News Summarization and Evaluation in the Era of GPT-3

Tanya Goyal[1] **Junyi Jessy Li**[2] **Greg Durrett**[1]
[1] Department of Computer Science [2] Department of Linguistics
The University of Texas at Austin
tanyagoyal@utexas.edu

Abstract

The recent success of zero- and few-shot prompting with models like GPT-3 has led to a paradigm shift in NLP research. In this paper, we study its impact on text summarization, focusing on the classic benchmark domain of news summarization. First, we investigate how zero-shot GPT-3 compares against fine-tuned models trained on large summarization datasets. We show that not only do humans overwhelmingly prefer GPT-3 summaries, but these also do not suffer from common dataset-specific issues such as poor factuality. Next, we study what this means for evaluation, particularly the role of gold standard test sets. Our experiments show that both reference-based and reference-free automatic metrics, e.g. recently proposed QA- or entailment-based factuality approaches, cannot reliably evaluate zero-shot summaries. Finally, we discuss future research challenges beyond generic summarization, specifically, keyword- and aspect-based summarization, showing how dominant fine-tuning approaches compare to zero-shot prompting.

To support further research, we release: (a) a corpus of 10K generated summaries from fine-tuned and zero-shot models across 4 standard summarization benchmarks, (b) 1K human preference judgments and rationales comparing different systems for generic- and keyword-based summarization.[1]

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.

1 Introduction

Fine-tuning pre-trained models on domain-specific datasets has been the leading paradigm in text summarization research in recent years (Lewis et al., 2020; Zhang et al., 2020; Raffel et al., 2020). These models generate high-quality summaries on standard benchmarks, but still require sizeable training datasets to adapt to new settings, e.g., summarizing data from a new source domain or producing a summary in a different style. The success of prompt-based models (GPT-3 (Brown et al., 2020), T0 (Sanh et al., 2022), PaLM (Chowdhery et al., 2022), etc.) provides an alternative approach, namely learning from natural language task instructions and/or a few demonstrative examples in the context without updating model parameters. While recent work (Zhao et al., 2021; Min et al., 2022; Ye and Durrett, 2022) has analyzed and evaluated this paradigm across a number of tasks, it has only been studied for text summarization with unreliable automatic metrics (He et al., 2022; Chowdhery et al., 2022; Ouyang et al., 2022) or in non-standard settings (Saunders et al., 2022).

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 (text-davinci-002) (Brown et al., 2020; Ouyang et al., 2022) as a case study. Figure 1 shows that GPT-3 summaries are extremely high-quality and adaptable to different summarization settings. Starting from these

[1] All data available at: https://tagoyal.github.io/zeroshot-news-annotations.html.

observations, we aim to answer three main questions. First, how do prompt-based GPT-3 summaries compare to those obtained from state-of-the-art fine-tuned summarization models (Zhang et al., 2020; Liu et al., 2022)? We compare these approaches using A/B testing on a new corpus of recent news articles, and find that our study participants overwhelmingly prefer zero-shot GPT-3 summaries across two different "styles" with different prompts (three-sentence and single-sentence). Moreover, these zero-shot summaries do not suffer from limitations due to low-quality training data that plague fine-tuned generic summarization models (Maynez et al., 2020; Goyal et al., 2022).

Second, are existing automatic metrics well-suited to evaluating zero-shot summaries? Recent work has shown that classic reference-based such as ROUGE (Lin, 2004) and BERTScore (Zhang* et al., 2020) are unreliable when small improvements are reported (Peyrard, 2019; Fabbri et al., 2021); however large differences, on the order of say 5 ROUGE points or greater, are considered to be correlated with human preferences (Bhandari et al., 2020; Deutsch et al., 2022). However, we find that the same is no longer true when evaluating zero-shot GPT-3 summaries. These summaries score much lower on automatic metrics (7 ROUGE-L points on average) than all prior state-of-the-art models while comfortably outperforming them on human evaluation. Furthermore, we show that recent reference-free metrics, e.g. QA-based metrics (Fabbri et al., 2022; Durmus et al., 2020) and trained factuality models (Kryscinski et al., 2020; Goyal and Durrett, 2020), similarly fail to adapt to this shift from the fine-tuned to zero-shot summarization space and need to be re-visited.

Finally, how can zero-shot summaries be used for use cases beyond generic summarization? We focus on keyword-based and aspect-based summarization. For keyword-based summarization, we find that zero-shot GPT-3 consistently generates more coherent and keyword-relevant summaries compared to current fine-tuned alternatives: crowd annotators prefer GPT-3 summaries over a baseline model (He et al., 2020) 70% of the time. We observe mixed results for the aspect-based setting, where GPT-3 summaries show frequent failure cases when prompted with simple prompts for aspect-based summarization.

Taken together, this evidence suggests that GPT-3 represents a fundamental paradigm shift in summarization, changing the what data we need (or don't need) and what approaches we can now explore. Evaluating these systems as they progress further will require a new framework distinct from the automatic metrics that have dominated the last decade of summarization research.

2 Models and Setup

2.1 Current Paradigms for Summarization

Recent zero- and few-shot prompting based models (Brown et al., 2020; Sanh et al., 2022), have shown impressive generalization capabilities on unseen tasks specified using prompts alone and without performing any gradient updates (Mishra et al., 2022). In this work, we want to compare their text summarization performance against the current state-of-the-art models.

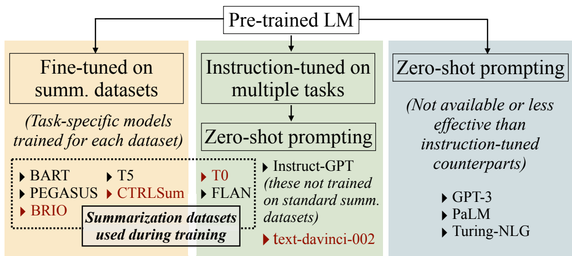


Figure 2: Broad categorization of available summarization systems; those compared in this work are highlighted in red.

Figure 2 shows the broad categories of all available summarization approaches, including current SOTA models and prompting-based models. The former set consists of **fine-tuned** language models, trained on a large number of article-summary pairs (e.g. BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020), BRIO (Liu et al., 2022)) to obtain dataset-specific systems. This category also includes models aimed at tasks beyond generic summarization, such as keyword- or query-based summarization, that still rely on standard datasets for training (He et al., 2020).

On the other extreme are **zero- or few-shot** models, (e.g. GPT3 (Brown et al., 2020), PaLM (Chowdhery et al., 2022)), that are not explicitly trained for any particular task, as discussed above. Recent work (Ouyang et al., 2022; Wei et al., 2022; Sanh et al., 2022) has improved on these models by introducing **instruction-tuned** models. Here, pre-trained language models are fine-tuned on multiple tasks (which may include summarization) using instruction templates in order to align their

Dataset	Avg. Words		% novel n-grams	
	Article	Summ	$n = 1$	$n = 2$
CNN	760.5	45.7	16.7	54.3
DailyMail	653.3	54.6	17.0	53.8
XSum (BBC)	431.1	23.2	35.7	82.4
Newsroom	658.6	26.7	18.9	47.5

Table 1: Basic statistics of standard summarization datasets: CNN/DM (Hermann et al., 2015; Nallapati et al., 2016), XSum (Narayan et al., 2018), Newsroom (Grusky et al., 2018). These show large variance in their summary properties and fundamentally differ in their definition of the "gold" standard.

training with inference time usage.

In this work, we compare the summarization performance of three models that are representative of this space of options:

1. **OpenAI's text-davinci-002**, a GPT-3 model (Brown et al., 2020) from the Instruct series (Ouyang et al., 2022). While we do not know the exact training details for this release of the model, the previous in the series (text-davinci-001) was fine-tuned on a combination of prompts submitted to their API and labeler prompts spanning multiple tasks. These tasks include summarization but not (to our knowledge) standard summarization datasets like CNN/DM (Hermann et al., 2015; Nallapati et al., 2016) or XSum (Narayan et al., 2018).

   We choose the text-davinci-002 version for our experiments in order to benchmark the best available prompt-based model.[2] We refer to this approach as GPT3-D2.

2. **BRIO** (Liu et al., 2022), a fine-tuned summarization model that reports state-of-the art results on both CNN/DM and XSum. We will use versions of this model fine-tuned on each of these two datasets.

3. **T0** (Sanh et al., 2022), a prompt-based model fine-tuned on multiple tasks including standard summarization datasets. This provides a useful point of comparison between task-specific fine-tuned (BRIO) and zero-shot models (GPT3-D2).

2.2 Using GPT3-D2 for zero-shot summarization

Fine-tuned models largely follow the "style" of reference summaries in their training data, and hence,

[2]We did not observe obvious quality differences in generated summaries between text-davinci-001 and text-davinci-002. Examples are included in Appendix C.

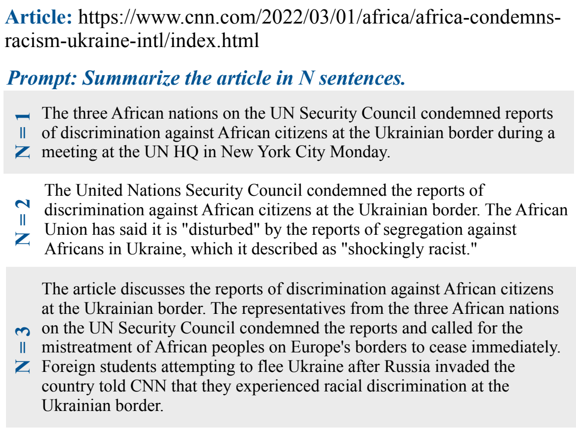


Figure 3: Illustration of length control using the task description / prompt for GPT3-D2. We found that the generated summaries followed the given sentence length constraint 98% of the time, allowing us to generate different length summaries emulating different datasets.

generated summaries show large variance between datasets (see Table 1 for basic summary statistics of standard summarization datasets). To ensure fair comparison between these and zero-shot GPT3-D2, we adapt the latter's prompt to align with dataset-specific styles.

Specifically, we follow prior work (Sanh et al., 2022) and use sentence-count length prompts to adapt to each dataset. Although these datasets also differ along other attributes, e.g. CNN/DM is lead-biased whereas XSum requires inferences drawing from a whole article, we do not attempt to control any other attributed of the summary. Figure 3 shows an example of different length GPT3-D2 summaries for the same news article, using the following prompt format:[3]

Article: {{article}}
Summarize the above article in N sentences.

