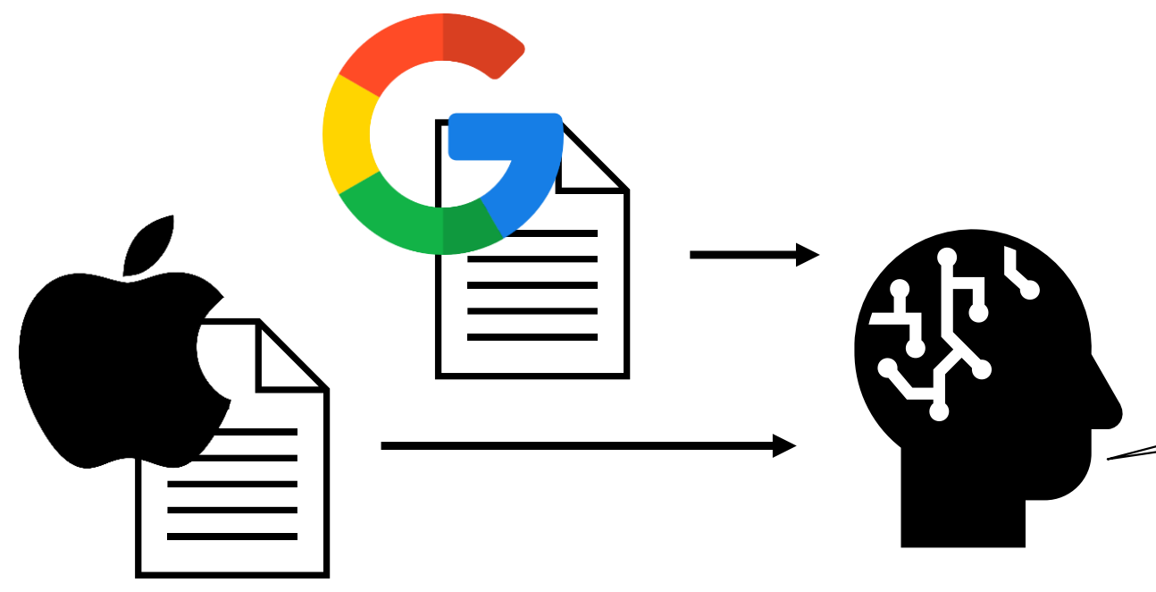
Benchmarking LLMs on the Semantic Overlap Summarization Task

John Salvador, Naman Bansal, Mousumi Akter, Souvika Sarkar, Anupam Das, and Shubhra Kanti Karmaker (“Santu”)
jms0256, nbansal, mza0170, szs0239, sks0086}@auburn.edu,
mousumi.akter@tuportmund.de, anupam.das@ncsu.edu

Introduction

- **Semantic Overlap Summarization** (SOS): given 2 narratives N_1 and N_2, create a summary that captures the overlapping information between N_1 and N_2.

- **Applications**:
  - Peer Reviewing
  - Security and Privacy
  - Journalism/News



Both Google and Apple will collect information including IP addresses, GPS location, etc.

Methodology

Dataset Creation
- Use existing privacy policy dataset as a base
- Group based on company sector (ex. *Food and Drink*)
- Further group on previously annotated categories (ex. Data Retention)
- Annotate paired data

Benchmarking
- Choose LLMs for evaluation, and target metrics: ROUGE, BERTscore, and Sem-F1

LLM Family	Model
Google PaLM2 (Anil et al., 2023)	chat-bison-001 (May 2023)
OpenAI (OpenAI, 2023)	gpt-3.5-turbo-0613 gpt-4-0613
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)

- Create diverse set of prompts.
- Evaluate and analyze data.

Introducing the *PrivacyPolicyPairs* (3P) Dataset

- Sourced from the OPP-115 Corpus

- 2 Source Documents, 3 Annotations

- 135 High Quality Samples

- Increases amount of SOS data and diversifies domains of available data (Privacy Policies *and* News data now available).

