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. While recent advancements in Large Language Models (LLMs) have achieved exceptional performance in numerous summarization tasks, a benchmarking study of the SOS task using LLMs is yet to be performed. As LLMs' responses are highly sensitive to variations in prompt design, a major challenge in conducting such a benchmarking study is to systematically explore a variety of prompts before drawing a reliable conclusion. Fortunately, the TELeR taxonomy has been recently proposed, which can be used to design and explore various prompts for LLMs. Using this TELeR taxonomy, this paper comprehensively evaluates 16 popular LLMs on the SOS Task. We evaluate and report on 905,216 LLM generated summaries using well-established metrics like ROUGE, BERTscore, and SEM-F_1 on two different datasets of alternative narratives and we also conduct human evaluation on 540 of those summaries for further analysis. We conclude the paper by analyzing the strengths and limitations of various LLMs in terms of their capabilities in capturing overlapping information[1].

1 Introduction

Large Language Models (LLMs) represent a groundbreaking advancement in the research landscape of Natural Language Processing (NLP) and Artificial Intelligence (AI). Trained on large bodies of text data, LLMs excel in generating coherent and human-like text. These models have been evaluated in a wide range of NLP tasks (Bubeck et al., 2023; Dai et al., 2022; Du et al., 2022; Smith et al., 2022) across several areas, including software development, law, and medicine (Schäfer et al., 2024; School, 2023; Thirunavukarasu et al., 2023). However, there are still areas and tasks where LLMs are yet to be rigorously evaluated. One such task is Semantic Overlap Summarization (SOS) (Bansal et al., 2022c; Karmaker Santu et al., 2018), where the goal is to summarize the common/overlapping information between two alternative narratives.

In this paper, we conduct a comprehensive benchmarking study of the SOS task using 16 popular LLMs. Conducting such a benchmarking study is challenging because of the large variations in LLMs' performance when different prompt types/styles are used and different degrees of detail are provided in the prompts. Indeed, Rodriguez et al. (2023) shows varying performance on the CM1 dataset (Hayes et al., 2006) across many different prompts with F1-scores ranging from 0.21 to 0.54, exhibiting a 0.33 point difference. To address this issue, Santu and Feng (2023) recently proposed a general taxonomy that can be used to design diverse prompts with specific properties in order to perform a wide range of complex tasks. Using this TELeR taxonomy, we devised a comprehensive set of prompts with different degrees of detail to perform the SOS task on two different alternative narratives datasets. One dataset is the previously introduced AllSides dataset released by Bansal et al. (2022c), and the second one is our original contribution with extensive human annotation efforts, which we name the *PrivacyPolicyPairs* (3P) dataset.

Figure 1 illustrates an example of the SOS task when comparing two alternative privacy policies in the 3P dataset, where the green text denotes the output (common information) from two input privacy policies, one from Google and one from Apple. In this case, we have two competing platforms that provide similar types of services (*e.g.* Cloud Storage or Streaming Services) and each company

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

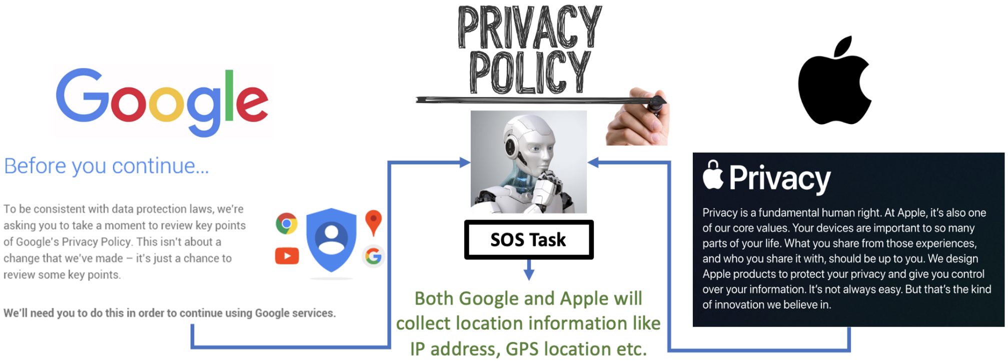

Figure 1: Example of a SOS task using two alternative privacy policy narratives.

lays out its practices when handling your private information. These documents contain important information regarding your privacy but can be long and cumbersome to read. The SOS task can help users by briefly identifying the common practices followed by each company.