We found that GPT3-D2 summaries faithfully follow the given length constraint in 98% of the test instances used in our human study data in Section 3.

Given this setup, we first compare the summary quality of the three summarization models through a human annotation study (Section 3). Then, we evaluate the current suite of summarization metrics for zero-shot summarization (Section 4). Finally, in

[3]Although prior work has shown that few-shot learning works better than zero-shot across tasks like inference, textual reasoning, and translation (Brown et al., 2020; Chowdhery et al., 2022; Ouyang et al., 2022; Sanh et al., 2022), we opt for zero-shot prompting as news articles are generally longer and multiple demonstration may not fit within the context.

Section 5, we briefly discuss GPT3-D2 performance on summarization tasks beyond generic summarization and new challenges.

3 Human evaluation of GPT3-D2 summaries

Generated summaries of fine-tuned models (Lewis et al., 2020; Zhang et al., 2020; Liu et al., 2022) emulate gold-standard summaries in their training datasets. In contrast, prompt-based GPT3-D2 models generate summaries based on how the given task description surfaces behavior learned during pre-training or instruction-tuning. In this section, we ask: how do these paradigms compare? Does learning from gold summaries lead to a better summarization model? To answer this, we conduct a human study to compare outputs of our 3 representative models and collect human preferences of quality.

3.1 Experimental Setup

Datasets for fine-tuning We choose two standard fine-tuning datasets whose summaries differ along multiple dimensions such as length and abstractiveness:

1. **CNN/DM** (Hermann et al., 2015; Nallapati et al., 2016) contains reference summaries that are approximately 3-4 sentences long. Summaries in this dataset are highly extractive and lead-biased.

2. **XSum** (Narayan et al., 2018) contains 1 sentence summaries of BBC news articles. In this dataset, references summaries, and consequently generated summaries from fine-tuned models are highly abstractive.

Datasets for evaluation Because GPT3-D2's pre-training and instruction-tuning datasets are unknown, it may have been trained on existing articles and summaries in the test splits of these standard benchmarks. We therefore run our human study on 100 recent articles from CNN[4] and BBC, collected between March 1, 2022 and June 31, 2022. We call these CNN-2022 and BBC-2022 respectively.

Model details We use the publicly released BRIO-XSum and BRIO-CNN/DM models to generate summaries.[5] For T0, we use a prompt we selected from its prompt repository for CNN/DM and XSum datasets.[6] Finally, to generate GPT3-D2 summaries, we set $N = 3$ for CNN and $N = 1$ for BBC in our standard sentence-count prompt template from Section 2.

For a maximally fair comparison in this "realistic" setting, we take some additional steps to improve the output of BRIO-XSum. In order to automate dataset creation, XSum removes the first sentence from news articles to use as the gold summary for training, then treats the rest of the sentences as the article to summarize. This setup differs from the real world usage of summarization systems where the complete article is summarized. Due to this mismatch, BRIO-XSum often generates very low quality outputs, e.g. *All images: Strule Shared Education Campus* in Figure 4, for around 30% of the articles. We manually identify these examples and first attempt to fix them by selecting a summary without such obvious failures from further down the beam (we use beam size $= 10$). However, if we cannot find a "better" summary, we remove the first sentence of the article and re-sample a new summary to align with its noisy training. This latter strategy often results in factually incorrect summary generations, as is well documented in prior research (Maynez et al., 2020; Goyal and Durrett, 2021).

Design of the human study We design an A/B test to collect preference annotations. For each given article, annotators are shown summaries from all three summarization systems (BRIO, T0 and GPT3-D2). They are then asked to select their most and least preferred summary or summaries. In addition to these multiple choice questions, we also ask for a free-text justification of both choices.

We make two design decisions for our human study: first, we do not provide annotators with specific definitions of summary quality to avoid introducing our own biases. It is also quite challenging to produce a unified definition of quality for the very different "styles" of summaries evaluated in this study. Instead, we ask them to rely on their own preferences based on summaries they would like to see if they were browsing the web, which

[4]Although the BRIO's CNN/DM model also includes DailyMail data in its training, we do not use this news source in our study as it is now widely considered to be unreliable. E.g. according to Media Bias / Fact Check site, DM's factual reporting is rated 'low' https://mediabiasfactcheck.com/daily-mail/.

[5]Models at: https://github.com/yixinL7/BRIO

[6]Repository with T0 prompts: https://github.com/bigscience-workshop/promptsource

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.

we believe to be a representative scenario for non-expert consumers of news summaries. Detailed task instructions are included in Appendix F.

Second, we allow multiple selections for both the best and worst summary questions to cater to scenarios in which different summarization systems output similar quality summaries without meaningful differences.

We hire crowd annotators through Prolific. For both CNN and BBC, we recruit 60 unique participants to annotate the 100 summaries in each dataset. Each annotator was asked to annotate 5 articles and each article was annotated by 3 annotators. Additionally, we use the Prolific's demographic filters to restrict participation to USA (or UK) residents for CNN (or BBC). We anticipate that residents from these respective countries are better positioned to understand country-specific news events and evaluate their summaries. Participants were paid approximately $11/hr for their work.

3.2 Results

Differences between summarization systems Figure 4 shows examples of generated summaries from all three summarization systems for both CNN and BBC articles. For CNN, we observe that fine-tuned BRIO summaries tend to be highly extractive and generally include a high number of named entities (dates, percentages, names), reflecting the data it was trained on. In contrast, GPT3-D2 summaries are more abstractive and less specific, but provide a more exhaustive overview of the article content. Table 2 provides quantitative evidence of this; we use percentage of novel n-grams to measure abstractiveness, and number of named entities per 100 words to measure specificity.

Model	Length Statistics #sent	#words/sent	% novel n-gms $n = 1$	$n = 2$	#NEs per 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.

For BBC, we observe inverse trends where BRIO and T0 are more abstractive compared to GPT3-D2. Again, this can be attributed to the XSum training data used to train both these prior models. For GPT3-D2 summaries, on the other hand, the level of abstractiveness does not differ between datasets. Finally, Table 2 shows that GPT3-D2 summaries tend to have longer sentences, and therefore

Dataset	BRIO		T0		GPT3	
	Best ↑	Worst ↓	Best ↑	Worst ↓	Best ↑	Worst ↓
CNN	36	24	8	67	58	9
BBC	20	56	30	29	57	15

Table 3: Percentage of times a summarization system is selected as the best or worst according to majority vote (may be tied). Human annotators have a clear preference for zero-shot GPT3-D2 for both CNN and BBC style summaries.

similar number of summary sentences often results in a longer summary for both datasets. We study the effect of this length difference on human preference judgments in Appendix B.

Which systems do humans prefer? Results of our human study are summarized in Table 3. We report the percentage of times a particular system is the most/least preferred model according to majority vote combining all three annotator's choices.[7] Across both datasets and styles, we observe a clear preference for GPT3-D2 summaries compared to the other two models. In fact, in both scenarios, the GPT3-D2 outperforms the next best model by at least 20 percentage points. This improvement is statistically significant according to a paired bootstrap test (CNN $p-$value $= 2 \times 10^{-3}$, BBC $p-$value $= 6 \times 10^{-4}$).

Note that the next best model differs between the two datasets. For BBC, annotators prefer T0 summaries over BRIO. Annotator rationales often mentioned misleading or incorrect information as the primarily reason for selecting BRIO as the worst summary, confirming the issues that have been observed with XSum-trained models (Maynez et al., 2020; Pagnoni et al., 2021; Goyal and Durrett, 2021). Although T0 also includes XSum training data, we hypothesize that its multi-task framework helps offset the noisy signal from XSum.

In contrast, annotators rate T0 as the worst summarization system for CNN. The most common rationales for these were shorter length and inclusion of irrelevant details, e.g. long quotes, while missing key points. Some annotators also commented that these T0 summaries were less coherent compared to the other models. Interestingly, we did not observe similar complaints for the single-sentence T0 summaries for BBC.

[7] As we allow multiple system selections, note that more that one system could be the majority. However, this is rare after majority vote: only 2% of the articles in CNN and 7% in BBC have multiple best summaries.

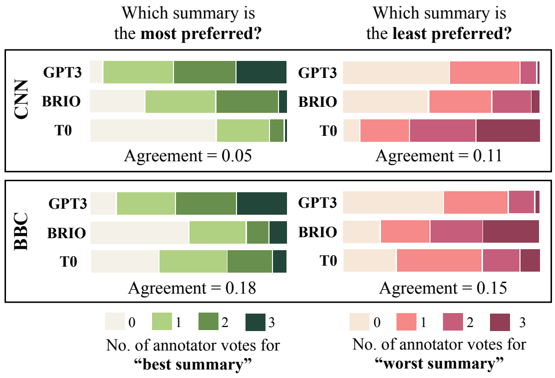


Figure 5: Annotator vote distribution for best and worst summaries across all datasets and models. Although GPT3-D2 is the clear winner according to majority vote, this choice is unanimous for less than 30% of the articles. This demonstrates the inherent variance in different annotators' definitions of "best summary", especially when comparing high-quality summaries from strong models.

Do annotators agree with each other? To study this, we plot the distribution of annotator votes for each summarization system and dataset in Figure 5. Additionally, we report the inter-annotator agreement, measured using Krippendorff's alpha with MASI distance (Passonneau, 2006), to account for multiple selections of best or worst summary allowed in our study design.

The vote distribution shows that although more annotators prefer GPT3-D2 summaries, this choice is only unanimous, i.e. supported by all three annotators, for less that 30% of the annotated articles. Conversely, although BRIO (or T0) summaries are less preferred than GPT3-D2 for the CNN (or BBC) dataset on aggregate, they were voted as the best summary by at least one annotator for more than 60% of the articles. This demonstrate two things: first, when comparing summaries from two strong models, the choice is inherently ambiguous (similar observations in Clark et al. (2021)). Second, these results and the diversity in the written rationales, show that there does not exist a universal definition of a "good" summary and that different summary properties appeal to different annotators. Regardless, the aggregate preference for GPT3-D2 is high enough across the board to give us confidence in its strength.

