A Prompt Design

Prompts for TriviaQA:

- **Flan-ul2 model** and **GPT-4**: Answer the following question in less than 5 words
  Q: {question}
  A:
- **Llama-2-13b-chat model** Answer these following question as succinctly as possible in less than 5 words
  Q: In Scotland a bothy/bothie is a?
  A: House
  Q: Who is Posh Spice in the spice girls pop band?
  A: Victoria Beckham
  Q: {question}
  A:
- **Mistral-7B-Instruct-v0.2 model** <s>[INST] Answer the following question as succinctly as possible in plain text and in less than 5 words. question [/INST]

Prompts for CoQA

- **Flan-ul2 model**, **Llama-2-13b-chat model** and **GPT-4**: Provide an answer in less than 5 words for the following question based on the context below: context: {context} Question: {question} Answer:
- **Mistral-7B-Instruct-v0.2 model** <s>[INST] Provide an answer in less than 5 words for the following question based on the context below:
  context: {context}
  Question: {question}
  Answer: [/INST]

Prompts for SQuAD

- **Flan-ul2 model**, **Llama-2-13b-chat model** and **GPT-4**: Provide an answer for the following question based on the context below, in less than 5 words:
- **Mistral-7B-Instruct-v0.2 model** <s>[INST] Provide an answer for the following question based on the context below, in less than 5 words:
  context: {context}
  Question: {question}
  Answer: [/INST]

Prompts for NQ: For all the models we used the following prompt:
Here are 5 Example Question Answer pairs:
Question: who makes up the state council in russia
Answer: governors and presidents
Question: when does real time with bill maher come back
Answer: November 9, 2018
Question: where did the phrase american dream come from
Answer: the mystique regarding frontier life
Question: what do you call a group of eels
Answer: bed
Question: who wrote the score for mission impossible fallout

Table 11: AUROCs on four Q&A and two summarization datasets (CNN, XSUM) using a total of six LLMs (Llama, Flan-ul2, Mistral, GPT-4, Pegasus, BART), where the number of queries to the LLMs is the same for the baselines and our method. Higher values are better. Best results **bolded**.

Dataset(LLM)	# of SS	Lexical Similarity	EigenValue	Eccentricity	Degree	SE	AVC	Ours
TriviaQA(Llama)	0.74	0.76	0.76	0.77	0.77	0.76	0.79	**0.88**
TriviaQA(Flan-ul2)	0.82	0.81	0.87	0.86	0.86	0.85	0.81	**0.95**
TriviaQA(Mistral)	0.65	0.72	0.76	0.75	0.75	0.68	0.73	**0.81**
TriviaQA(GPT-4)	0.89	0.91	0.91	0.92	0.91	0.92	0.94	**0.96**
SQuAD(Llama)	0.65	0.72	0.74	0.58	0.72	0.61	0.61	**0.83**
SQuAD(Flan-ul2)	0.6	0.7	0.67	0.65	0.67	0.63	0.66	**0.8**
SQuAD(Mistral)	0.59	0.7	0.67	0.65	0.67	0.62	0.64	**0.84**
SQuAD(GPT-4)	0.79	0.82	0.84	0.79	0.83	0.81	0.86	**0.91**
CoQA(Llama)	0.61	0.74	0.76	0.76	0.77	0.64	0.78	**0.92**
CoQA(Flan-ul2)	0.61	0.76	0.78	0.78	0.79	0.63	0.76	**0.87**
CoQA(Mistral)	0.56	0.74	0.79	0.77	0.79	0.59	0.75	**0.81**
CoQA(GPT-4)	0.81	0.86	0.88	0.87	0.88	0.89	0.91	**0.95**
NQ(Llama)	0.65	0.75	0.75	0.73	0.74	0.68	0.74	**0.85**
NQ(Flan-ul2)	0.76	0.76	0.86	0.86	0.86	0.81	0.84	**0.93**
NQ(Mistral)	0.66	0.73	0.77	0.77	0.78	0.68	0.75	**0.83**
NQ(GPT-4)	0.81	0.85	0.85	0.85	0.88	0.89	0.9	**0.93**
CNN (Pegasus)	0.51	0.67	0.73	0.72	0.72	0.55	0.73	**0.77**
CNN (BART)	0.51	**0.59**	0.52	0.48	0.54	0.53	0.5	0.57
XSUM (Pegasus)	0.51	0.58	0.69	0.70	0.71	0.54	0.71	**0.73**
XSUM (BART)	0.51	**0.59**	0.54	0.52	0.52	0.52	0.53	0.57

Answer: Lorne Balfe
Now answer the following Question succinctly, similar to the above examples:
Question: {question}
Answer:

Prompt for GPT-4 as-a-judge: Please provide a score between 0 and 1 of how similar the summaries are. 1 indicating very similar and 0 indicating very different.