For evaluation, we report well-established metrics like ROUGE, BERTscore, and SEM-F_1 on Allsides and 3P datasets for each combination of LLMs and prompt style, totaling 905,216 unique samples for analysis. We further evaluate a subset of samples using human annotators to truly gauge the capabilities of LLMs in capturing and synthesizing overlapping information from multiple narratives. We conclude the paper by analyzing the strengths and limitations of various LLMs for the same task.

2 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).

The SOS Task: *Semantic Overlap Summarization* can be framed as a multi-document summarization task, *i.e.*, multi-seq-to-seq task (Goldstein et al., 2000; Yasunaga et al., 2017; Zhao et al., 2020; Ma et al., 2020; Meena et al., 2014; Lebanoff et al., 2018; Fabbri et al., 2019). However, unlike typical multi-document summarizing tasks, *SOS* aims to summarize multiple alternative narratives with an overlapping constraint (Bansal et al., 2022c), *i.e.*, the output should only contain the common information from both input narratives (Santu et al., 2018). The availability of data for this task is relatively small so recently Bansal et al. (2022a) proposed a method for training models by utilizing synthetically generated data.

Transformers and LLMs: Encoder-decoder-based transformer models have recently gained a lot of attraction, especially for abstractive summarization tasks, (Rush et al., 2015; Chopra et al., 2016; Zhou et al., 2017; Paulus et al., 2017). Training a generic language model on a large corpus of data and then transferring/fine-tuning it for the summarization job has become a standard approach (Radford et al., 2019; Devlin et al., 2019; Lewis et al., 2019; Xiao et al., 2020; Yan et al., 2020; Zhang et al., 2019; Raffel et al., 2019). Transformer architecture features more parallelizable training, better-scaling properties, and a built-in attention mechanism, allowing large language models (LLMs) to emerge. Made up of billions of parameters, many LLMs like GPT-2 (Radford et al.), and LLaMA (Touvron et al., 2023a) have showcased strong abilities at generating text. Then, with the introduction of Reinforcement Learning From Human Feedback (Ouyang et al., 2022), LLMs became even more powerful, allowing users to interact with them as data with natural language queries.

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.

3 Background

3.1 Semantic Overlap Summarization Task

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).

Furthermore, SOS adheres to the commutative property, meaning the order of input documents doesn't affect the output summary; $D_A \cap_O D_B = D_B \cap_O D_A$. To facilitate research in this area, Bansal et al. (2022c) introduced the AllSides dataset for training and evaluation, which we also used for evaluation in this work.

3.2 Prompting With TELeR Taxonomy

As shown in Figure 6, 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.2.

4 The Benchmark Datasets

4.1 The AllSides Data

The AllSides dataset is collected from AllSides.com, a third-party online news forum known for presenting news and information from various political perspectives. Bansal et al. (2022c) crawled news articles from AllSides.com to build the dataset, focusing on narratives covering $2,925$ events. These articles provide contrasting perspectives from media outlets affiliated with "Left" and "Right" political leanings. Additionally, each event includes a factual description labeled as "Theme" by AllSides, which can serve as a neutral perspective for readers (ground truth for common facts).

The test set is comprised of 137 narrative pairs where each sample has 4 references: one from AllSides and 3 from human annotators, totaling 548 reference summaries for the test set.

4.2 The *PrivacyPolicyPairs* (3P) Data

For a more comprehensive evaluation, we introduce the *PrivacyPolicyPairs* (3P) dataset, an additional evaluation set for the SOS task, which contains 135 human annotated samples. 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

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 1: 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.

- Other

These annotations were also associated with a *text span* in the privacy policy to denote where the labels were relevant.

Our motivation behind introducing a new dataset for SOS evaluation is the following: 1) it extends 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 for the two datasets; 2) The 3P dataset represents a new type of documents in the form of semi-structured privacy policies as opposed to the news articles that make up the AllSides data; 3) News datasets are abundant, and LLMs are extensively pretrained on them in comparison to relatively infrequent privacy policy data; hence the 3P dataset is supposedly more challenging for LLMs.

