SummQual: A Dataset of Human Evaluation on Large-Scale Language Model Summarization Quality

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown impressive performance on various natural language processing tasks, including text summarization. However, evaluating the quality of the summaries generated by LLMs is challenging, as automatic metrics often do not correlate well with human judgments. In this work, we present SummQual, the largest human evaluation dataset of multi-domain summarization systems to date, featuring 6k document-summary pairs in the test set and 30k training pairs. Our dataset evaluates several state-of-the-art LLM systems, such as GPT-4, Bard, and Vicuna. Unlike most existing datasets that focus on the news domain, our dataset covers nine diverse domains: Wikipedia, News TV, Pubmed, Reddit, Youtube videos, supreme court cases, clinical dialogues, and financial reports. To avoid overlap with LLMs' training data, SummQual collects documents from the most recent public online sources, starting from the year 2023. Furthermore, this dataset contains not only common summary quality annotations, e.g., relevance and coherence, but also fine-grained human feedback on hallucinated spans. We believe SummQual can elicit a deeper understanding of LLM's summarization capability and promote research in text summarization as well as hallucination detection and mitigation.

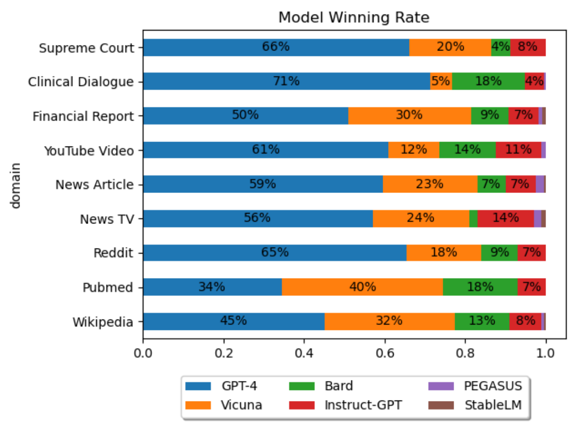

Figure 1: Percentage of each model's summary selected as the best one by annotators in SummQual test set. GPT-4 works the best in almost all domains, and Vicuna performs similarly to GPT-4 in Pubmed and Wikipedia domains.

1 Introduction

Summarization systems generate a succinct overview of a long document. The summary should be well-organized, easy to read, and concisely cover the main points of the document. Summarization has applications in many areas to save time for users when reading news, reports, forums, conversations, etc. With the rapid development of large-scale language models (LLMs) such as Instruct-GPT (Ouyang et al., 2022), GPT-4 (OpenAI, 2023), Bard[1], PaLM (Chowdhery et al., 2022), Vicuna (Zheng et al., 2023), and LLaMA (Touvron et al., 2023), one can instruct an LLM to produce high-quality summaries under zero-shot and few-shot paradigms. As summary quality steadily improves, how to evaluate summarization quality becomes a challenge, as most of these systems are capable of generating readable summaries, but that are not necessarily faithful to the original document, not informative enough to convey the key information, or containing minor errors. These issues typically require careful human judgment over different dimensions, which prompts the advent of various summarization datasets.

However, existing human evaluation datasets of summarization systems have several shortcomings. First, they heavily focus on news articles (Grusky et al., 2018; Fabbri et al., 2021; Zhang et al., 2023), while other common real-world domains such as clinical dialogue, financial reports, or online videos are less explored. Second, with the prevalence of LLMs, there still lacks a systematical human evaluation and comparison of these LLMs on the sum-

[1]https://bard.google.com/

marization task. Third, as LLMs are usually pre-trained on large-scale text data, it is possible that existing summarization datasets, such as CNN/DM dataset (Hermann et al., 2015), have already been covered in the pre-training data, undermining the validity of using these datasets for a fair comparison. To address these challenges, we construct the large-scale SummQual dataset for a comprehensive human evaluation of the summarization capability of the latest prevalent LLMs.

To construct this dataset, we first select nine popular real-world domains from which to collect document data. For the well-explored domains, such as news (Fabbri et al., 2021) and forum posts (Stiennon et al., 2020), we re-crawl the latest documents from the same source but released in 2023 to avoid overlap with LLM training data. We further explore several more challenging domains, such as Youtube Video transcripts, doctor-patient dialogues, U.S. Supreme Court conversations, financial reports, etc. In total, SummQual contains 6k documents

We select five LLMs: GPT-4, Instruct-GPT, Google Bard, Vicuna, Stable-LM, and one state-of-the-art supervised model (PEGASUS) to generate candidate summaries for each document. For the five LLMs, we produce zero-shot summaries by merely providing a general instruction. As for PEGASUS, we fine-tune it on the CNN/DM dataset.

