you will read two documents and give brief answers to user questions
  - you are a machine who is given 3 inputs: document 1, document 2, and the instructions. your output will adhere to these 3 inputs.
  - you will be given 2 documents and a set of instructions. follow the instructions as closely as possible.
  - you will be given 2 documents and a set of instructions. your response to these instructions will rely on the material covered in the 2 documents.

In-Context Learning Template: We use the following for our in-context learning template:

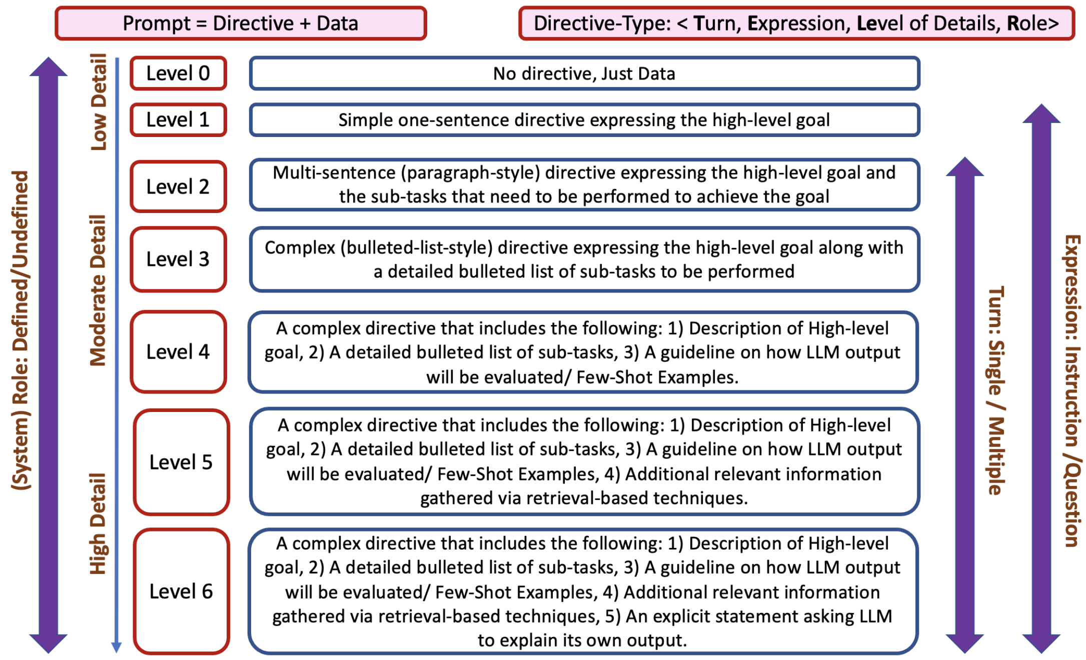

Figure 3: **TELeR** Taxonomy proposed by Santu and Feng (2023): (<**T**urn, **E**xpression, **Le**vel of Details, **R**ole>)

- Document 1: {{**Example Document 1**}}
  Document 2: {{**Example Document 2**}}
  Summary: {{**Example Reference**}}

  Document 1: {{**Document 1**}}
  Document 2: {{**Document 2**}}
  Summary:

TELeR Level 0 Template: With no possibility for variation, our TELeR L0 template is written as follows:

- {**Document 1**} {**Document 2**}

TELeR Level 1 Template: For our TELeR L1 templates we have 3 AllSides-only items, 3 3P-only items, and 5 general-purpose items.

- **AllSides**
  - Document 1: {{**Document 1**}}
    Document 2: {{**Document 2**}}

    In one sentence, please tell me the overlapping information between article 1 and article 2
  - Document 1: {{**Document 1**}}
    Document 2: {{**Document 2**}}

    summarize the overlapping information between the articles
  - Document 1: {{**Document 1**}}
    Document 2: {{**Document 2**}}

    output the overlapping information of the events covered in these articles
- **3P**
  - Policy 1: {{**Document 1**}}
    Policy 2: {{**Document 2**}}

    In one sentence, please tell me the overlapping information between policy 1 and policy 2
  - Policy 1: {{**Document 1**}}
    Policy 2: {{**Document 2**}}

    summarize the information that the two policies share
  - Policy 1: {{**Document 1**}}
    Policy 2: {{**Document 2**}}

    what is the shared information between the two policies
- **Both**
  - Document 1: {{**Document 1**}}
    Document 2: {{**Document 2**}}

    In one sentence, please tell me the overlapping information between Document 1 and Document 2
  - Document 1: {{**Document 1**}}
    Document 2: {{**Document 2**}}

    summarize the overlapping information between the documents.
  - Document 1: {{**Document 1**}}
    Document 2: {{**Document 2**}}

    output the overlapping information between the documents.
  - Document 1: {{**Document 1**}}
    Document 2: {{**Document 2**}}

    output the common information between the documents.
  - Document 1: {{**Document 1**}}
    Document 2: {{**Document 2**}}

    output only the overlapping information

TELeR Level 2 Variations: For our TELeR L2 templates we have 3 AllSides-only items, 3 3P-only items, and 3 general-purpose items.

