News Summarization and Evaluation in the Era of GPT-3

Introduction

- Fine-tuning pre-trained models on domain-specific datasets has been the leading paradigm in text summarization research in recent years.
- The success of prompt-based models provides an alternative approach.
- In this paper, we conduct the first systematic study of the impact of prompt-based models on the text summarization research space, using an Instruct-tuned 175B GPT-3 model as a case study.

CNN article: https://www.cnn.com/2022/09/09/politics/judge-throws-out-trumps-rico-lawsuit-against-hillary-clinton-and-democrats/index.html

Length-constrained prompt

Summarize the above article in 2 sentences.

GPT-3: Donald Trump's lawsuit against Hillary Clinton, the Democratic National Committee, and several other individuals and entities has been dismissed by a federal judge. The judge ruled that Trump's claims are unsupported by any legal authority and that many of his specific factual assertions are implausible or unsupported.

Keyword-constrained prompt

Summarize the above article briefly focusing on Alina Habba.

GPT-3: Alina Habba is an attorney for Donald Trump. She stated that Trump will "immediately move to appeal this decision." Habba also said that the judge's opinion is "rife with erroneous applications of the law" and that it "disregards the numerous independent governmental investigations which substantiate our claim that the defendants conspired to falsely implicate our client and undermine the 2016 Presidential election."

Figure 1: Examples of zero-shot summarization using GPT-3. We can generate summaries following style constraints or queries included in the prompts, allowing us to emulate a range of existing fine-tuned systems.

Models and Setup

- Recent zero and few-shot prompting based models, have shown impressive generalization capabilities on unseen tasks specified using prompts alone and without performing any gradient updates.
- The former set consists of fine-tuned language models, trained on a large number of article-summary pairs, PEGASUS to obtain dataset-specific systems.
- On the other extreme are zero- or few-shot models, that are not explicitly trained for any particular task.