For human annotations, we follow SummEval (Fabbri et al., 2021) and score the quality of each summary with a numeric rating from 1 to 5 for each of the five aspects: coherence, consistency, coverage, fluency, and overall. To ensure the quality of annotation, we further ask labelers to find all the inconsistent spans from the summaries that contradict the source documents (e.g., hallucinated facts).

Based on SummQual, we compare the summarization quality of various LLMs. The final win rate of each LLM is shown in Figure 1. GPT-4 produces the most favorable summary in all domains except Pubmed, followed by Vicuna, which performs similarly to GPT-4 in the domains of Pubmed and Wikipedia. We also conduct a thorough meta-evaluation of existing summarization evaluation metrics. The results show that LLM-based evaluation (Liu et al., 2023) outperforms other automatic metrics in terms of correlation with human judgments.

Overall our contributions are as follows:

1. We comprehensively evaluate the summarization ability of LLMs on 9 domains, including medical, legal, finance, etc.
2. To make a fair comparison of LLMs, we are the first to re-crawl the latest data from the year 2023 for evaluation.
3. SummQual is the largest scale human evaluation data set on multi-domain summarization systems.
4. SummQual provides fine-grained multi-aspect quality ratings and marked inconsistent spans in each summary.
5. We conduct a broad evaluation of the latest meta-evaluation metrics, such as G-Eval, UniEval, BartScore, etc. G-Eval and Vicuna trained with SummQual work the best.

2 Related Work

Summarization Meta-Evaluation Dataset is to evaluate how well the metrics on summarization evaluation align with human ratings, such as Rouge (Lin, 2004), BLEURT (Sellam et al., 2020), BARTScore (Yuan et al., 2021), UniEval (Zhong et al., 2022), GEval (Liu et al., 2023), etc. Meta-evaluation datasets need to provide human scores regarding the quality of summarization systems. NEWSROOM (Grusky et al., 2018) contains 60 news articles with summaries scored by humans regarding coherence, fluency, informativeness, and relevance. Rank19 (Falke et al., 2019) ranks the correctness of the summaries from 200 news in CNN-DM dataset. (Stiennon et al., 2020) provided 64k pairs of human comparisons on summaries of Reddit posts. SummEval (Fabbri et al., 2021) and RealSumm (Bhandari et al., 2020) collected human judgments of summarization systems on 100 documents from CNN/DailyMail dataset. Zhang et al. (2023) gave an expert review of LLM systems on 50 articles from each of the CNN/DM and XSUM evaluations. All the existing summarization meta-evaluation datasets come from a single domain, either news or forums. And many other real and challenging problems deserve to be explored, such as summarizing a video or a financial report. Our work will evaluate the summarization systems on multiple domains involving thousands of documents.

Large-scale language models (LLMs) have shown impressive summarization ability and even achieve human performance in the news domain (Zhang et al., 2023). In previous works, most