Constructing the 3P Dataset: The 3P dataset includes pairs of passages taken from the OPP-115 corpus and tasks the annotator with finding the semantically overlapping information between them. A data sample is shown in Table 1. Each sample comprises 2 source documents (two alternative privacy policy narratives), the category they fall under, and 3 reference overlap summaries. The company names and word counts are also included.

When curating this dataset, we wanted to ensure each passage pair had some degree of overlap. To facilitate this goal, we reversed the process followed by the original authors and grouped the documents back into their respective sectors. Then, we built pairs of passages for each document in each sector according to the categories they were originally labeled with. This process resulted in 6110 passage pairs across all sectors.

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

Of the sectors, we chose to focus on three: *eCommerce*, *Technology*, and *Food and Drink*, due to their popularity as well as diversity among each other. From these sectors, we collected 346 passage pairs to annotate. For this task, three human annotators were asked to write a summary of common information present in each document pair. Conflicting summaries arose when there was no overlap or the annotators considered shared words as overlap. To address these issues, we retain only the policy pairs where at least two annotators wrote at least 15 words as their reference summaries. After annotating, resolving conflicts, and removing samples with no overlap, the process yielded us 3 annotations per passage pair for a total of 405 annotations for 135 high-quality samples. The final dataset statistics are listed in Table 2.

5 Methodology

5.1 Large Language Models Evaluated

We choose to test our datasets using 7 families of instruction-tuned LLMs, totaling 16 models. All evaluated models are listed in Table 3. For the commercial LLMs (OpenAI and Google), we used their provided APIs for summary generation, but for open-source LLMs, we used the huggingface transformers library (Wolf et al., 2020) to access model weights and perform generation on a server with $4XA4500$ 20GB GPUs. For additional speedup, we leveraged 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., 2023b)	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 3: The list of models evaluated in this paper. We use 7 families of models, 2 of which are closed source, and 5 open source. Parameter counts of open source models are included in parentheses. OpenAI and Google have not reported the parameter counts of their models.

We prompted LLMs in a zero-shot setting with TELeR as 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. 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 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).

5.2 Designed Prompts

For each template, we use the following outline for our prompt design.

- **TELeR Level 0**: {**Document 1**} {**Document 2**}
- **TELeR Level 1**:

  Document 1: {**Document 1 Content**}
  Document 2: {**Document 2 Content**}
  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 Doc. 1 Content**}
  Document 2: {**Example Doc. 2 Content**}
  Summary: {**Example Summary**}

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

To ensure comprehensive prompt engineering, we created groups of templates for TELeR levels 0 through 4, In-Context Learning (Brown et al., 2020) formats, and also for system roles. In each template group, we create variations of prompts that follow their respective formats. For example, the group of TELeR L1 prompts is comprised of 5 general prompts, 3 AllSides-specific prompts, and 3 3P-specific prompts. Then, to construct our final set of prompts, we took all possible combinations of system roles and prompts. A breakdown of the variation counts for each group is shown in Table 4. Using this prompting strategy, we've created 56,576 unique prompts for each of our 16 evaluated LLMs, totaling 905,216 evaluation samples. See appendix A.2 for the exact prompts that were used.

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 4: 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.2 for exact prompt contents.

5.3 Evaluation

5.3.1 Automatic Evaluation Metrics

ROUGE: ROUGE (Lin, 2004) is a family of metrics that score the lexical overlap between the generated text and the reference text. We used 3 variations, R-1, R-2, and R-L, which are widely adopted for evaluating text summarizing tasks. However, despite its popularity, works like Akter et al. (2022b) and Bansal et al. (2022b) show that ROUGE is an unsuitable metric for comparing semantics.

BERTscore: BERTscore is a metric that utilizes contextual embeddings from transformer models like BERT to evaluate the semantic similarity between the generated text and reference text. For this study, we compute BERTscore with the hashcode `roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.40.2)-rescaled`.

SEM-F1: While ROUGE and BERTscore are useful and powerful metrics, SEM-F1 was specifically designed for the SOS task. SEM-F1 leverages rigorously fine-tuned sentence encoders to evaluate the SOS task using sentence-level similarity unlike BERTscore which utilizes token-level similarity. For this study, we compute SEM-F1 with 3 underlying models: USE (Cer et al., 2018), RoBERTa (Zhuang et al., 2021), and DistilRoBERTa (Sanh et al., 2019).