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

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

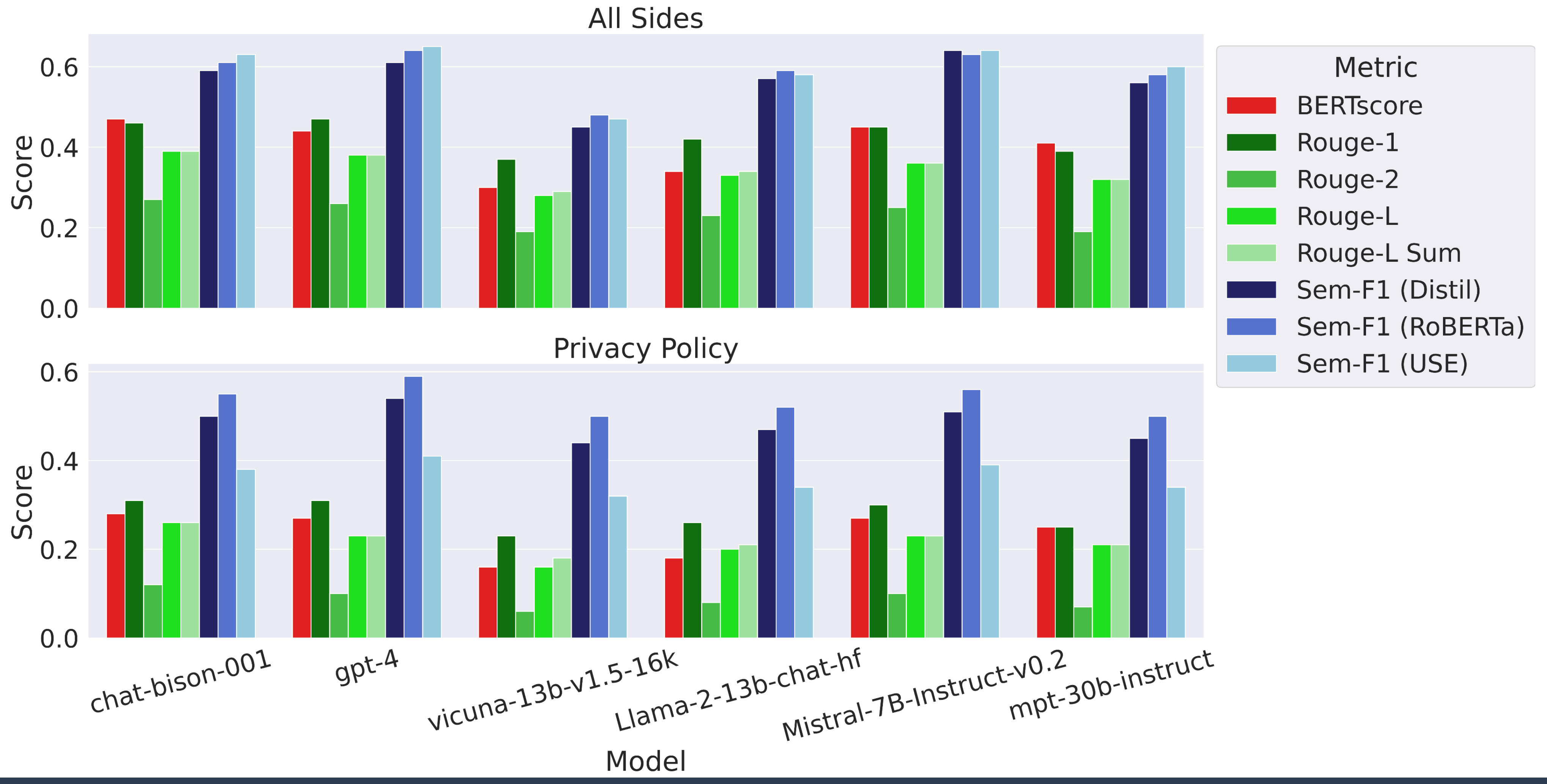
A_1	A_2	A_3
We work to protect the security of your information during transmission by using Secure Sockets Layer (SSL) software, which encrypts information you input. We reveal only the last four digits of your credit card numbers when confirming an order. Of course, we transmit the entire credit card 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.

Benchmark Results

- Commercial LLMs such as GPT-4 and PALM2 generally outperform open source LLMs.