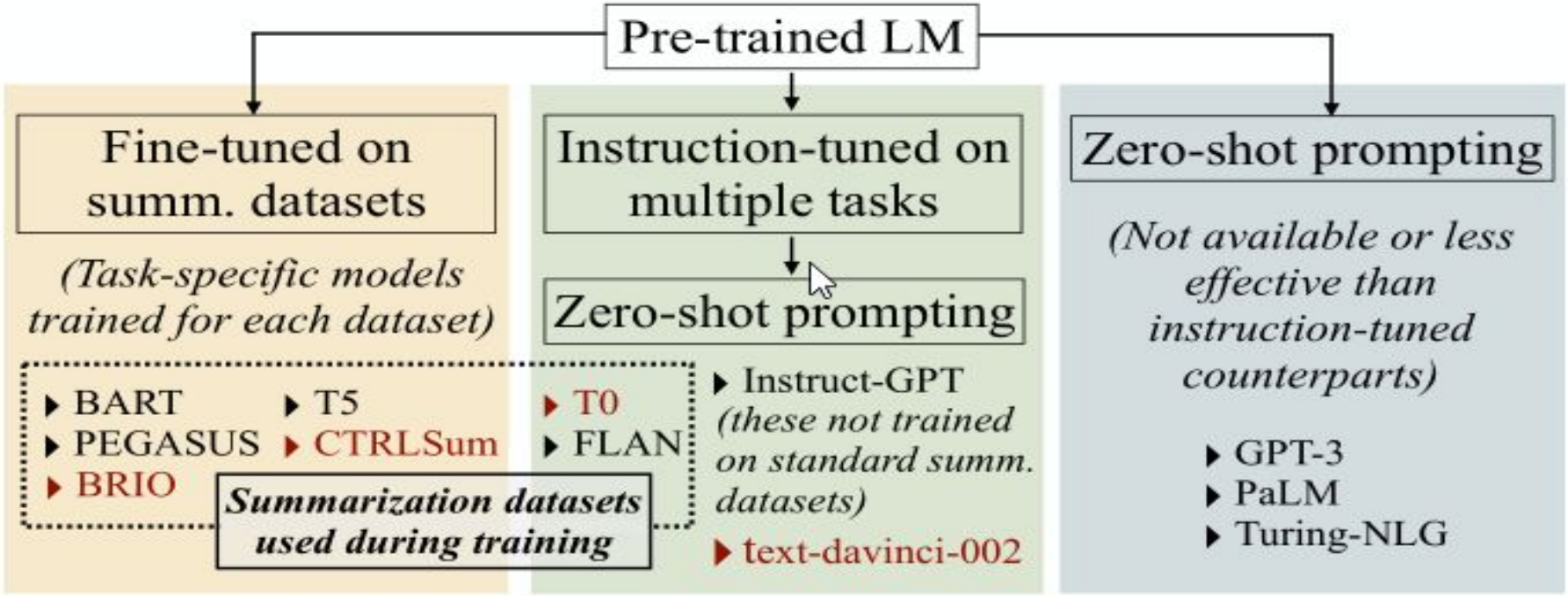

Figure 2: Broad categorization of available summarization systems; those compared in this work are highlighted in red.

Experimental Setup

- We choose two standard fine-tuning datasets whose summaries differ along multiple dimensions such as length and abstractiveness:
  - 1. CNN/DM contains reference summaries that are approximately 3-4 sentences long. Summaries in this dataset are highly extractive and lead-biased.
  - 2. XSum contains 1 sentence summaries of BBC news articles. In this dataset, references summaries, and consequently generated summaries from fine-tuned models are highly abstractive.

CNN Article: *(CNN) Mortgage rates fell slightly this week, marking the third consecutive week of declines. But with rates hovering above 5% and home prices well above where they were this time last year, prospective buyers are finding it increasingly difficult to afford a home. The 30-year, fixed-rate mortgage averaged 5.09% in the week ending June 2, down from 5.10% the week before, according to Freddie Mac. It is still well above the 2.99% average from this time last year. […] Rising prices had already been pushing many prospective buyers to the sidelines […] The Fed has been seeking to tame inflation by raising interest rates over the past couple of months. And the central bank has signaled there are more rate hikes to come. […]*

BRIO	The 30-year, fixed-rate mortgage averaged 5.09% in the week ending June 2. It's the third consecutive week of declines. But rates are still above 5% and home prices are well above where they were this time last year. Rising prices have pushed some prospective buyers to the sidelines.
T0	30-year, fixed-rate mortgage averaged 5.09% in the week ending June 2. It is still well above the 2.99% average from this time last year. Rising prices had already been pushing many prospective buyers to the sidelines.
GPT3	Mortgage rates have fallen slightly in the past few weeks, but they are still well above last year's levels. This is making it difficult for many prospective buyers to afford a home. The Fed has been raising interest rates in an effort to control inflation, and more rate hikes are expected.

BBC Article: *The full opening of a landmark shared education campus in Omagh is facing another delay. Education Minister Michelle McIlveen has now said the Strule Shared Education Campus is planned to open in September 2026. The minister clarified the new date in response to an assembly question from the SDLP MLA Daniel McCrossan. The campus had originally been due to open in 2020, but has been delayed a number of times. The estimated cost of the project has also risen substantially to about £230m according to figures previously released by the minister in November 2021. […]*

BRIO	All images: Strule Shared Education Campus. ***After removing first sentence of input article:*** The full opening of the Strule shared education campus in Omagh, County Tyrone, has been delayed to September 2026.
T0	Strule, Northern Ireland's biggest ever school-building project, is now scheduled to open in 2026, the Education Minister has said.
GPT3	The Strule Shared Education Campus is facing another delay, and is now planned to open in September 2026.

Figure 4: Examples of CNN-style and BBC/XSum-style summaries for the three systems. For CNN, we observe that models fine-tuned on the `CNN/DM` training set reflect its dataset biases; summaries are highly extractive, specific and lead-biased. On the other hand, `GPT3-D2` summaries contain fewer specific details but cover more content.