5.3.2 Human Evaluation

For our human evaluation strategy, we recruited 3 volunteer annotators. These annotators evaluated 15 randomly chosen dataset samples, where 7 were picked from the AllSides dataset and 8 were picked from the 3P dataset. For each dataset sample, if we consider all prompts for every possible template combination and every model, it will amount to $3,328$ annotations required for each of the 15 dataset samples, making human judgment very time-consuming. To solve this challenge, we reduce this number by 1) choosing a subset of models to evaluate and 2) choosing the best-evaluated prompts for each of the 6 template groups based on their average performance in terms of automated metrics on both datasets. This strategy reduced the number of summaries per sample from $3,328$ to 36 summaries per sample, giving us a total of 540 annotations per human annotator. The humans were tasked to score the summaries on a scale of 0-5 based on how well they captured the overlapping information of the 2 given source documents. After individually scoring the summaries, the annotators sat together to negotiate a final score to assign to each sample, giving us 2160 annotation scores across all samples.

6 Results

For clarity, we show our results for automatic evaluation scores on the largest or newest models of each family in Figure 2. This Figure shows the highest scores achieved by each model over the set of given TELeR prompts. The first two columns are for the commercial LLMs (`GPT-3.5-Turbo` and `Gemini-Pro`), and the 5 columns on the right show our open-source LLMs. In general, we observe that `GPT-3.5-Turbo` and `Gemini-Pro` beat the open-source models across all benchmarks but we also find that `Mistral-7B-Instruct-v0.2` exhibits competitive performance across all metrics despite being only a 7-billion parameter model. For a comprehensive breakdown of the model scores achieved by each LLM and for each metric, refer to appendix (Figure 4).

Annotator/Metric Agreement In Figure 3, we show Pearson correlation and Kendall's τ correlation to compare human annotator scores with our automatic evaluation metrics. We denote the final annotation score by Ann_{comb}. From this table, all metrics have a relatively low correlation with human judgments, again demonstrating the limitations of automated metrics for evaluating text generation. Interestingly, we see that SEM-F1 best correlates with the human annotators, demonstrating its superior quality over ROUGE and BERTscore

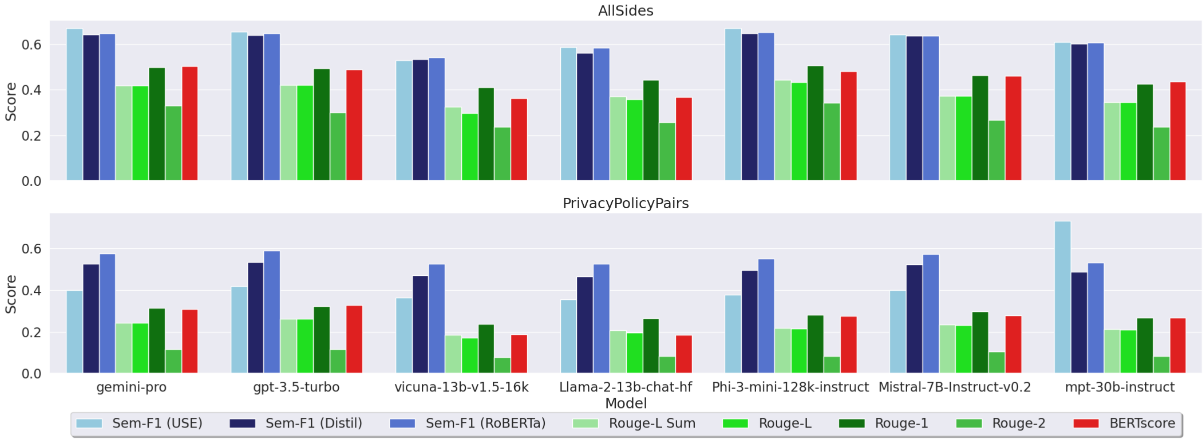

Figure 2: Best scores over each TeLER prompt level for the largest model of each family of LLMs and for each dataset. Red shows BERTscore, green shows ROUGE, and blue shows Sem-F1. A full breakdown of max scores obtained by each model is shown in Appendix A.1