- **AllSides**
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    these articles share similarities. output the information that is shared between them. keep your output short. to be as accurate as possible, cover the "who, what, when, where, and why of the shared information.
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    who or what are the common subjects of the two documents? what events are common between the documents? do the documents mention any locations that are the same between the two? give your response in a single sentence.
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    summarize the overlap
- **3P**
  - Policy 1: **{{Document 1}}**
    Policy 2: **{{Document 2}}**

    These policies are categorized under "Category". Describe the common aspects of these two policies in terms of this category. make sure to include the shared entities, actions and scope of the documents. Do not make any mention of information that is not shared between them. Keep your response short
  - Policy 1: **{{Document 1}}**
    Policy 2: **{{Document 2}}**

    These policies are categorized under "Category". Describe the common aspects of these two policies in terms of this category. make sure to include the shared entities, actions and scope of the documents. Do not make any mention of information that is not shared between them. give your response in a single sentence.
  - Policy 1: **{{Document 1}}**
    Policy 2: **{{Document 2}}**

    These privacy policy excerpts are tagged with the category: "Category". summarize the overlapping information between the documents. to be as accurate as possible, cover the who, what, when, where, and why of the common information.
- **Both**
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    summarize the overlapping information between the two documents. explain the who, what, when, where, and why to give full context.
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    summarize the overlapping information between the two documents. explain the who, what, when, where, and why to give full context. the output should be two sentences at most.
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    output the shared information between the documents. do not include any information outside of the shared information. keep your response short.

TELeR Level 3 Variations: For our TELeR L3 templates we have 3 AllSides-only items, 3 3P-only items, and 2 general-purpose items.

- **AllSides**
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    please answer the following:
    - who or what are the common subjects of the two documents
    - what events are common between the documents
    - do the documents mention any locations that are the same between the two
    - keep your response brief. 2 sentences max.
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    Consider the following questions and respond in a single sentence:
    - who or what are the common subjects of the two documents
    - what events are common between the documents
    - do the documents mention any locations that are the same between the two
- **3P**
  - Policy 1: **{{Document 1}}**
    Policy 2: **{{Document 2}}**

    These policies are categorized under "Category". With this in mind, please answer the following:
    - Describe the common aspects of these two policies in terms of this category.
    - make sure to include the shared entities, actions and scope of the documents.
    - Do not make any mention of information that is not shared between them.
    - Do not respond in a list format and instead respond normally.
    - Keep your response to 3 sentences at most
  - Policy 1: **{{Document 1}}**
    Policy 2: **{{Document 2}}**

    These policies are labelled under the "Category" category. With this in mind, use a single sentence that answers the following:
    - Describe the common aspects of these two policies in terms of this category.
    - make sure to include the shared entities, actions and scope of the documents.
    - Do not make any mention of information that is not shared between them.
    - Do not respond in a list format and instead respond normally.
  - Policy 1: **{{Document 1}}**
    Policy 2: **{{Document 2}}**

    These policies are labelled under the "Category" category. With this in mind, use a single sentence that answers the following:

- summarize the information that is shared between the policies
- cover the who, what, when, where, and why of the common information
- respond in as few sentences as possible

- **Both**
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    please answer the following:
    - who or what are the common subjects of the two documents
    - what events are common between the documents
    - do the documents mention any locations that are the same between the two
    - keep your response brief. 2 sentences max.
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    Consider the following questions and respond in a single sentence:
    - who or what are the common subjects of the two documents
    - what events are common between the documents
    - do the documents mention any locations that are the same between the two

TELeR Level 4 Variations For our TELeR L4 templates we have 3 AllSides-only items, 3 3P-only items, and 2 general-purpose items.

- **AllSides**
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    your goal is to describe all the common information between the given documents. to accomplish this you will need to answer the following:
    - who or what are the common subjects of the two documents
    - what events are common between the documents
    - do the documents mention any locations that are the same between the two
    - keep your response brief. 2 sentences max.