How do these results impact the field? Progress in text summarization research in the last five years has been enabled by the construction of large-scale text summarization

Dataset	Model	Overlap-Based ROUGE(1/2/L)	METEOR	BLEU	Similarity-Based BERTScore	MoverScore	QAEval EM	F1
CNN	PEGASUS	34.85/14.62/28.23	.24	7.1	.858	.229	.105	.160
	BRIO	38.49/17.08/31.44	.31	6.6	.864	.261	.137	.211
	T0	35.06/13.84/28.46	.25	5.9	.859	.238	.099	.163
	GPT3-D2	31.86/11.31/24.71	.25	3.8	.858	.216	.098	.159
DailyMail	PEGASUS	45.77/23.00/36.65	.33	12.2	.865	.308	.159	.229
	BRIO	49.27/24.76/39.21	.37	11.7	.871	.331	.175	.259
	T0	42.97/19.04/33.95	.28	8.9	.863	.290	.121	.184
	GPT3-D2	38.68/14.24/28.08	.26	6.6	.859	.248	.101	.159
XSum	PEGASUS	47.97/24.82/39.63	.36	9.8	.901	.362	.145	.221
	BRIO	49.66/25.97/41.04	.39	10.6	.901	.372	.139	.224
	T0	44.20/20.72/35.84	.34	8.0	.896	.340	.125	.208
	GPT3-D2	28.78/7.64/20.60	.19	2.2	.869	.197	.066	.119
Newsroom	PEGASUS	39.21/27.73/35.68	.39	.14	.873	.272	0.182	0.253
	BRIO	-	-	-	-	-	-	-
	T0	25.64/9.49/21.41	.20	.04	.849	.145	.080	0.125
	GPT3-D2	27.44/10.67/22.18	.22	.05	.859	.159	.089	0.142

Table 4: Performance of different summarization systems measured using reference-based automatic metrics. Across all datasets, we observe that automatic metrics report substantially worse results for GPT3-D2 summaries compared to fine-tuned models. This directly contradicts the human preference results from Section 3, demonstrating that these reference-based metrics cannot reliably compare the quality of zero-shot summaries against fine-tuned summaries.

datasets that involved scraping news articles and pairing them with any available summary-like data (Hermann et al., 2015; Narayan et al., 2018; Grusky et al., 2018). The CNN/DM dataset considers bullet points accompanying news articles as its summary. These "gold" standard summaries provided useful training signal to train impressive supervised models (Lewis et al., 2020; Zhang et al., 2020; Liu et al., 2022) and hence, their quality or alignment with human preferences was largely ignored.

We found that, despite its popularity, XSum is largely unsuitable for fine-tuning models like BRIO for realistic summarization settings. Even though a CNN/DM-trained BRIO model performed better, the results of our human study question the continued utility of hill-climbing on this dataset, as it seems users may simply prefer a different style of summary altogether. In fact, this preference for GPT3-D2 is much larger than incremental improvements reported in other human evaluation settings, e.g. improvements on XSum on the GENIE leaderboard (Khashabi et al., 2021). Furthermore, as we we will see in Section 5, the greater flexibility of GPT3-D2 compared to these systems makes it more suitable for news summarization tasks beyond generic summarization.

If a system designer collects a large-scale dataset of high-quality summaries that they wish to emulate, we believe a fine-tuned system may outperform GPT3-D2. However, better-trained models on datasets collected via "incidental" supervision are less likely to help.

4 Can current automatic metrics evaluate GPT3-D2 summaries?

Automatic metrics proposed for summarization evaluation can be broadly divided into two categories: (1) **reference-based**, that compare generated summaries against available gold summaries, and (2) **reference-free** that only rely on the input document. Here, we compare their performance at evaluating zero-shot GPT3-D2 summaries.

Experimental Setup We evaluate automatic metrics using summaries from 4 different summarization datasets, listed in Table 1. For each dataset, we construct our evaluation sets by randomly sampling 500[8] articles from the standard test split.[9] We compare the same 3 summarization systems from Section 3 in our analysis. Additionally, we also

[8]This size is chosen to give sufficient statistical power (Card et al., 2020) while keeping costs for GPT3-D2 evaluation low to enable others to compare on this subset. We outline costs in Appendix D.

[9]Note that these standard datasets were released before 2020. Therefore, it is possible that some article-summary pairs in our test set overlap with GPT3-D2's training data. However, we do not observe a qualitative difference in GPT3-D2's performance on these older articles.

Dataset	Model	Overall Quality SUPERT	Overall Quality BLANC	Factuality (QA-based) QuestEval	Factuality (QA-based) QAFactEval	Factuality (NLI-based) FactCC	Factuality (NLI-based) DAE	Factuality (NLI-based) 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.

report results using the fine-tuned PEGASUS model (Zhang et al., 2020), as BRIO fine-tuned models are not available for all datasets.

We publicly release this corpus of summarization outputs to standardize the test sets and support future research into GPT3-D2 based summarization. Link: https://tagoyal.github.io/zeroshot-news-annotations.html.

4.1 Reference-based metrics

Here, we study if the gold summaries of the standard datasets are useful for evaluation, especially when evaluating zero-shot summaries that are not trained to emulate the gold. We benchmark the performance of 3 different summarization metrics: (1) **overlap-based** metrics, specifically ROUGE (Lin, 2004) METEOR (Banerjee and Lavie, 2005) and BLEU (Papineni et al., 2002). (2) **similarity-based** metrics, that compute similarity between embeddings representations of generated and reference summaries. Specifically, we report BERTScore (Zhang* et al., 2020) and MoverScore (Zhao et al., 2019). (3) a **QA-based** metric, specifically QAEval (Deutsch et al., 2021). Although most QA-metrics are reference-free (discussed in Section 4.2), QAEval uses the reference summaries to indicate saliency. We report both exact match (EM) and F1 components of QAEval.

Results Table 4 outlines the results. It shows that BRIO and PEGASUS models, fine-tuned to emulate the reference summaries, outperform GPT3-D2 summaries according to all reference-based automatic metrics. The difference in their assigned scores is very high, e.g. >7 ROUGE-L points between GPT3-D2 and BRIO. For comparison, these reported scores for GPT3-D2 are even lower than the trivial Lead-3 baseline reported in prior work (Fabbri et al., 2021; Grusky et al., 2018). This clearly demonstrates that **current automatic reference-based metrics cannot be used to reliably measure summary quality under the zero-shot prompting paradigm**.

Amongst prompting-based models, we observe that T0 summaries report better metric scores than GPT3-D2 for all datasets except Newsroom. Interestingly, out of the four datasets evaluated here, Newsroom is the only one not used to train the T0 model. This further shows that access to dataset-specific reference summaries during training improves performance according to these metrics, rendering them unsuitable for zero-shot evaluation.

4.2 Reference-free metrics

Next, we investigate whether current reference-free evaluation metrics reflect the human preference rankings between summarization systems, as observed in Section 3. Here, we study 2 categories of metrics: (1) **quality metrics**, specifically SUPERT (Gao et al., 2020), which evaluates generated summaries against automatically identified salient sentences in the input, and BLANC (Vasilyev et al., 2020), which evaluates summaries on language understanding tasks. We refer readers to the original papers for detailed explanation of these. (2) **factuality metrics**, that are evaluate whether generated summaries contain incorrect information with respect to the source article. We report the performance of summarization systems using two QA-based metrics: QuestEval (Scialom et al., 2021) and QAFactEval (Fabbri et al., 2022). Additionally, we also benchmark entailment-based metrics: FactCC (Kryscinski et al., 2020), DAE (Goyal and Durrett, 2020, 2021) and SummaC (Laban et al., 2022).[10] These entailment-based models are designed for classification into factual or non-factual; therefore, we use $P(\text{factual} \mid \text{article}, \text{summary})$ to score generated summaries.

Results Table 5 outlines the scores for each summarization system according to the above reference-free metrics. Ideally, we want the relative rankings of different systems according to these metrics to correspond to human preferences, i.e. GPT3-D2 > BRIO > T0 for CNN/DM[11] and GPT3-D2 > T0 > BRIO for XSum.[12]

Overall, we observe that none of the reference-free metrics we evaluate follow these trends for both CNN/DM and XSum datasets. In particular, we observe that GPT3-D2 summaries report low factuality scores (except XSum) even though we rarely found any factual errors in our qualitative analysis of its generated summaries.

Interestingly, we noticed a roughly inverse relation to abstractiveness; summarization systems that generated more abstractive summaries (see Table 2) were generally scored lower by all automatic reference-based metrics. For instance, GPT3-D2 is scored lower than BRIO by both quality metrics for all datasets except XSum; the latter is the only dataset for which GPT3-D2 summaries are less abstractive. Such shortcomings of reference-free evaluation metrics due to spurious correlations have also been studied in prior work (Durmus et al., 2022). These issues become more exaggerated when the summarization systems being compared exhibit very different properties.

[10]Exact model versions and configurations used for these are outlined in Appendix A.

[11]Although the human study in Section 3 is only run on CNN articles, the underlying fine-tuned model is same for both CNN and DM. Therefore, it we can reasonably expect it to display similar quality differences with respect to GPT3-D2.

[12]Note that while annotators were not explicitly asked to rate factuality, we instructed them to carefully check factuality and appropriately downvote non-factual summaries.

Discussion On the surface, the failure of reference-free metrics at evaluating GPT3-D2 summaries is more surprising that reference-based metrics as the later explicitly compares generated summaries with references that GPT3-D2 is not trained to imitate. Therefore, GPT3-D2 understandably scores lower than fine-tuned systems.

However, we note two different issues with reference-free metrics: (1) Some of these, e.g. FactCC and DAE, use reference summaries as positive examples to train the metric. Therefore, although "reference-free" at test time, they are still trained to reward the summary properties seen in the standard summarization benchmarks. (2) Even completely reference-free metrics, e.g. QuestEval and QAFactEval, have only been evaluated on reference-based benchmarks and fine-tuned models. Therefore, the choice of different components, such as question answering or question generation models to use, etc. has been dictated by the error space of prior fine-tuned models (Tang et al., 2022). These decisions also now need to be re-visited to incorporate GPT3-D2 evaluation; we leave this for future work.

5 Beyond Generic Summarization

Previously, we observed that GPT3-D2 models faithfully follow simple "style" instructions in the given prompts. This provides a promising direction to tackle other use cases in news summarization beyond the generic summarization task from Section 3.