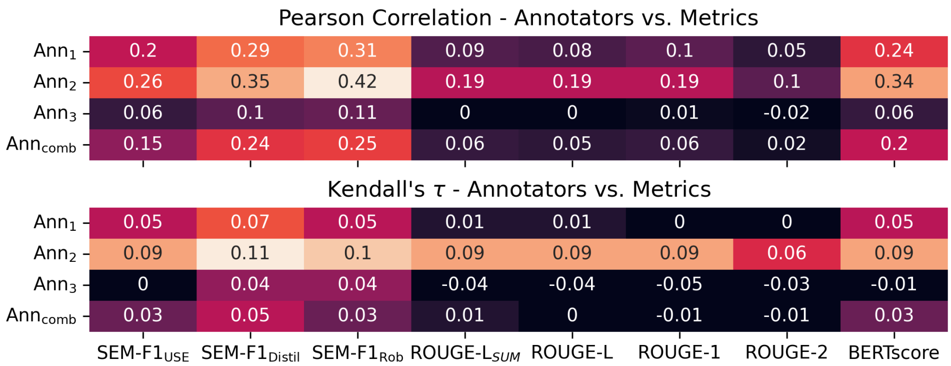

Figure 3: Pearson correlation and Kendall's τ scores between annotator scores and automatic evaluation metrics (higher is better). The "comb" subscript shows the combined score where the annotators sat with each other to settle on a final score for each annotation sample.

for multi-document summary evaluation.

AllSides Vs. 3P: In Table 5, we show average scores across all LLMs for each template prompt. The highest scores for each column are bolded. From this table, it is evident that scores on the 3P dataset tend to trail significantly behind the AllSides dataset. Additionally, we observe that for automatic evaluation, TELeR level 1 consistently performs the best across the AllSides dataset, while on the 3P dataset, the SEM-F1 scores suggest better performance in TELeR level 2 and 4.

Human Preference on Model and Template: While Table 5 shows that the automatic evaluations tend to have a preference towards TELeR L1 prompts, Table 6 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.

7 Discussion

The main takeaways from our study are as follows:

> ***Finding-1**: "3P" dataset is harder than "AllSides" for LLMs in the context of SOS task.*

To elaborate, Table 5 shows a clear difference in scores between the AllSides data and the 3P data. These differences can possibly be explained for the following reasons. The average document word count for both datasets has a significant difference but is well within the context windows of LLMs. For the AllSides data, the average is 504.51 while for the 3P data, it's 662.01. Another difference worth noting is the amount of overlapping

Dataset	Template	BERTscore	R-1	R-2	R-L Sum	R-L	Sem-F1 (Distil)	Sem-F1 (RoBERTa)	Sem-F1 (USE)
AllSides	ICL	0.453	0.46	0.267	0.367	0.367	0.621	0.639	0.651
	L0	0.391	0.399	0.209	0.315	0.291	0.6	0.618	0.614
	L1	**0.503**	**0.507**	**0.342**	**0.442**	**0.433**	**0.646**	**0.652**	**0.671**
	L2	0.437	0.466	0.278	0.388	0.369	0.631	0.636	0.653
	L3	0.461	0.465	0.273	0.377	0.376	0.639	0.646	0.658
	L4	0.477	0.467	0.268	0.376	0.376	0.641	0.647	0.656
Privacy Policy Pairs (3P)	ICL	0.262	0.278	0.092	0.219	0.219	0.499	0.55	0.371
	L0	0.138	0.226	0.067	0.187	0.164	0.509	0.567	0.399
	L1	**0.329**	**0.324**	**0.118**	**0.262**	**0.262**	**0.535**	0.588	0.419
	L2	0.267	0.307	0.109	0.254	0.234	0.531	**0.589**	0.414
	L3	0.256	0.278	0.08	0.211	0.21	0.517	0.578	0.385
	L4	0.299	0.314	0.112	0.244	0.243	**0.535**	0.577	**0.734**

Table 5: Average scores per metric broken down by level and dataset. TELeR Levels are denoted by "Lx" and In-Context Learning is denoted by "ICL". The highest of each metric and dataset are in bold.