Table 12: AUARCs on four Q&A and two summarization datasets (CNN, XSUM) using a total of six LLMs (Llama, Flan-ul2, Mistral, Pegasus, BART), where the number of queries to the LLMs is the same for the baselines and our method. Higher values are better. Best results **bolded**.

Dataset(LLM)	# of SS	Lexical Similarity	EigenValue	Eccentricity	Degree	SE	AVC	Ours
TriviaQA(Llama)	0.76	0.8	0.81	0.8	0.8	0.79	0.8	**0.83**
TriviaQA(Flan-ul2)	0.7	0.72	0.73	0.73	0.73	0.71	0.72	**0.74**
TriviaQA(Mistral)	0.55	0.63	**0.64**	**0.64**	**0.64**	0.58	0.63	**0.64**
TriviaQA(GPT-4)	0.8	0.84	0.84	0.84	0.82	0.84	0.85	**0.89**
SQuAD(Llama)	0.3	0.36	0.37	0.28	0.36	0.36	0.31	**0.68**
SQuAD(Flan-ul2)	0.73	0.95	0.83	0.82	0.83	0.78	0.83	**0.96**
SQuAD(Mistral)	0.72	0.93	0.82	0.82	0.82	0.76	0.83	**0.96**
SQuAD(GPT-4)	0.7	0.72	0.72	0.63	0.66	0.69	0.71	**0.83**
CoQA(Llama)	0.56	0.67	0.67	0.67	0.67	0.61	0.66	**0.71**
CoQA(Flan-ul2)	0.7	0.79	**0.8**	0.79	0.79	0.73	0.77	**0.8**
CoQA(Mistral)	0.46	0.62	0.64	0.63	**0.64**	0.51	0.62	0.61
CoQA(GPT-4)	0.68	0.73	0.72	0.73	0.74	0.72	0.76	**0.86**
NQ(Llama)	0.37	0.41	0.42	0.41	0.41	0.39	0.42	**0.45**
NQ(Flan-ul2)	0.41	0.44	**0.47**	0.46	0.45	0.44	0.45	**0.47**
NQ(Mistral)	0.32	0.38	0.40	0.40	0.39	0.36	0.39	**0.42**
NQ(GPT-4)	0.69	0.73	0.74	0.74	0.74	0.73	0.72	**0.79**
CNN (Pegasus)	0.45	0.51	0.53	0.43	0.52	0.48	0.47	**0.74**
CNN (BART)	0.21	0.22	0.21	0.21	0.21	0.23	0.23	**0.34**
XSUM (Pegasus)	0.16	0.17	0.19	0.17	0.17	0.21	0.19	**0.27**
XSUM (BART)	0.21	0.22	0.20	0.21	0.22	0.23	0.22	**0.35**

Table 13: AUROCs on two summarization datasets (CNN, XSUM) with GPT-4 as a judge. Higher values are better. Best results **bolded**.

Dataset(LLM)	# of SS	Lexical Similarity	EigenValue	Eccentricity	Degree	SE	AVC	Ours
CNN (Pegasus)	0.54	0.65	0.76	0.77	0.75	0.61	0.75	**0.81**
CNN (BART)	0.55	**0.64**	0.55	0.52	0.58	0.56	0.54	**0.64**
XSUM (Pegasus)	0.56	0.62	0.72	0.74	0.73	0.6	0.75	**0.79**
XSUM (BART)	0.55	**0.63**	0.56	0.54	0.55	0.56	0.59	0.61

Table 14: AUARCs two summarization datasets (CNN, XSUM) with GPT-4 as a judge. Higher values are better. Best results **bolded**.

Dataset(LLM)	# of SS	Lexical Similarity	EigenValue	Eccentricity	Degree	SE	AVC	Ours
CNN (Pegasus)	0.49	0.55	0.58	0.49	0.57	0.52	0.53	**0.77**
CNN (BART)	0.25	0.26	0.27	0.26	0.26	0.27	0.29	**0.35**
XSUM (Pegasus)	0.19	0.22	0.23	0.2	0.21	0.23	0.21	**0.29**
XSUM (BART)	0.26	0.26	0.25	0.27	0.27	0.27	0.26	**0.37**

Table 15: ECEs two summarization datasets (CNN, XSUM) with GPT-4 as a judge. Lower values are better. Best results **bolded**.

Dataset(LLM)	# of SS	Lexical Similarity	EigenValue	Eccentricity	Degree	SE	AVC	Ours
CNN (Pegasus)	0.18	0.14	0.11	0.1	0.09	0.15	0.07	**0.05**
CNN (BART)	0.48	0.17	0.24	0.25	0.22	0.22	0.22	**0.14**
XSUM (Pegasus)	0.18	0.18	0.13	0.11	0.09	0.17	0.1	**0.06**
XSUM (BART)	0.23	0.19	0.21	0.23	0.23	0.22	0.2	**0.16**