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 (He et al., 2020), query-focused (Baumel et al., 2014; He et al., 2020), or aspect-focused summarization (Krishna and Srinivasan, 2018; Ahuja et al., 2022), amongst others. Here, we evaluate GPT3-D2 performance at two of these use cases.

In **keyword-based summarization**, the output summaries must succinctly summarize the input document focusing on a given keyword; these generally correspond to specific entities or events directly mentioned in the document. In contrast, the control units in **aspect-based summarization** are high-level topics that can be common across multiple similar types of documents. For e.g., for the input article in Figure 1, *Donald Trump* or *Russian*

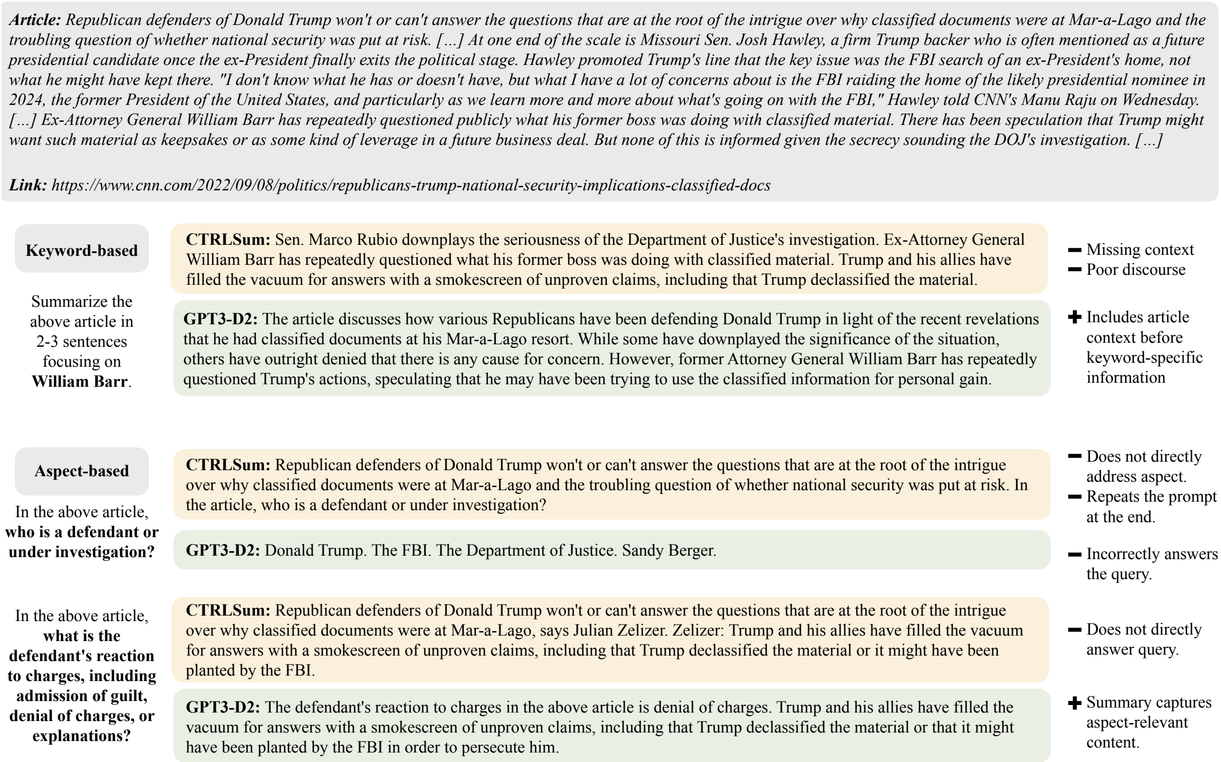


Figure 6: Comparison of keyword- and aspect-based summaries using `GPT3-D2` and `CTRLSum` models. The `GPT3-D2` prompt is shown on the left with the corresponding keyword or aspect bolded. For keyword-based summarization, the `GPT3-D2` summary presents appropriate context before the keyword-specific information. However, for aspect-based summarization, it does not always generate factually correct summaries, as shown in the first aspect example. We observe that `CTRLSum` performs poorly for both these settings.

interference in 2016 elections are keyword controls whereas *charges against the defendants* is a higher-level aspect that can serve as the query for any news article discussing a lawsuit or investigation.

5.1 Qualitative Analysis

Baseline Model for comparison We use the recently proposed `CTRLSum` (He et al., 2020), a fine-tuned `BART` model, as our baseline. It can be flexibly adapted for both keyword- and aspect-based settings by including a prompt as additional input to the encoder. We use the prompt template recommended in the original paper.[13]

Control Units For the keyword-focused setting, we use named entities extracted from the input article as the control units. For aspect-focused summarization, we directly use the aspects introduced in the guided summarization task from TAC 2011.[14] It defined 5 broad categories of newswire articles, such as accidents and natural disasters, investigations and trial, etc., and multiple aspects for each category. For example, the *"investigations and trials"* category includes aspects such as *"who is the defendant or under trial?"*, *"who is investigating, prosecuting, judging?"*, and so on.

Qualitative Analysis Figure 6 shows examples of keyword- and aspect-focused summaries using `GPT3-D2` and the baseline `CTRLSum` model. The keywords or aspects are highlighted in bold within the `GPT3-D2` prompt displayed on the left.

In this example, representative of average `GPT3-D2` quality, the keyword-focused `GPT3-D2` summary first gives a brief overview of the article setting before providing keyword-relevant information. In contrast, the `CTRLSum` summary exhibits poor discourse structure and reads like a list of facts stapled together.

The figure also shows aspect-focused summaries for two aspects associated with the "investigations and trial" category most appropriate for the chosen article. We see mixed results here for `GPT3-D2`; it generates a factually incorrect summary for the first

[13]Trained model publicly released at: `https://github.com/salesforce/ctrl-sum`.

[14]`https://tac.nist.gov/2011/Summarization/Guided-Summ.2011.guidelines.html`

aspect, listing multiple people from the input article as defendants instead of only "Donald Trump". For the second aspect, it correctly maps the high-level concept "defendant" to "Donald Trump" in the input article and generates the correct answer to the input query: *"The defendant's reaction to charges in the above article is denial of charges"*.

On the other hand, CTRLSum fails to generate aspect-focused summaries for both cases. We believe that it struggles to align high-level concepts and explicit entities in the article due to a lack of such aspect-specific examples in its training data. Instead, it generates summaries focusing on lexically similar words, i.e. "defenders" for both cases.

Based off of GPT3-D2's promising keyword-focused summarization capabilities observed above, we next conduct a human study to systematically compare it against the CTRLSum baseline. We leave further explorations of aspect-based summarization to future work, given the mixed to poor results for both models at this task.

5.2 Human Study: Keyword-focused summarization

Task Setup Similar to Section 3, we design an A/B test to compare the two models. We use the same set of 100 CNN[15] articles as Section 3. We randomly extract 2 distinct named entities from each article. In the study interface, the annotator is shown the article-keyword pair and GPT3-D2 and CTRLSum summaries corresponding to it. They are asked to select the summary that best summarizes the input article while focusing on the given keyword. Exact task instructions are included in Appendix F.

Again, we run this study using the Prolific platform. We recruit 60 participants to annotate the 100 articles; each article is annotated by 3 annotators which includes annotations for 2 separate keywords. Each annotator evaluates 5 articles.

Results Figure 7 shows the distribution of annotator votes between the GPT3-D2 and CTRLSum models. Annotators show a clear preference for GPT3-D2. In fact, for nearly 70% of all article-keyword pairs, GPT3-D2 is preferred over CTRLSum by a majority of the annotators. The main rationales given for this choice were better contextualization of keyword-related information and better coherence in GPT3-D2 summaries.

[15]We run this study using only CNN articles as the baseline CTRLSum model is trained on CNN/DM.

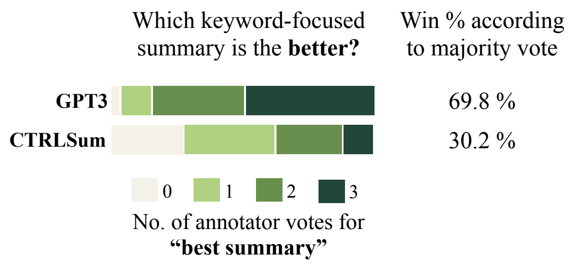


Figure 7: Distribution of annotator votes for the keyword-focused summarization task. Annotators prefer GPT3-D2 summaries over CTRLSum for approximately 70% of all article-keyword pairs, showing unanimous preference more than half the time.

Impact These results show that zero-shot GPT-3 models present a promising alternative to fine-tuned models for such specialized summarization tasks that can be easily described using textual prompts. One of the major drawbacks of fine-tuned models is that they are constrained by what data is available and how it can be transformed to create new task-specific training data. CTRLSum relied on the SQuAD question answering dataset (Rajpurkar et al., 2016) because the required "queries" or "questions" were unavailable at scale for summaries in standard summarization datasets. In contrast, zero-shot models are not constrained by the availability of task-specific data and can flexibly adapt to new tasks. Future research should focus on further exploring these capabilities and possible improvements on currently "unsolved" tasks such as aspect-based or plan-based summarization.

6 Discussion and Related Work

In recent years, research in text summarization (Rush et al., 2015; Nallapati et al., 2016; See et al., 2017; Lewis et al., 2020; Zhang et al., 2020; Liu et al., 2022) has typically relied on comparisons with gold test sets for evaluation, possibly augmented with reference-free metrics for dimensions like factuality. This paper shows that **all these metrics are completely ineffective at evaluating GPT-3 summaries**. Although issues with these metrics, particularly low correlation with human judgments, have also been studied earlier (Fabbri et al., 2021; Deutsch and Roth, 2021), they are considered reliable when comparing systems in different score ranges (Peyrard, 2019; Deutsch et al., 2022). However, GPT-3 challenges these established practices and evaluation protocols, and poses an urgent need for better evaluation.

This brings us to manual evaluation, generally considered to be the gold standard for generation evaluation. The majority of summarization research now reports results from a human study in addition to automatic metrics, but there is a general lack of consensus on what dimensions to evaluate, task design, and other factors (Hardy et al., 2019). This presents difficulties in conducting reliable and reproducible comparisons between systems (Karpinska et al., 2021), another factor contributing to the popularity of automatic metrics. Although recent efforts like GENIE (Khashabi et al., 2021) have taken steps to standardize manual evaluation protocols across systems, its annotation is not universally affordable and the quality is not strictly monitored. We hope that future work addresses these challenges and democratizes human evaluations.