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

Table 6: 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 TELeR prompt.

tokens present in each dataset. Utilizing the NLTK library (Bird et al., 2009) for tokenization, we find that the AllSides dataset has an average of 23.36% unique token overlap between source documents while it's 17.95% only for 3P. Aside from the numerical differences, it is also important to note the compositional differences between these datasets. The AllSides data is comprised of news articles that cover the same events. By the nature of covering the same events, it is more likely for overlapping spans of text to occur in the form of quotes or headlines. In contrast, although the documents of the 3P dataset are similar in terms of their document structure and the types of excerpts they fall under, but they are not essentially written about the same subjects. Each document pair in the 3P dataset represents two different companies whose policies can differ greatly while using similar language.

> ***Finding-2**: SEM-F1 remains the recommended evaluation metric for the SOS task.*

SEM-F1 showcases the best correlation with human annotators, while ROUGE continues to show its limitations, which is consistent with the previous findings from the literature (Akter et al., 2022a). Figure 3 supports these claims, showing correlations between annotators and ROUGE that even go into the negatives. This figure also further highlights the work that needs to be done in order to improve automatic evaluation to the point where we can rely on them more seriously and let go of expensive human evaluation.

> ***Finding-3**: Use* `gpt-3.5-turbo` *with TELeR L2 prompts for best results in the SOS task*

As of the time of writing, closed-source commercial LLMs remain the top performers in text generation. However he quality of open-source models are not too far off from `gpt-3.5-turbo` according to our human evaluators as noted in Table 6 and most of our evaluated models even compete with `gemini-pro`. Aside from model preferences it is also important to note that In-Context Learning styled prompts have been shown as a less effective prompting method compared to TELeR in the contstrained multi-document summarization setting.

8 Conclusion

In this study, we provide a comprehensive look into the capability of LLMs for the Semantic Overlap Summarization (SOS) task. To facilitate robust evaluation, we test on a previously created dataset and additionally introduce the *PrivacyPolicyPairs* (3P) dataset. We use the TELeR prompting taxonomy to devise a set of hand-crafted prompts that generate the highest scores we could achieve with pre-trained instruction-tuned LLMs and found that: 1) the 3P dataset is a harder benchmark for LLMs 2) SEM-F1 is still the best method for evaluating on SOS but far from ideal and 3) based on our testing methodology, the best summarization results in a zero-shot setting can be accomplished using `gpt-3.5-turbo` with TELeR L2 style prompts.

9 Limitations

The primary limitation of this work is the size of the dataset. At only 135 samples, it is not feasible to train a model on just the 3P data alone. 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. In the future, a larger scale effort will be needed to increase the space of data for the SOS task.

Another limitation is that we did not perform any fine-tuning on these 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.

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].

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. 2022a. Revisiting automatic evaluation of extractive summarization task: Can we do better than ROUGE? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1547–1560, Dublin, Ireland. Association for Computational Linguistics.

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker Santu. 2022b. Revisiting automatic evaluation of extractive summarization task: Can we do better than rouge? In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1547–1560. Association for Computational Linguistics.

Sanghwan Bae, Taeuk Kim, Jihoon Kim, and Sang-goo 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. Learning to generate overlap summaries through noisy synthetic data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, page 11765–11777, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Naman Bansal, Mousumi Akter, and Shubhra Kanti Karmaker Santu. 2022b. 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. 2022c. 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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, page 169–174, Brussels, Belgium. Association for Computational Linguistics.

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

Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

Jade Goldstein, Vibhu O Mittal, Jaime G Carbonell, and Mark Kantrowitz. 2000. Multi-document summarization by sentence extraction. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*.

J.H. Hayes, A. Dekhtyar, and S.K. Sundaram. 2006. Advancing candidate link generation for requirements tracing: the study of methods. *IEEE Transactions on Software Engineering*, 32(1):4–19.

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.

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].

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.

Logan Lebanoff, Kaiqiang Song, and Fei Liu. 2018. Adapting the neural encoder-decoder framework from single to multi-document summarization. *arXiv preprint arXiv:1808.06218*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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*.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2020. Multi-document summarization via deep learning techniques: A survey. *arXiv preprint arXiv:2011.04843*.