    For Example:
    Doc1: i have a dog. it's pretty fast.
    Doc2: i have a dog. he is a slow runner
    Reference Summary: i have a dog.
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    your goal is to describe all the common information between the given documents. to accomplish this you will need to answer the following:
    - who or what are the common subjects of the two documents
    - what events are common between the documents
    - do the documents mention any locations that are the same between the two

    your response will be evaluated according to how similar it is to a "reference summary".
    Example:
    Question: what is common between the sentence "the dog is slow" and "the dog is fast"
    Reference Summary: Both sentences talk about the speed of a dog
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    your goal is to describe all the common information between the given documents in one sentence. your single-sentence response will need to capture the following:
    - the common events
    - common people
    - common locations
    - the overlapping narrative of the documents

    your response will be evaluated according to how similar it is to a "reference summary".
    Example:
    Doc1: the dog is slow
    Doc2: the dog is fast
    Reference Summary: Both sentences talk about the speed of a dog
- **3P**
  - Policy 1: **{{Document 1}}**
    Policy 2: **{{Document 2}}**

    your goal is to describe all the common information between the given privacy policies. to accomplish this you will need to answer according to the following:
    - Describe the common aspects of these two policies in terms of this category.
    - make sure to include the shared entities, actions and scope of the documents.
    - Do not make any mention of information that is not shared between them.
    - Do not respond in a list format and instead respond normally.
    - Keep your response to 3 sentences at most

    your response will be evaluated according to how similar it is to a "reference summary".
    For example, an output of "cat" could be compared to "light" to get a score of 0 but that same output could be compared to "cat" to receive a score of 100. These reference summaries are usually quite short so it is important to keep your response to 3 sentences or less.

    your response will be evaluated according to how similar it is to a "reference summary". Example:
    Doc1: the dog is slow
    Doc2: the dog is fast
    Reference Summary: Both sentences talk about the speed of a dog
  - Policy 1: **{{Document 1}}**
    Policy 2: **{{Document 2}}**

    your goal is to describe all the common information between the given documents in one sentence. your single-sentence response will need to include the following:
    - common aspects related to the given category
    - common entities
    - common applications

    your response will be evaluated according to how similar it is to a "reference summary".

    Example Documents:
    Doc1: the dog is slow
    Doc2: the dog is fast

Example Response:
Both sentences talk about the speed of a dog

- Policy 1: **{{Document 1}}**
  Policy 2: **{{Document 2}}**

  your goal is to describe all the common information between the given documents in one sentence. your single-sentence response will need to include the following:
  - common aspects related to the given category
  - common entities
  - common applications

  your response will be evaluated according to how similar it is to a "reference summary".

  Example Documents:
  Doc1: the dog is slow
  Doc2: the dog is fast

  Example Response:
  Both sentences talk about the speed of a dog

- **Both**
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    Write a summary of the given documents that follows these instructions:
    - who or what are the common subjects of the two documents
    - what events are common between the documents
    - do the documents mention any locations that are the same between the two
    - keep your response brief. 2 sentences max.

    your response will be evaluated according to how similar it is to a "reference summary".
    For Example:
    Doc1: i have a dog. it's pretty fast.
    Doc2: i have a dog. he is a slow runner
    Reference Summary: i have a dog.
  - Document 1: **{{Document 1}}**
    Document 2: **{{Document 2}}**

    Summarize the overlapping information between these documents. your summary should follow these instructions:
    - exclude any information that is similar but differing or contradictory
    - write the summary as if you were summarizing a single document.
    - your summary should be short. keep it within 2 sentences.

    your response will be evaluated according to how similar it is to a "reference summary".
    For Example:
    Doc1: i have a dog. it's pretty fast.
    Doc2: i have a dog. he is a slow runner
    Reference Summary: i have a dog.

A.7 Annotation Details

3P Dataset Annotations When constructing the 3P dataset, annotators were instructed as follows:

> 1) You are given a list of document pairs. For each document pair, read and understand the overlapping information between doc1 and doc2.
>
> 2) Write a summary that only includes the overlapping information you have identified.
>
> What is overlapping information? Any information, statement, or fact that is shared between two or more documents example: 'John doe is on a trip to Las Vegas' and 'John Doe went to see the fight in Vegas' shares the information 'John Doe is in Las Vegas'
>
> What DOES NOT qualify as overlapping information: shared mentioning of names example: 'John Doe is a pilot ' and 'John Doe has never been to Canada' does not have any overlapping information