The ultimate test of summarization systems is with actual users using the systems in practice. Jones (2007) discusses the need to align task formulations with actual applications scenarios ("purpose factors"). However, the research in text summarization until now has been constrained to certain problems or domains by the heavy dependence on large-scale training data: for example, producing a bullet-point summary of a news article has emerged as standard due to availability of data from CNN, not because it is shown to be the best way to present information.

Now, the success of zero-shot models like GPT3-D2 can allow realistic use-cases to drive research in a more top-down way. We already show that GPT3-D2 improves upon prior keyword-focused summarization systems that were trained on artificially adapted training data. In future research, we are interested in tackling other real world use cases, such as update summarization and plan- or aspect-based summarization. Additionally, adapting GPT3-D2 to documents longer than the allowed context, or structured inputs such as tables, presents research challenges beyond the current capabilities of GPT-3 and would be interesting to study.[16]

7 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.

[16] We very briefly discuss long document summarization with GPT-3 in Appendix E.

References

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECTNEWS: Aspect-oriented summarization of news documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Tal Baumel, Raphael Cohen, and Michael Elhadad. 2014. Query-chain focused summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 913–922.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, and Pengfei Liu. 2020. Metrics also disagree in the low scoring range: Revisiting summarization evaluation metrics. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5702–5711.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052, Seattle, United States. Association for Computational Linguistics.

Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

Esin Durmus, Faisal Ladhak, and Tatsunori B Hashimoto. 2022. Spurious correlations in reference-free evaluation of text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1443–1454.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alexander R Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards new frontiers in unsupervised evaluation metrics for multi-document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354.

Tanya Goyal and Greg Durrett. 2020. Evaluating factuality in generation with dependency-level entailment. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3592–3603.

Tanya Goyal and Greg Durrett. 2021. Annotating and modeling fine-grained factuality in summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1449–1462.

Tanya Goyal, Jiacheng Xu, Junyi Jessy Li, and Greg Durrett. 2022. Training dynamics for text summarization models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2061–2073.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. Highres: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. CTRLsum: Towards generic controllable text summarization. *arXiv*.

Pengcheng He, Baolin Peng, Liyang Lu, Song Wang, Jie Mei, Yang Liu, Ruochen Xu, Hany Hassan Awadalla, Yu Shi, Chenguang Zhu, et al. 2022. Z-Code++: A pre-trained language model optimized for abstractive summarization. *arXiv preprint arXiv:2208.09770*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems*, 28.

Karen Spärck Jones. 2007. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using mechanical turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel S. Weld. 2021. Genie: A leaderboard for human-in-the-loop evaluation of text generation. *ArXiv*, abs/2101.06561.

Kundan Krishna and Balaji Vasan Srinivasan. 2018. Generating topic-oriented summaries using neural attention. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1697–1705.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346.

Wojciech Kryscinski, Nazneen Fatema Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir R Radev. 2021. BookSum: A collection of datasets for long-form narrative summarization.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What makes In-context Learning Work? . *arXiv preprint*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Rebecca J Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations*.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yahvuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2022. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *arXiv preprint arXiv:2205.12854*.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. SummN: A multi-stage summarization framework for long input dialogues and documents: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan, and Dragomir Radev. 2021. An exploratory study on long dialogue summarization: What works and what's next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4426–4433.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 563–578.

Yao Zhao, Mohammad Saleh, and Peter J Liu. 2020. SEAL: Segment-wise extractive-abstractive long-form text summarization. *arXiv preprint arXiv:2006.10213*.

A Implementation Details

Prompts Used To generate GPT3-D2 summaries for all experiments in this paper, we use the standard prompt format outlined in Section 2. We set $N = 3$ for CNN and DailyMail, $N = 2$ for Newsroom, and $N = 1$ for XSum/BBC. For the latter, the prompt is slightly modified to *"Summarize the above article briefly in 1 sentence."*

For T0, we use the following prompts: a) CNN/DM: *"Summarize the article below in 3 to 4 sentences?"*, b) Newsroom: *"Summarize the article below in 2 to 3 sentences?"*, and c) XSum/BBC: *"Summarize the article below in 1 sentence?"*

Factuality Metrics In Section 4.2, we evaluated several recently proposed factuality metrics. We note that multiple versions have been released for some of these models in recent years. Here, we specify the versions used in our experiments to ensure reproducibility of results:

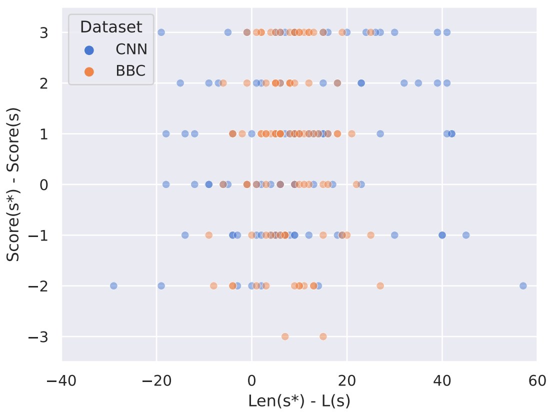


Figure 8: Correlation between summary length and annotator score (computed as the no. of "best summary" votes. For each example, plot the difference in length (x-axis) and annotator score (y-axis) between the GPT3-D2 summary and the next best system's summary.

1. **QuestEval**: We use version 0.2.4 of the questeval python package and report numbers using the precision-only setting.

2. **DAE**: We use the updated version of the DAE model trained for document-level factuality. Latest code and model released at https://github.com/tagoyal/factuality-datasets.

3. **SummaC**: We use the SummaC-Conv model (model_name = 'vitc') and sentence-level granularity in our experiments.

Keyword-based data For our keyword-based human study, we extracted two named entities per article, as discussed in Section 5. In practice, we constrained the first keyword to be lead-biased, i.e. it was extracted from the first three sentences of the article, and the second keyword was extracted from the remaining article. As CNN-based summarization models are generally lead-biased, this allowed us to benchmark models under both settings.

B Are annotator judgments of quality correlated with length?

In Section 3, results of the human study showed that annotators provide shorter length as one of the main reasons for selecting T0 summaries as the worst for the CNN dataset. Here, we investigate if the choice between GPT3-D2 and BRIO is similarly influenced by their length differences; GPT3-D2 summaries are on average 9 words longer.

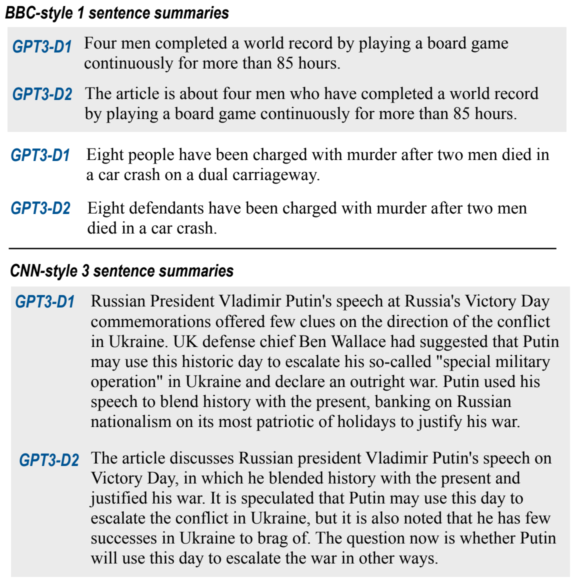


Figure 9: Examples of generated summaries using the text-davinci-001 (GPT3-D1) and text-davinci-002 (GPT3-D2) versions. The figure shows both BBC and CNN-style summaries.

To study this, we plot the difference in summary length against the difference in annotator score (measured as the no. of votes for a summarization system) between the best summarization system (GPT3-D2) and the next best system (BRIO for CNN and T0 for BBC). The resulting plot is shown in Figure 8. In general we observe low correlation between these; Pearson's ρ is 0.17 for CNN and .02 for the BBC dataset. These correlation values cannot solely explain the large differences in annotator judgments reported in the human study results of Section 3; additional quality factors must have influenced this choice. Anecdotally, we observe that the GPT summaries are slightly less information dense; our impression is that these contain a similar level of information content, but are easier to read and understand despite being a bit more verbose.

C Qualitative differences between GPT-3 versions

Figure 9 shows examples comparing summaries from text-davinci-001 (GPT3-D1) to those from GPT3-D2. For BBC-style single sentence summaries, we observed that the two models generated very similar summaries with high content and lexical overlap. More variance is observed for CNN-style summaries. In our anecdotal assessment, GPT3-D1 generated more detailed summaries while those from GPT3-D2 are less information dense.

Input Article: https://www.cnn.com/2021/09/07/opinions/covid-19-good-and-bad-news-ranney/index.html

Common prompt for all segments:

Summarize the above article briefly in 2-3 sentences.

Summary for input segment 1: The article discusses how the current situation with Covid-19 cases rising again is similar to last year in some ways, but different in others. One big difference is that there are now vaccines available which decrease the risk of severe illness.

Summary for input segment 2: The article discusses how the current situation with Covid-19 cases rising again is similar to last year in some ways, but different in others. One big difference is that there are now vaccines available which decrease the risk of severe illness. The article advises that everyone should take precautions such as wearing masks and opening windows to let fresh air in.

Summary for input segment 3: The article discusses the challenges of living through the pandemic and the importance of science and community. It also highlights the need for patience, hope, and humility as we enter the next phase of the pandemic.

Figure 10: Illustrative example of `GPT3-D2` summary of a long source article generated using the segment-then-summarize pipeline.

D Human study and API costs

At the time of running our experiments, `GPT-3` API's text-davinci-002 version was priced at $0.06 per 1K tokens. New pricing information is available at: `https://openai.com/api/pricing/`.

In our experiments, we generated around 2600 `GPT3-D2` summaries across all experiments in Section 3 (human study), Section 4 (evaluation of metrics) and Section 5 (keyword-based human study). We spent a total of approximately $150 on API requests.