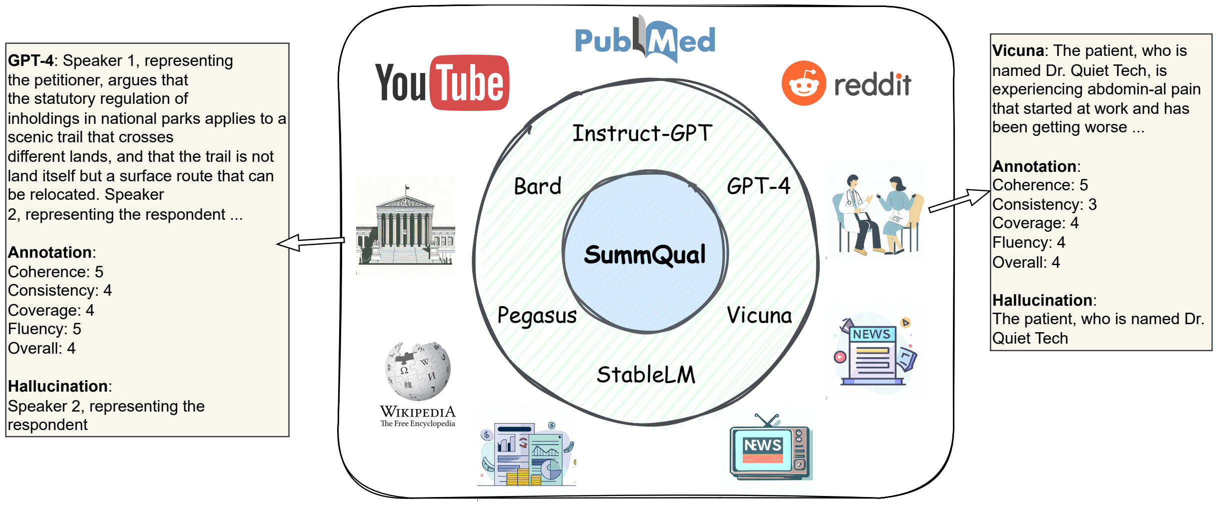

Figure 2: SummQual consists of human evaluation on the summarization ability of 5 LLMs and 1 supervised model in 9 different domains. Human labelers need to provide evaluation scores regarding the aspects of coherence, consistency, coverage, fluency, and overall. The highlighted phrases in red are the human-annotated hallucinated spans.

of the evaluations are based on supervised models (Fabbri et al., 2021), such as PEGASUS (Zhang et al., 2020), or the OpenAI models (Zhang et al., 2023), such as GPT-3 (Brown et al., 2020) or Instruct-GPT (Ouyang et al., 2022). With the rapid progress of LLM development, more powerful models come out, such as GPT-4 (OpenAI, 2023), and there is also a series of open-sourced LLM released, such as LLaMA-based models (Touvron et al., 2023). In this work, we will broadly evaluate the summarization ability of the most powerful LLMs. Due to the strong zero-shot ability of LLMs, we can now go beyond the restrictions of training data, as many domains don't have data with well-written summaries. We will use a simple and effective method to prompt LLMs on summarization. This work will extend the human evaluation of LLMs zero-shot learning ability on summarization.

3 SummQual

An overview of the data and summarization systems for annotation is shown in Figure 2, and statistical analysis of the dataset is shown in Table 1.

3.1 Data Domain

SummQual contains documents from a wide variety of domains, covering both traditionally popular summarization dataset domains such as news articles and underexplored real-world domains such as financial reports.

News Article comes from the CNN website. All the documents are well organized and written by experts. The news domain is the most widely explored area of human judgment on summarization systems (Grusky et al., 2018; Fabbri et al., 2021). The public dataset, like CNN/DM (Hermann et al., 2015), was collected in 2015, a long time ago. To ensure LLMs never see the evaluation data, we crawl news in 2023 for annotation.

News TV [2] also comes from CNN in the year 2023. We use the transcripts of the video as source documents, the structure of which is more complex than the news articles. The transcripts include not only dialogues but also some video clips inserted. It requires summarization systems to understand the structure of the document better.

Wikipedia is one of the most popular public resources for research. Most of LLMs are also pretrained on Wikipedia. We randomly pick Wikipedia pages created in 2023 as source documents to generate summaries.

Pubmed comes from full-text archive of biomedical and life science journals in 2023. As journal papers are written by professional researchers, it is hard for labelers to understand the whole paper fully. Thus we use the abstract of each paper as source documents to summarize.

Youtube Video are searched based on the keywords of hot topics, such as Games, Comedy, Education, Fitness, Food and Drinks, Travel, DIY, Tutorials, Music and Dance, etc. For each topic, we select

[2] http://transcripts.cnn.com

the top 20 relevant videos created in 2023 [3] and crawl the corresponding transcripts [4] as source documents.

Reddit is a source of user posts on a wide range of topics. We search the posts on Reddit for some hot topics [5], such as funny, gaming, NBA, trashy, pics, mildlyinteresting, Showerthoughts, etc. As Reddit data is public and widely used for LLM pretraing, we only leave the posts created in 2023 for summarization, and to keep the posts diverse, we only leave the top 10 posts for each topic.

Clinical Dialogue contains dialogues between doctors and patients. As most of the dialogues are private and not released, we searched the clinical dialogues that appeared on YouTube on hand. After confirming the video happens in a hospital / clinic, we crawl the corresponding transcript.

Supreme Court contains the transcripts of the conversations that take place in the US Supreme Court. Each case can have more than one session of oral arguments. Although the data was explored for argument mining, it hasn't been explored for summarization, due to the limitation of human summaries. We fully make of the Supreme Court Oral Arguments Corpus [6].

Financial Report[7] is a company annual report. As there are many tables in the reports, we only select the report that can be easily transferred to uniform txt format, but not PDF version. And for each page, we will ask summarization systems to give a summary. This is also a rarely touched area for summarization systems. It needs models to understand the numbers in the table better.

For all the source documents above, they may involve thousands of words, especially some transcripts. Due to the encoding length limitation of LLMs, such as 2048 tokens for LLaMA, we only use the first 700 words split by space for summarization systems. The number of annotated documents from each domain is shown in Table 1. We treat financial report, supreme court, and clinical dialogue as out-of-domain data and don't provide any training data in the domains.