Model Summary Annotations As covered in Section 3.3, we chose our human evaluation samples by 1) evaluating a subset of data that correspond to 15 samples (7 from AllSides and 8 from 3P) out of the 272 test set samples between AllSides and 3P), 2) evaluating only the largest/newest models from each model family, and 3) evaluating only the summaries that correspond to the best performing prompts within each TELeR level. To clarify point 3, each TELeR level has a set of templates, as shown in Table 6. TELeR L1, for example, has 8 prompt and 8 system role templates that can be used to prompt the models on the AllSides dataset. All possible combinations for TELeR L1 prompt and system role templates give us 64 unique prompts to be applied to the entire dataset. After collecting responses and evaluating the average performance for each of the 64 unique prompts, the samples associated with the prompt that yielded the best performance over the AllSides dataset were chosen for human annotation.

When evaluating the summaries generated by the LLMs, annotators were instructed as follows:

> 1) You are given a list of document pairs. For each document pair, read and understand the overlapping information between doc1 and doc2.
>
> 3) Read each of the corresponding 'response' entries and assign a score between 0 and 5 (decimal values indluded)

based on how well you think it covers the overlapping information * decimal values such as 1.23 are acceptable scores.

What is overlapping information? Any information, statement, or fact that is shared between two or more documents example: 'John doe is on a trip to Las Vegas' and 'John Doe went to see the fight in Vegas' shares the information 'John Doe is in Las Vegas'

What DOES NOT qualify as overlapping information: shared mentioning of names example: 'John Doe is a pilot ' and 'John Doe has never been to Canada' does not have any overlapping information

A.8 Additional Results

Human Preference on Model and Template: While Table 7 shows that the automatic evaluations tend to have a preference towards TELeR L1 prompts, Table 3 shows that human annotators actually tend to prefer TELeR L2 prompts instead. However, this preference is only 0.04 points ahead of the next best. The table also indicates the annotators' preference towards `gpt-3.5-turbo` for the commercial LLMs. Then, for the open-source LLMs, `mpt-30b-chat` was the most preferred, with an average annotator score of 3.39. However, it is important to note that `Phi-3-mini-128k-instruct` and `Mistral-7B-Instruct-v0.2` match and beat `gemini-pro`, respectively, according to humans.

Dataset	Tmplt.	R-L Sum	R-L	R-1	R-2	BLEU	METEOR	chrF	TER ↓	S-F1	BERTsc	BLEURT	MoverScore	SMS
AllSides	L0	0.212	0.192	0.279	0.135	0.0009	0.337	36.115	1353.976	0.476	0.173	-0.637	0.548	0.546
	L1	**0.276**	0.258	0.356	**0.188**	0.0010	**0.407**	42.538	833.364	**0.524**	0.281	**-0.474**	0.568	0.561
	L2	0.257	0.243	0.339	0.170	0.0010	0.386	40.701	827.023	0.516	0.240	-0.558	0.562	0.549
	L3	0.273	**0.263**	**0.358**	0.175	0.0012	0.406	**42.696**	590.499	0.499	**0.297**	-0.505	**0.569**	**0.565**
	L4	0.259	0.250	0.335	0.162	**0.0015**	0.372	39.775	**514.080**	0.457	0.244	-0.646	0.561	0.548
	ICL	0.214	0.202	0.286	0.129	0.0010	0.342	36.837	942.628	0.423	0.179	-0.768	0.543	0.542
Privacy Policy Pairs (3P)	L0	0.109	0.096	0.134	0.042	0.0008	0.218	22.929	2243.971	0.412	-0.004	-0.682	0.520	0.510
	L1	**0.157**	0.147	**0.199**	**0.062**	0.0011	**0.265**	30.684	1057.247	**0.440**	**0.116**	**-0.545**	0.534	**0.518**
	L2	0.145	0.136	0.188	0.053	0.0008	0.254	29.823	1130.120	0.441	0.085	-0.605	0.531	0.515
	L3	0.151	0.145	**0.199**	0.048	0.0011	0.248	31.943	700.396	0.413	0.112	-0.599	0.532	0.513
	L4	0.152	**0.148**	**0.199**	0.049	**0.0015**	0.237	**30.729**	**590.374**	0.393	0.104	-0.661	0.529	0.505
	ICL	0.120	0.112	0.155	0.042	0.0010	0.219	25.154	1198.308	0.389	0.059	-0.715	**0.561**	0.477

Table 7: Average scores per metric broken down by level and dataset. Higher is better for all metrics except TER which is denoted by the ↓. TELeR Levels are denoted by "Lx" and In-Context Learning is denoted by "ICL". The best of each metric and dataset are in bold.