For the human study, we paid participants $4 per task (each task involved annotation for 5 articles). On average, this translated to $11/hr of work. The combined cost for the generic summarization (Section 3) and the keyword-based summarization (Section 5) studies was $1020, including platform costs and bonus payments.

E Long document summarization using `GPT3-D2`

Summarization of long documents has attracted significant interest in recent years (Cohan et al., 2018; Kryscinski et al., 2021). Here, we study how a naive zero-shot application of `GPT-3` performs at long-document summarization.

First, we extract text from a long input article from the CNN website.[17] Next, we follow the commonly used segment-then-summarize procedure from prior work (Zhao et al., 2020; Zhang et al., 2022). We divide the input article into 3 disjoint segments, summarize each segment separately and concatenate these outputs to form the final summary.

Figure 10 shows the prompt used and the generated summaries for each segment. While individual segment summaries are high quality, we can see that the concatenated summary is not coherent and includes repeated "introductory" sentences outlining similar content. Related to this, it also does not cover all important aspects of the input article as a majority of its 'length budget' is spent on a high-level overview. We also observed that the generated summaries for long documents often focus on less unimportant parts of the document, e.g. *"...everyone should take the precaution of ... opening windows to let the fresh air in"* in the illustrated example. This is, in part, due to the segmentation of the input article: `GPT3-D2` still exhibits some lead bias and treats the beginning of the input segment as more salient. Therefore, the exact segmentation of the article also dictates the quality of the final summary, and cannot be readily fixed by altering the prompt.

These observations show that while `GPT3-D2` produces superior segment-level summaries, it is more difficult to adapt it to "non-natural" text inputs without fine-tuning. Therefore, techniques that have shown promising results for fine-tuned models, e.g. segment-then-summarize or extract-then-abstract (Zhang et al., 2021) approaches, are not as effective when directly applied with zero-shot models.

F Task Instructions

Task instructions provided to crowd annotators for the generic summarization task setting are shown in Figure 14 and those for the keyword-based setting are shown in Figure 15.

G Examples of generated summaries

We show examples of generated summaries for articles for generic summarization for CNN-2022 and BBC-2022 in Figures 11 and 12. It includes summaries from the 3 different summarization models evaluated in the human study in Section 3.

Examples of keyword-focused summaries are shown in Figure 13 for CNN. It includes summaries generated by `GPT3-D2` and `CTRLSum` models.

[17] Article link: `https://www.cnn.com/2021/09/07/opinions/covid-19-good-and-bad-news-ranney/index.html`

Input Article	Generated Summaries
(CNN) Toronto Police fatally shot a man who was seen carrying a firearm near three schools in the Scarborough area Thursday, police said. Officers responded to reports of a man carrying a firearm and "there was an interaction" between officers and the man, according to the Special Investigations Unit in the Canadian province of Ontario. At some point during the interaction, two officers shot at the suspect, hitting him at least once, police said. The suspect was pronounced dead at the scene. At least four schools in the area were under lockdown Thursday as authorities tried to determine the extent of the threat, according to a series of tweets from the Toronto District School Board. The incident comes as the US is grappling with the shock of two recent mass shootings: its deadliest school shooting in nearly a decade in Uvalde, Texas, and a racist shooting at a supermarket in Buffalo, New York. "We certainly understand the trauma and how traumatic this must have been for staff, students and parents given the two recent events that have happened in the United States," said Toronto Police Chief James Ramer at a news briefing after the incident, referencing the Uvalde and Buffalo shootings. He added that he understands the community's concern as the armed suspected was very close to schools. The schools under lockdown included William G. Davis Junior Public School, Joseph Howe Senior Public School, Charlottetown Junior Public School and Sir Oliver Mowat Collegiate Institute, according to a tweet from the school board. "I was really scared, I just started praying," one student told CNN affiliate CTV. Ontario's Special Investigations Unit is now investigating the fatal shooting, according to a news release. "Four investigators and three forensic investigators have been assigned to the case," the release said. Ontario premier Doug Ford tweeted his thanks to police and emergency services for their "quick action." "Thank you to police and emergency services for your quick action today in Scarborough," he said. "We're extremely grateful for everything you do to keep our communities safe."	**BRIO:** Suspect was seen carrying a firearm near three schools in the Scarborough area of Toronto, police say. At least four schools are under lockdown in the area. The incident comes as the U.S. is reeling from two recent mass shootings. The suspect was shot and killed by police. **T0:** NEW: "I was really scared, I just started praying," one student says. Shooting comes after two mass shootings in the United States. The gunman was spotted near three schools in the Scarborough area. **GPT3-D2:** The article discusses a police shooting that took place near several schools in Scarborough, Canada. The suspect was shot and killed by officers after an interaction during which the suspect was seen carrying a firearm. The incident is being investigated by the Special Investigations Unit.
(CNN) The owner of seven nursing homes across Louisiana that evacuated residents to a warehouse as Hurricane Ida approached last year has been indicted on felony charges after seven residents died at the temporary shelter, officials said. Bob Glynn Dean was arrested and charged with eight felony counts of cruelty to persons with infirmities, five felony counts of Medicaid fraud and two felony counts of obstruction of justice, according to a Wednesday news release from Louisiana Attorney General Jeff Landry. Five of the seven deaths at the warehouse shelter were considered storm-related, state health officials said. In total, more than 800 residents were taken to the facility ahead of the storm. A joint investigation by the attorney general's Medicaid Fraud Control Unit (MFCU) and the Louisiana Bureau of Investigations (LBI) "revealed Dean refused to move his residents out of the warehouse following Hurricane Ida, billed Medicaid for dates his residents were not receiving proper care, and engaged in conduct intended to intimidate or obstruct public health officials and law enforcement," the AG's news release said. Dean's attorney, John McClindon, told CNN Wednesday Dean plans to plead not guilty to all charges he's facing and said Dean's mental health will factor into the case. "I don't think it's any secret that Bob's mental health is going to be an issue in this case," McClindon said. "Bob clearly has some cognitive impairments and did on the day of this incident." McClindon said Landry informed him of an arrest warrant for Dean Monday and his client flew from Georgia to self-surrender. He made an initial court appearance in Louisiana Wednesday and was released on $350,000 bond, McClindon said. McClindon called the charges "very defensible" and said "the evidence will bear out eventually." Residents kept in 'unsafe, unsanitary, and unhealthy' conditions, officials said The nursing home residents were taken to the warehouse in Independence, about 57 miles east of Baton Rouge, ahead of Hurricane Ida's landfall on August 29. The state health department said it soon started to hear about deteriorating conditions at the warehouse. CNN obtained the logs of 61 calls from the warehouse to 911 operators. At least 30 of the calls asked for assistance with medical episodes before and after landfall, including calls for seizures, stopped breathing, and one instance in which a caller says a diabetic patient needed transport because they had "not eaten due to them having no more supplies." "Let's be clear; there is no emergency-preparedness plan that allows for residents to be kept in such an unsafe, unsanitary, and unhealthy condition," Stephen Russo, director of legal, audit and regulatory affairs for the health department, said last year. "The lack of adequate care for these residents is inhumane, and goes against the rules, regulations, and applicable statutes." The seven facilities involved had their licenses revoked and cannot repatriate or admit residents, officials said at the time. The homes also had their Medicaid provider agreements terminated, the health department said. The Attorney General's Office investigation is ongoing and additional legal action may be filed in the future, the Wednesday release said. The next court date for Dean has not been set, but McClindon said it will most likely happen in the next 60 days.	**BRIO:** Owner of seven nursing homes evacuated residents to a warehouse as Hurricane Ida approached last year. Seven residents died at the temporary shelter. Bob Glynn Dean's attorney says he plans to plead not guilty to all charges. Five of the seven deaths at the warehouse shelter were considered storm-related, officials say. **T0:** Bob Glynn Dean faces eight felony counts of cruelty to persons with infirmities. Dean's attorney says his client's mental health will be an issue in the case. Seven nursing home residents died at the warehouse shelter during Hurricane Ida. **GPT3-D2:** The owner of seven nursing homes in Louisiana has been indicted on felony charges after seven residents died at a temporary shelter during Hurricane Ida. The owner, Bob Glynn Dean, is facing eight felony counts of cruelty to persons with infirmities, five felony counts of Medicaid fraud, and two felony counts of obstruction of justice. Dean's attorney says Dean's mental health will be a factor in the case.
(CNN) Global leaders and defense officials had spent weeks speculating what Russian President Vladimir Putin might reveal about his Ukraine plans in a speech at Russia's Victory Day commemorations Monday. They'll have to keep guessing -- the leader offered few clues on the direction of the conflict. UK defense chief Ben Wallace had suggested that Putin may use this historic day to escalate his so-called "special military operation" in Ukraine and declare an outright war. Even if that had been Putin's plan, he was unlikely to follow through after Wallace's comments, not wanting to appear to his Western foes as such an easy nut to crack. Instead, the Russian president used his speech to blend history with the present, banking on Russian nationalism on its most patriotic of holidays to justify his war. In his reverence for Soviet war heroes who helped defeat Nazi Germany in World War II -- the reason Russia celebrates Victory Day -- Putin referred to new Nazi threats in Ukraine, repeating his baseless justification for the invasion as an operation to "denazify" the nation. In reference to the threat of NATO troops in Europe, Putin said: "Everything indicated that a clash with neo-Nazis, Banderites [Ukrainian nationalists], on whom the United States and their younger partners counted on, would be inevitable." "Danger was increasing every day. Russia repelled this aggression in a preventative way. This was the only correct decision, and it was a timely decision. The decision of an independent, sovereign and powerful nation," he said. Putin had few other options than to use his speech to keep selling his war to his own people. He has so few successes in Ukraine to brag of, after all. All he can do now is to keep Russians on his side as they suffer the economic hardship of crippling sanctions and international isolationism. The question now is whether Putin will use this day -- or this week even -- to escalate the war in other ways. There are growing concerns that Russian forces will turn again to standoff weapons -- aerial strikes and long-range missiles, for example -- that can be fired from afar, as they so often do when they are on the backfoot. That's worrying, as those attacks are indiscriminate and tend to cause huge civilian tolls. A bombing of a school in Luhansk, eastern Ukraine, that is feared to have killed at least 60 sheltering people over the weekend is just one example. After Russia's failure to take territory in Ukraine's north and around the capital, Kyiv, it's struggling even in the east and south, where it has had a presence through pro-Russian rebels for years. The possibility that Russia may win nothing, or very little, in Ukraine is real. Whether anything changes this Victory Day or not, a new chapter in the war will inevitably need writing soon.	**BRIO:** Russian President Vladimir Putin gave a speech at Russia's Victory Day commemorations on Monday. Peter Bergen says Putin gave few clues about his Ukraine plans in his speech. He says Putin used his speech to blend history with the present to justify his war in Ukraine. The question now is whether Putin will escalate the war in other ways, he says. **T0:** Vladimir Putin offered few clues about the direction of the conflict in Ukraine. Putin used his speech to blend history and the present to justify his war. **GPT3-D2:** The article discusses Russian president Vladimir Putin's speech on Victory Day, in which he blended history with the present and justified his war. It is speculated that Putin may use this day to escalate the conflict in Ukraine, but it is also noted that he has few successes in Ukraine to brag of. The question now is whether Putin will use this day to escalate the war in other ways.