[3]Search package: https://serpapi.com/integrations/python

[4]Transcript crawling package: https://pypi.org/project/youtube-transcript-api/

[5]Crawling package: https://praw.readthedocs.io/en/stable/

[6]https://convokit.cornell.edu/documentation/supreme.html

[7]https://www.microsoft.com/investor/reports/ar22/download-center/

3.2 Summarization Systems

As we are working on the summarization of multiple challenging domains, we need the model to have a strong zero-shot learning ability. Meanwhile, it should have the ability to control the output length. We hope to compare summaries with similar lengths. After exploration, we select GPT-4, Instruct-GPT (text-davinci-002), Google Bard (versions earlier than May)[8], Vicuna-7B, and StableLM-7B as LLM summarization systems. To have better understanding LLMs, we also involve one supervised summarization model, PEGASUS. As most of the domains we collected don't have human written summaries to fine-tune, we only fine-tune PEGASUS on the most widely explored dataset CNN/DM (Hermann et al., 2015)[9].

Different LLMs may generate summaries of different lengths, and annotators may leverage this to bias their ratings across systems. Thus, we design a normalization process to make LLMs generate summaries of a similar number of sentences as follows. We first ask GPT-4 to generate the summary with the following prompt: "This is a summarization system. {{domain name}}: {{document}} Summarize the {{domain name}}:", where {{domain name}} tells LLMs where the document comes from, such as news article, clinical dialogue, Reddit post, Youtube video transcript, etc., and {{document}} is the corresponding content of the document. We will use GPT-4 summary as a reference to control the summary length. Based on our observation, LLMs such as Instruct-GPT, Google Bard, and Vicuna are much more likely to follow the requirement of the number of sentences than that of the number of words. So we tokenize the GPT-4 generated summaries and count the number of sentences. Then add the sentence number requirement to the prompt of the other LLMs as follows: "This is a summarization system. {{domain name}}: {{document}} Summarize the {{domain name}} in {{sentence number}} sentences:". For PEGASUS, we don't add any restrictions but directly feed the documents into the well-trained model to generate the summaries. We show the summary length of different systems in Table 1.

3.3 Data Annotation

We hire four native English speakers for data annotation and one project manager to control the

[8]https://bard.google.com/

[9]https://huggingface.co/google/pegasus-cnn_dailymail

Domain	Court	Clinic	Finance	News	NewsTV	Wiki	Pubmed	Youtube	Reddit
				Number of documents					
Train	0	0	0	882	888	882	882	416	882
Test	148	136	164	100	100	100	100	100	100
				Number of words per document or per summary					
Document	650	505	609	419	700	394	225	544	265
Instrcut-GPT	85.7	50.6	67.8	65.9	63.3	81.3	57.6	48.9	45.1
GPT-4	114	75.1	94.9	74.6	81.3	83.6	61.9	76.0	61.8
Vicuna	131	93.3	120	88.6	76.5	114	86.2	86.6	61.3
Bard	125	57.6	84.8	57.1	60.4	82.7	69.7	53.0	48.4
StableLM	474	399	561	115	136	118	96.0	105	98.2
PEGASUS	51.1	66.3	53.1	51.1	42.0	46.9	41.8	63.2	64.0

Table 1: Statistics of SummQual. The table on the top is the number of annotated documents in 9 different domains. On the bottom is the number of words in either the source document or the summaries generated by different summarization systems.

annotation pipeline and quality. All the hiring comes from a professional data annotation company, and the whole dataset annotation takes around two months and $37k. For the test set (1k documents), each source document and all the associated summaries are independently annotated twice by different labelers. The final test set uses the averaged scores from two annotations as the human score for each document-summary pair. For the training set (5k documents), each document and the summaries are only annotated once.

During the annotation, each instance contains two parts: 1) A document that contains 50-700 words. 2) Multiple similar-length summaries generated from different summarization models. The labelers are asked to provide the following two types of annotations:

Summary Quality Scoring The labeler reads the document and summaries and then assigns a score from 1-5 (bad, fair, relatively good, good, excellent) for each summary with regard to the following criteria:

1. Coherence: the summary should be well-organized and easy to read;
2. Consistency: the summary should not contradict the information in the document;
3. Coverage: the summary should cover the main points of the document;
4. Fluency: the summary should be written in good English;
5. Overall: In general, how good the summary is.

We also ask labelers to note down the best summary, which is especially helpful when there are more than one summaries with the same highest overall score.