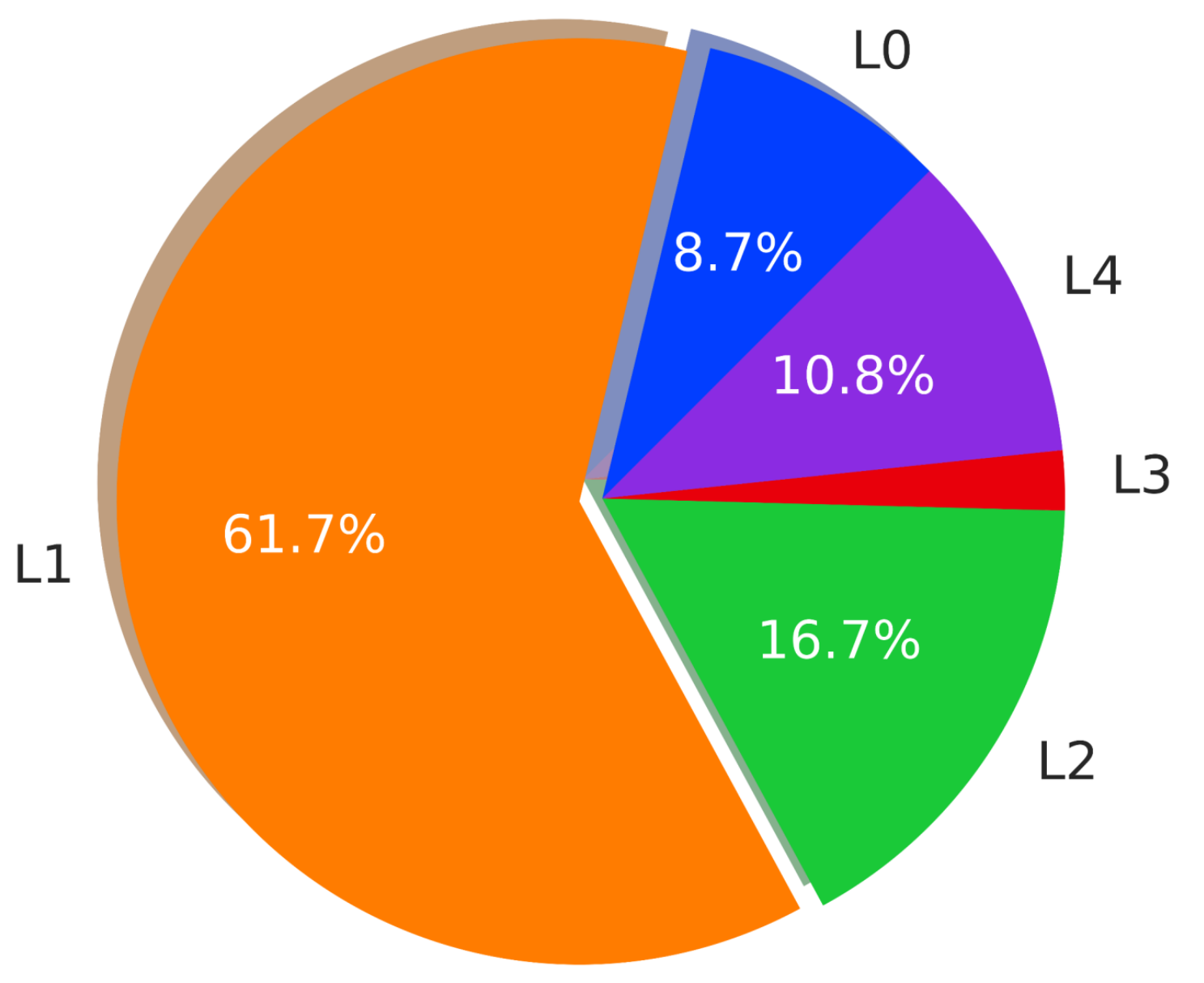
- Mistral-7B-Instruct-v0.2 score best among open source models

- 3P Dataset is *Harder* than the previously introduced AllSides dataset for the SOS task.