Figure 11: Examples of generated summaries for the CNN-2022 dataset using 3 different summarization systems.

Input Article	Generated Summaries
Four men have completed what they hope will be a world record by playing a board game continuously for more than 85 hours. The foursome, of Gloucester, completed their mammoth challenge on Monday night with just a couple of hours of sleep. Lea Poole, Dale Poole, Adam Bircher and Luke de Witt Vine played the game Dune, watched by supporters and an online audience. The world record attempt was carried out in aid of Alzheimer's Research UK. The rules of the challenge stated the men were allowed to accumulate five minutes respite for every hour they played. They had to beat the previous world record, held by a team in The Netherlands, which played a board game for 80 hours in January 2017. Guinness World Records stipulated they would need to play for at least five more hours to be considered for a world record. The men's attempt has yet to be ratified. Alzheimer's Research UK was selected to benefit from the challenge as Lea and Dale Poole's father suffers from the condition. Dale Poole said: "It's a bit of a rollercoaster really. It's very flattering. People have been sending in their support and donations and it's really humbling." The team were not allowed full meals, but just snacks throughout the attempt, which included 79 games in total. "You can be within one hour quite wide awake and alert, and by sitting down and not having to think for a few seconds, that's it, you're now tired and there's no way of breaking that weird cycle," said Lea Poole. The men completed the challenge at 23:00 GMT on Monday but decided to carry on and finish the game they were on, sipping on champagne to celebrate. Mr Burcher said he was "tired, beyond measure" but "really quite satisfied and pleased with how far we've come". "If you'd told me last year I'd be doing something like this I would have called you crazy," he added. Mr De Witt Vine said it had been a "difficult and trying experience but I'm glad we've done it." The group have raised almost £1,000 for Alzheimer's UK, as well as a smaller amount for Herefordshire Mind, a mental health charity.	**BRIO:** A team of four men have set a new world record by playing a board game for 90 hours. **T0:** Four men have broken the world record for the longest time spent playing a board game. **GPT3-D2:** The article is about four men who have completed a world record by playing a board game continuously for more than 85 hours.
A man standing trial for the murder of his partner's three-year-old son had convictions for common assault, burglary and possession of heroin with intent to supply, a jury has heard. Kemarni Watson Darby suffered fatal abdominal injuries at his mother's flat where he lived with Nathaniel Pope. Alicia Watson, 30, from West Bromwich, is also on trial for murder. Mr Pope, 32, from Wolverhampton, and Miss Watson both deny that charge and another of child cruelty. During his cross-examination, Miss Watson's QC, Charles Sherrard, told the court Mr Pope had been sentenced to four months' imprisonment for common assault in May 2011, 12 months for burglary in September 2011, and 20 months for possessing heroin with intent to supply in July 2013. Mr Pope conceded he had "just told some of the truth" about himself in evidence heard last week. Mr Sherrard referred to another arrest in March 2021. He said: "On the journey to Oldbury custody suite and on arrival at Oldbury custody suite you were asked by police officers several times if you had anything secreted on your person that you needed to present to them. "Each time you said to them 'no'. The custody suite police officers were authorised to conduct a strip search of you. "Two police officers then took you into a room and asked again if you had any drugs secreted on your person. Once again, you said 'no'. "A rock of 9.16 grams of crack cocaine was found in a white tissue secreted between your buttocks." Mr Sherrard also referred to Mr Pope's evidence last week that he was "merely a recreational user of cannabis". He asked: "(In) 2013 you're a dealer of heroin and in 2021 it's crack cocaine. Are you saying that in between we should read into it there would have been nothing of the sort?" Mr Pope answered: "There was nothing of the sort." The trial continues.	**BRIO:** A man accused of murdering a toddler had a string of convictions, Wolverhampton Crown Court has heard. **T0:** A man accused of murdering a toddler in Wolverhampton had convictions for assault and drugs offences, a court has heard. **GPT3-D2:** The man on trial for the murder of his partner's 3-year-old son had a history of convictions for assault, burglary, and drug dealing.
A yellow warning for snow and ice has been issued for much of the North East. The Met Office warning is due to come into force at 20:00 BST on Wednesday and will run until 10:00 on Thursday. There will probably be icy patches on untreated roads and paths with higher elevation roads will be "affected by snowfall", the Met Office said. Motorists are being urged to take care. The cold snap comes days after the region basked in warm sun and highs of 20C (68F). The weather warning will cover Northumberland, County Durham, Tyneside, Darlington and Teesside. The wintry weather is expected to last until the weekend when slightly warmer temperatures will come in from the west, bringing unsettled conditions.	**BRIO:** Parts of the UK are set to be hit by icy conditions, forecasters have warned. **T0:** The wintry weather is set to return to the North East of England on Wednesday night into Thursday morning. **GPT3-D2:** The Met Office has issued a warning for snow and ice in the North East, which is expected to last until the weekend.

Figure 12: Examples of generated summaries for the BBC-2022 dataset using 3 different summarization systems.

Input Article

A coalition of thousands of Etsy sellers signaled support for a one-week strike starting on Monday — the same day the online marketplace known for its unique handicrafts will start hiking the fees it charges those who use its platform to earn a living. An online petition started by Etsy (ETSY) shop owner Kristi Cassidy urging the company to cancel the fee increases — which tick up from 5% to 6.5% starting Monday — has garnered nearly 50,000 signatures. Of those signatories, some 18,500 come from people who have identified as Etsy sellers who support the strike, according to Etsy shop operator and strike participant Mattie Boyd. "We feel like we deserve a seat at the table," Boyd told CNN Business. "And we hope these demands are met, that's our immediate goal. But, generally, there's got to be some kind of change, where there's some kind of dialogue, or Etsy sellers have some kind of representation where these decisions are being made." Sellers participating in the strike are putting their shops on "vacation mode" for a week starting Monday, according to Cassidy's petition, a temporary setting that lets users essentially put their Etsy shop on hold for a designated period of time. Etsy CEO Josh Silverman announced the fee increases in a memo to sellers in late February. The letter touted Etsy's massive growth over the past two years, boasting how active sellers last year increased their sales by "23% on average compared to 2019, and in 2021 alone, we showed more than 90 million active buyers worldwide that there's an alternative to big-box, automated shopping." Silverman then announced plans to "make significant investments in marketing, seller tools, and creating a world-class customer experience so we can continue this tremendous growth." "To support this goal, on April 11 we will increase our current 5% transaction fee to 6.5%," Silverman wrote. Etsy is the main source of income for Boyd, who operates a shop via the online retailer featuring homemade graphic T-shirts and other "niche" items that Boyd says are "geared towards people who are members of the queer and trans community, and who are also into punk rock and metal." Demands listed in the petition include canceling the fee increases passed onto sellers; creating a comprehensive plan to crack down on "reseller" shops (people selling mass-produced goods that they have not designed themselves); improve and expedite the support systems for sellers who have had their business disrupted by Etsy's automated tools; end the "Star Seller" program that Etsy uses to rate sellers; and to let sellers opt out of offsite ads for their products. Boyd said many sellers felt like they weren't given a fair way to give feedback on the sudden fee-hike announcement, which marks the first increase since 2018. While the other demands in the petition are issues Boyd said have been brewing for a while, "I think for a lot of us it was that 30% increase [to fees] that really lit the match," Boyd added. "The strike officially is meant to go from April 11, today, through April 18," Boyd told CNN Business. "But people are being encouraged to participate for as much time as they feel like they can, and no one's being shamed if they can't do the whole week." With sellers spread out across the country, Boyd said organizers are using a slew of tech tools to rally together and support each other — including Reddit channels, Discord chats, and Instagram. Raina Moskowitz, the chief operating officer for Etsy, told CNN Business in a statement via email on Monday that "sellers' success is a top priority for Etsy." "We are always receptive to seller feedback and, in fact, the new fee structure will enable us to increase our investments in areas outlined in the petition, including marketing, customer support, and removing listings that don't meet our policies," the statement added. "We are committed to providing great value for our 5.3 million sellers so they are able to grow their businesses while keeping Etsy a beloved, trusted, and thriving marketplace." The seller strike on Etsy notably comes amid a wave of workplace activism seen at a slew of major companies over the past year -- from Starbucks (SBUX) to Amazon. Earlier this month, Amazon (AMZN) workers in Staten Island, New York, voted to form the e-commerce giant's first-ever US labor union in a landmark election. Amazon has since filed an appeal, calling for a do-over of the entire vote.

Generated Summaries

Keyword: Kristi Cassidy

CTRLSum: Etsy (ETSY) shop owner Kristi Cassidy started a petition urging the company to cancel the fee increases. Sellers participating in the strike are putting their shops on "vacation mode" for a week starting Monday. The strike comes amid a wave of workplace activism seen at a slew of major companies.

GPT3-D2: Kristi Cassidy is an Etsy shop owner who is organizing a one-week strike against the company starting on Monday. The strike is in response to Etsy's decision to raise the fees it charges sellers, which will increase from 5% to 6.5%. Cassidy's petition urging the company to cancel the fee increases has garnered nearly 50,000 signatures.

Keyword: Silverman

CTRLSum: Etsy CEO Josh Silverman announced the fee increases in a memo to sellers in late February. "To support this goal, on April 11 we will increase our current 5% transaction fee to 6.5%," Silverman wrote. "We feel like we deserve a seat at the table," Etsy shop operator Mattie Boyd told CNN Business.