Results

Differences between summarization systems. Figure shows examples of generated summaries from all three summarization systems for both CNN and BBC articles.

Model	Length Statistics		% novel n-gms		#NEs per
	#sent	#words/sent	$n = 1$	$n = 2$	100 words
CNN					
BRIO	3.7	15.8	12.1	36.2	12.9
T0	2.7	14.9	16.4	45.2	12.8
GPT3-D2	2.9	23.4	16.3	40.7	10.5
BBC					
BRIO	1.0	20.2	24.6	61.2	9.1
T0	1.0	20.0	26.3	66.7	9.8
GPT3-D2	1.0	27.7	16.4	42.3	8.5

Table 2: Statistics for generated summaries evaluated in the human study across all datasets and summarization systems. We observe that GPT3-D2 generated summaries nearly always follow the sentence length constraints in their prompts.

Can current automatic metrics evaluate GPT3-D2 summaries?

- Automatic metrics proposed for summarization evaluation can be broadly divided into two categories:
  - reference-based , that compare generated summaries against available gold summaries,
  - reference-free that only rely on the input document. Here, we compare their performance at evaluating zero-shot GPT3-D2 summaries.

Dataset	Model	Overall Quality		Factuality (QA-based)		Factuality (NLI-based)		
		SUPERT	BLANC	QuestEval	QAFactEval	FactCC	DAE	SummaC
CNN	PEGASUS	.5466	.0605	.7373	4.4071	.3743	.8223	.1138
	BRIO	.5586	.0802	.7334	3.8332	.1817	.7577	-.0532
	T0	.5330	.0558	.7799	3.7517	.2012	.7556	-.0605
	GPT3-D2	.5560	.0749	.7249	3.6399	.2428	.6671	-.0729
DailyMail	PEGASUS	.6433	.1137	.7536	4.4677	.5152	.8497	.2402
	BRIO	.6360	.1217	.7415	4.1362	.3699	.8118	.0153
	T0	.5995	.0889	.7803	3.9827	.2431	.8043	.0478
	GPT3-D2	.6118	.0983	.7461	3.8279	.2697	.6990	.0365
XSum	PEGASUS	.4439	.0249	.8233	2.0089	.2465	.3598	-.2993
	BRIO	.4459	.0230	.8305	1.8626	.2031	.3040	-.3292
	T0	.4538	.0238	.7957	2.0330	.2219	.3392	-.3037
	GPT3-D2	.5060	.0594	.8064	2.9492	.3977	.6372	-.2626
Newsroom	PEGASUS	.6286	.1131	.7118	4.2120	.7218	.7956	.2418
	BRIO	-	-	-	-	-	-	-
	T0	.5433	.0640	.7511	3.5799	.2828	.7376	.0261
	GPT3-D2	.5408	.0599	.7160	3.2336	.3988	.6564	-.0729

Table 5: Performance of different summarization systems, as scored by automatic reference-free evaluation metrics from the summarization literature. Similar to reference-based metrics, these also generally fail to produce the same system rankings as human preferences reliably across datasets.

Beyond Generic Summarization

- Different users can have very different information needs from the same article, all of which cannot be satisfied with a single generic summary.
- Prior work has introduced several task formulations to address this gap, including keyword-focused query-focused or aspect-focused summarization, amongst others.
- Here, we evaluate GPT3-D2 performance at two of these use cases.
    - keyword-based summarization
    - aspect-based summarization

Conclusion

- In this work, we performed the first systematic study comparing zero-shot GPT-3 and fine-tuned models at the news summarization task.
- We analyzed the impact of zero-shot models on the summarization field, including training paradigms and evaluation practices.
- Finally, to support further research and analysis into zero-shot summarization, we release a large corpus of generated summaries for multiple zero-shot and fine-tuned models, as well as human preference judgments comparing these systems.