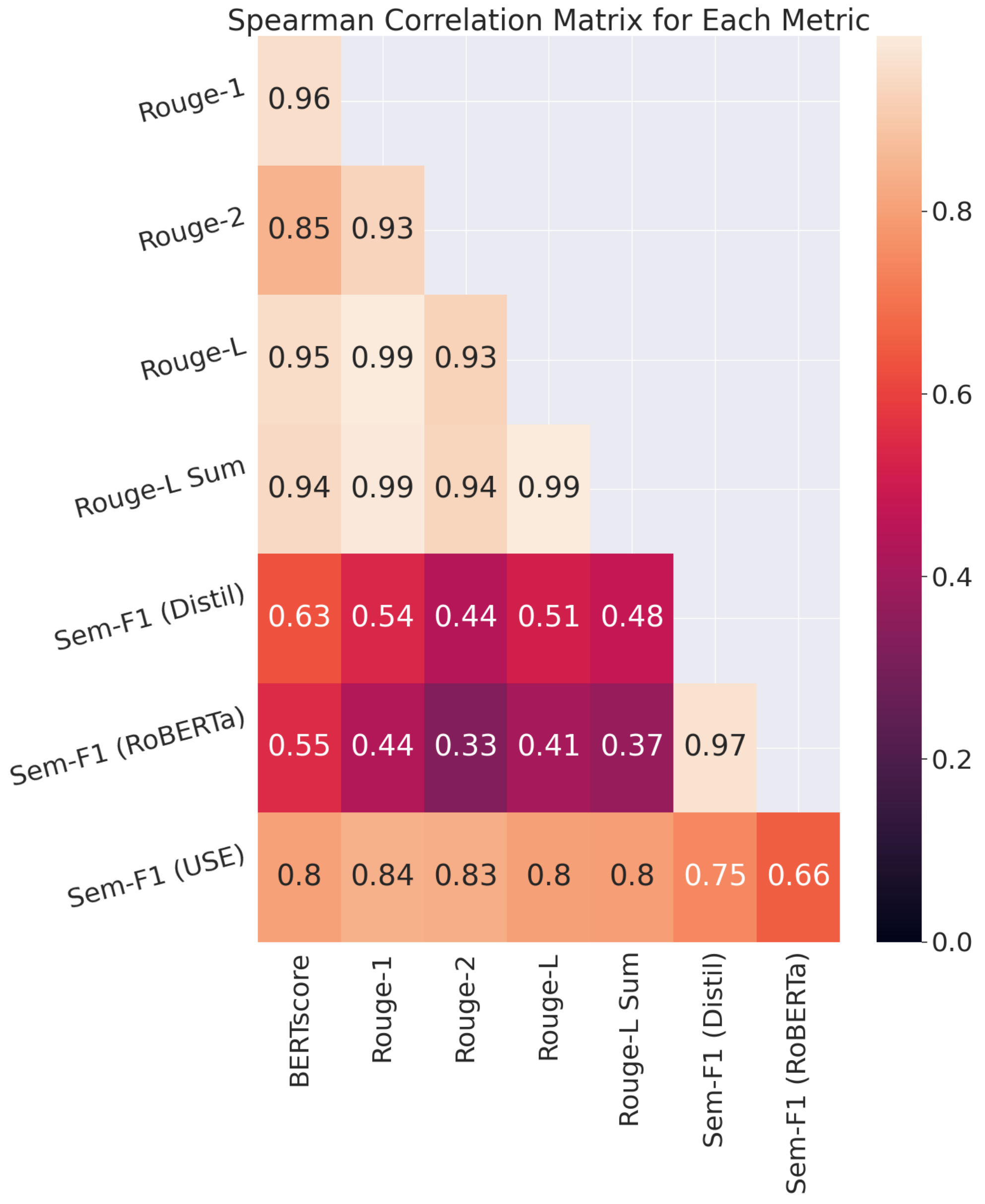
Observations and Limitations

TELeR Level 1 prompts consistently scored heist for each metric



Dataset	Level	BERT score	R-1	R-2	R-L	R-L-Sum	Sem-F1 (Distil)	Sem-F1 (RoBERTa)	Sem-F1 (USE)
Privacy Policy Pairs (3P)	0	-0.069	0.119	0.035	0.085	0.095	0.398	0.444	0.288
	1	0.139	**0.235**	**0.073**	**0.181**	**0.185**	0.438	0.473	0.323
	2	**0.160**	0.223	0.059	0.158	0.166	0.447	0.501	0.330
	3	0.135	0.209	0.053	0.149	0.155	0.452	0.504	0.333
	4	0.145	0.214	0.058	0.152	0.157	**0.462**	**0.511**	**0.339**
AllSides	0	0.105	0.255	0.123	0.177	0.190	0.475	0.493	0.481
	1	**0.265**	**0.365**	**0.199**	**0.290**	**0.292**	**0.511**	**0.526**	**0.525**
	2	0.227	0.327	0.154	0.243	0.249	0.443	0.473	0.475
	3	0.239	0.331	0.152	0.243	0.252	0.437	0.474	0.472
	4	0.249	0.332	0.159	0.245	0.251	0.453	0.489	0.486

Good agreement between Metrics with the exception of Sem-F1 (Distl, RoBERTa)


Spearman Correlation Matrix for Each Metric
Rouge-1
Rouge-2
Rouge-L
Rouge-L Sum
Sem-F1 (Distil)
Sem-F1 (RoBERTa)
Sem-F1 (USE)
BERTscore
Rouge-1
Rouge-2
Rouge-L
Rouge-L Sum
Sem-F1 (Distil)
Sem-F1 (RoBERTa)