GPT3-D2: The article discusses a one-week strike being led by Etsy sellers in response to fee increases that will go into effect on Monday. The strike was organized in response to a memo from Etsy CEO Josh Silverman announcing the fee increases. Etsy sellers are demanding that the fee increases be cancelled, among other things.

New York (CNN Business) As Russia's assault on Ukraine continues, American bar and restaurant owners are hoping a small word change will help show their solidarity with the Ukrainian people. In a move reminiscent of the "freedom fries" fad of the early aughts, they're taking Moscow Mules off the menu and replacing them with Kyiv Mules. Small American businesses, such as independent bar or restaurant owners, may not have any direct business ties to Russia, but many feel strongly about the violent attack on Ukrainian cities and citizens. Replacing "Moscow" with "Kyiv" in their vodka-ginger-lime cocktails is one way to show support for Ukraine. Bond Bar, in San Francisco, has renamed its Moscow Mule the Kyiv Mule. "It's just a little token of acknowledgment to the Ukrainian people," said owner Andrea Minoo. "We're just trying to raise awareness, and to let people know, we're in support [of Ukraine]." She wants Ukrainians to know that "we see what's happening, we wish we could do more." Bond Bar doesn't serve Russian vodka, Minoo noted, so it's not replacing any ingredients in its Kyiv Mule. Madrone Art Bar, also in San Francisco, did serve Russian vodka until this past weekend, when owner Michael Krouse decided to take it off the menu. First, he had to figure out which of the roughly 10 vodkas he carries were actually Russian. Many top-selling vodka brands that trace their origins to Russia are now distilled in multiple countries, including the United States. Stoli Vodka, for example, is actually made in Latvia, and the company's headquarters are in Luxembourg. After some research, Krouse removed Russian Standard, one of the few vodka brands that actually is Russian-made, from his bar. Then he decided to rename Madrone's Moscow Mule the Kyiv Mule and looked for a Ukrainian vodka to make it with. The bar unveiled the reconstituted cocktail on Instagram this week. "Introducing the 'Kyiv Mule' made with Prime Ukrainian vodka!,'" a Wednesday post reads, adding that "$2 of each Kyiv Mule sale will be donated to the Ukraine Crisis Fund." The Kyiv Mule costs $12. Krouse said he was feeling sad and helpless about the situation in Ukraine when he decided to make those moves. Those changes were "at least something that we could do," he said. Making a gesture Em Chamas Brazilian Grill in Kansas City, Missouri, said in a Facebook post last week that its Moscow Mule will be replaced by a "Snake Island Mule," in "support of the Ukrainian resistance and in honor of the brave soldiers of Snake Island." Snake Island, also known as Zmiinyi Island, sits about 30 miles off the southern tip of the Ukrainian mainland in the northwestern Black Sea. Last week, a handful of Ukrainian fighters on the island stood up to Russian warships. They were feared dead, but the Ukrainian Navy released a statement Monday saying the troops were "alive and well" after being forced to surrender "due to the lack of ammunition." Back in Kansas City, the Em Chamas post also urged customers against "directing animosity towards local restaurants or businesses that are Russian-themed and/or are owned/operated by those of Russian [descent]." The eatery, which charges about $11 for cocktails, plans to donate profits from the Snake Island Mule to a Ukrainian charity, according to the Facebook post. Ronnie Heckman, owner of Caddies on Cordell, a bar and grill in Bethesda, Maryland, has not only swapped out Moscow Mules for Kyiv Mules, he's gone so far to replace Black Russians and White Russians with Black Ukrainians (vodka and Kahlua on the rocks) and White Ukrainians (made with vodka, Kahlua and cream). "It's a gesture," he said. Caddies is also donating part of the revenue from those drinks to Ukrainian aid, he said. Heckman hopes that if enough restaurant and bar owners take references to Russia off their menus, they can send a message to Russian leadership. The attack on Ukraine "makes no sense," he said. "It's wrong." CNN's Jordan Valinsky, Brad Lendon, Tim Lister and Josh Pennington contributed to this report.

Keyword: Ukraine

CTRLSum: Bar owners in the U.S. are replacing "Moscow" with "Kyiv" in their vodka-ginger-lime cocktails. They're doing it to show their solidarity with the Ukrainian people. Russia's assault on Ukraine is "wrong," says Ronnie Heckman, owner of a Maryland bar.

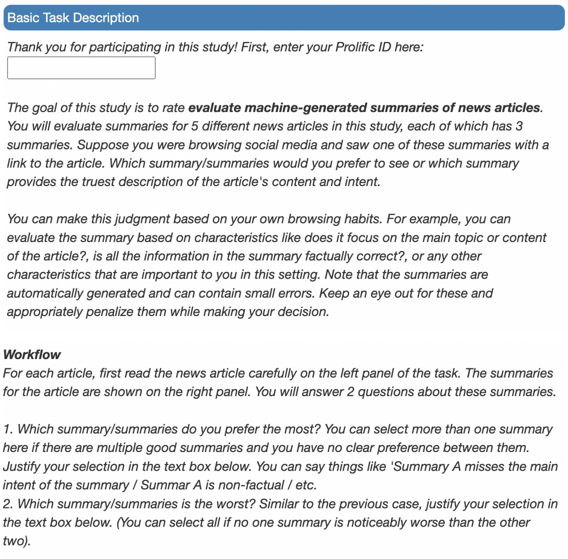
GPT3-D2: The article discusses how American bar and restaurant owners are changing the name of the Moscow Mule cocktail to the Kyiv Mule to show support for Ukraine. The Kyiv Mule is made with Ukrainian vodka and ginger-lime, and some businesses are donating part of the proceeds to Ukrainian aid. The change is meant to be a small gesture to show support for the Ukrainian people as Russia continues its assault.

Keyword: Brad Lendon

CTRLSum: Bar owners are replacing "Moscow" with "Kyiv" in their vodka-ginger-lime cocktails. CNN's Jordan Valinsky, Brad Tim Lendon, Josh Pennington contributed to this report. The attack on Ukraine "makes no sense," says Ronnie Heckman, owner of Caddies on Cordell.

GPT3-D2:The article discusses how American bar and restaurant owners are swapping out Moscow Mules for Kyiv Mules in order to show support for Ukraine. This is in response to Russia's recent assault on the country. CNN's Brad Lendon contributed to the report.

Figure 13: Examples of keyword-focused summaries for CNN articles from 2022.



Basic Task Description

Thank you for participating in this study! First, enter your Prolific ID here:

*The goal of this study is to rate **evaluate machine-generated summaries of news articles**. You will evaluate summaries for 5 different news articles in this study, each of which has 3 summaries. Suppose you were browsing social media and saw one of these summaries with a link to the article. Which summary/summaries would you prefer to see or which summary provides the truest description of the article's content and intent.*

You can make this judgment based on your own browsing habits. For example, you can evaluate the summary based on characteristics like does it focus on the main topic or content of the article?, is all the information in the summary factually correct?, or any other characteristics that are important to you in this setting. Note that the summaries are automatically generated and can contain small errors. Keep an eye out for these and appropriately penalize them while making your decision.

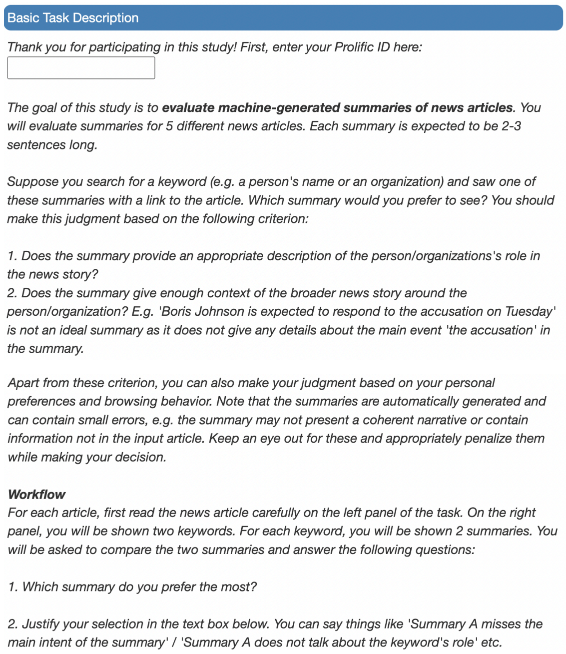
Workflow

For each article, first read the news article carefully on the left panel of the task. The summaries for the article are shown on the right panel. You will answer 2 questions about these summaries.

1. Which summary/summaries do you prefer the most? You can select more than one summary here if there are multiple good summaries and you have no clear preference between them. Justify your selection in the text box below. You can say things like 'Summary A misses the main intent of the summary / Summar A is non-factual / etc.

2. Which summary/summaries is the worst? Similar to the previous case, justify your selection in the text box below. (You can select all if no one summary is noticeably worse than the other two).

Figure 14: Screenshot of the task instructions for the generic summarization setting.



Basic Task Description

Thank you for participating in this study! First, enter your Prolific ID here:

*The goal of this study is to **evaluate machine-generated summaries of news articles**. You will evaluate summaries for 5 different news articles. Each summary is expected to be 2-3 sentences long.*

Suppose you search for a keyword (e.g. a person's name or an organization) and saw one of these summaries with a link to the article. Which summary would you prefer to see? You should make this judgment based on the following criterion:

1. Does the summary provide an appropriate description of the person/organizations's role in the news story?

2. Does the summary give enough context of the broader news story around the person/organization? E.g. 'Boris Johnson is expected to respond to the accusation on Tuesday' is not an ideal summary as it does not give any details about the main event 'the accusation' in the summary.

Apart from these criterion, you can also make your judgment based on your personal preferences and browsing behavior. Note that the summaries are automatically generated and can contain small errors, e.g. the summary may not present a coherent narrative or contain information not in the input article. Keep an eye out for these and appropriately penalize them while making your decision.

Workflow

For each article, first read the news article carefully on the left panel of the task. On the right panel, you will be shown two keywords. For each keyword, you will be shown 2 summaries. You will be asked to compare the two summaries and answer the following questions:

1. Which summary do you prefer the most?

2. Justify your selection in the text box below. You can say things like 'Summary A misses the main intent of the summary' / 'Summary A does not talk about the keyword's role' etc.

Figure 15: Screenshot of the task instructions for the keyword-based setting.