Yogesh Kumar Meena, Ashish Jain, and Dinesh Gopalani. 2014. Survey on graph and cluster based approaches in multi-document text summarization. In *International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014)*, pages 1–5. IEEE.

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*.

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

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. ArXiv:2203.02155 [cs].

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019. Better language models and their implications. *OpenAI Blog https://openai.com/blog/better-language-models*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

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.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

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.

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.

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.

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*.

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*.

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 Moufarek, 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, Xiang-Hai 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 Hirnschall, 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.

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, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. ArXiv:2302.13971 [cs].

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. 2023b. Llama 2: Open foundation and fine-tuned chat models. ArXiv:2307.09288 [cs].

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*.

Dongling Xiao, Han Zhang, Yukun Li, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-gen: An enhanced multi-flow pre-training and fine-tuning framework for natural language generation. *arXiv preprint arXiv:2001.11314*.

Yu Yan, Weizhen Qi, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. *arXiv preprint arXiv:2001.04063*.

Michihiro Yasunaga, Rui Zhang, Kshitijh Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. *arXiv preprint arXiv:1706.06681*.

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*.

Jinming Zhao, Ming Liu, Longxiang Gao, Yuan Jin, Lan Du, He Zhao, He Zhang, and Gholamreza Haffari. 2020. Summpip: Unsupervised multi-document summarization with sentence graph compression. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1949–1952.

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*.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. *arXiv preprint arXiv:1704.07073*.

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*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, page 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Appendix

A.1 Additional Figures

Figure 4 shows a comprehensive breakdown of the best scores obtained by each model for each dataset. Figure 5 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 Prompt Design

The prompt designs for each group mostly follow the format covered in section 3.2 but the entire taxonomy is best laid out by Figure 6. The exact prompts are laid out in the following passage.

System Role Templates 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:

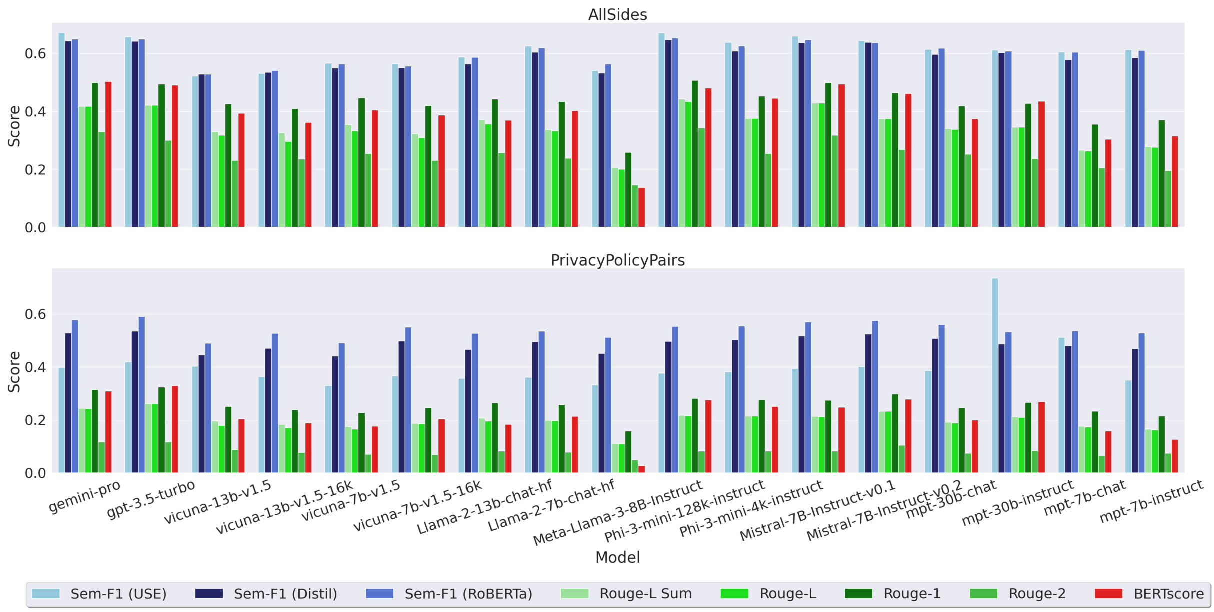

Figure 4: Best scores over each TeLER prompt level for all 16 evaluated LLMs and for each dataset. Red shows BERTscore, green shows ROUGE, and blue shows Sem-F1.

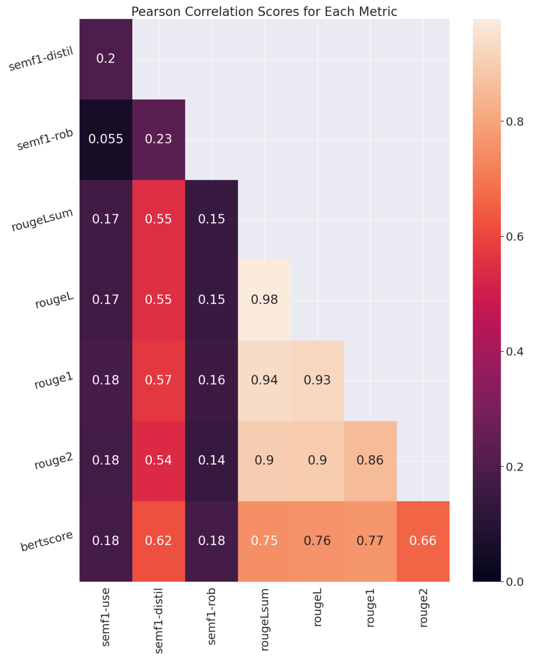

Figure 5: Correlation scores between all evaluation metrics.

- 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**

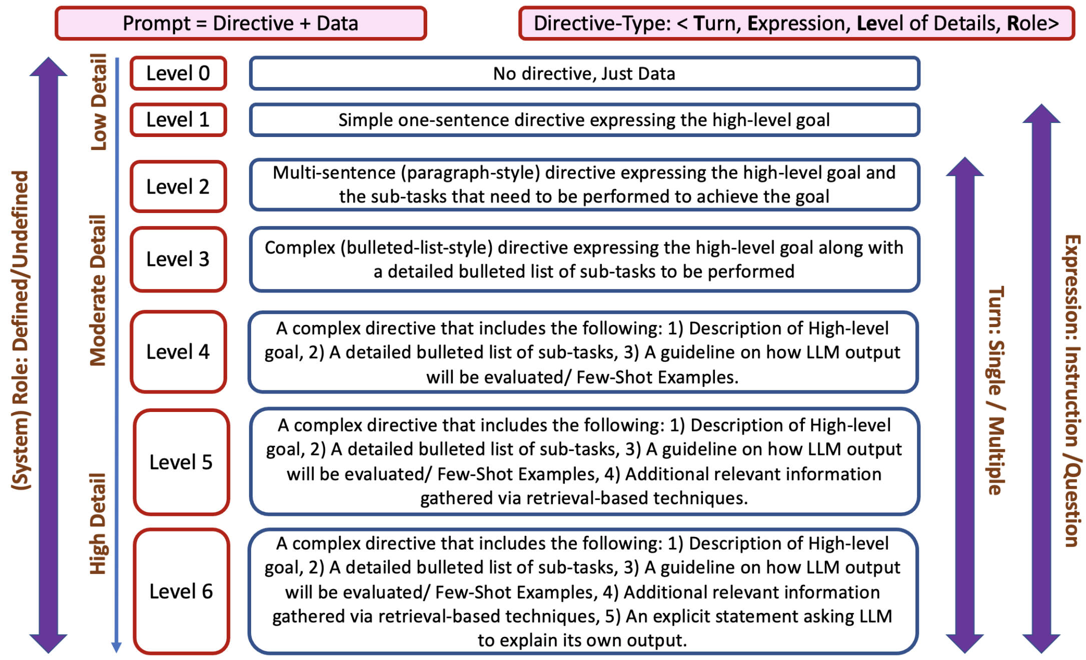

Figure 6: **TELeR** Taxonomy proposed by Santu and Feng (2023): (<**T**urn, **E**xpression, **Le**vel of Details, **R**ole>)

- 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 Templates: 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 Templates: 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 Templates 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.