We follow SummEval dataset (Fabbri et al., 2021) to compute the agreement between labelers on test set. For each document, we compute the Pearson correlation coefficient between two labelers on the scores of six summaries from different summarization systems. We compute the correlation for each evaluation criterion independently. After computing the scores for all the documents, we average all the Pearson correlation coefficients by criteria as follows: coherence 0.7, consistency 0.83, coverage 0.75, and fluency 0.75, overall 0.79. Besides the Pearson correlation, we also compute the Krippendorff's alpha coefficient (Krippendorff, 2011) regarding the overall scoring, 0.69, which is on the same level as the inter-agreement of SummEval dataset.

Hallucination Annotation As hallucination has become an important issue in LLM's output (Agrawal et al., 2023), we ask labelers to highlight the spans in all the summaries that are inconsistent with the document. For example, if the document says *"President Biden will visit Europe next week"* while the summary says *"President Biden will visit Asia next week"*, then "Asia" is an inconsistent span. The factually correct spans but not based on the original document also need to be highlighted. We compute the summary level agreement on whether the summary has at least one hallucination span. Thus, for each document-summary, there would be a binary label indicating the hallucination or not. The Cohen's kappa coefficient between the hallucination annotations of two labelers on the test set is 0.5.

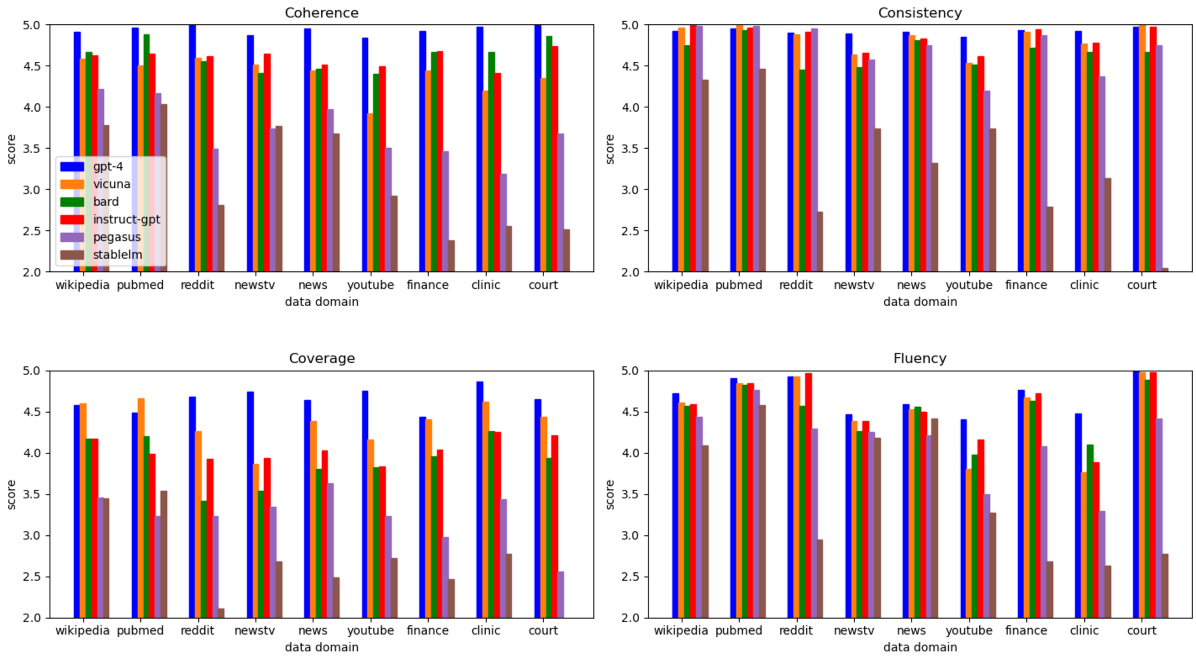

Figure 3: Human rating on different evaluation metrics on the test set.

4 Experiment and Analysis

4.1 Summarization System Analysis

Comparing Summarization Systems Given the annotators' selection of the best summary for each document, GPT-4 outperforms all other systems on all domains except the PubMed domain (Fig. 1). On the more challenging domains, such as clinical dialogue, supreme court, News TV, and YoutuTube Video, GPT-4 is selected to be the best on over 50% of the documents. For widely explored areas, such as the News Article and Reddit, GPT-4 also has a stable lead.

Besides GPT-4, we are surprised to find that Vicuna works the second best overall and even better than GPT-4 on Pubmed. Vicuna (7B) is a language model with a much smaller number of parameters than Instruct-GPT (175B), yet with higher summary quality. We attribute this great performance to the high-quality instruction data used during instruction tuning.

Vicuna is then followed by Bard, Instruction-GPT, PEGASUS, and StableLM. We hypothesize that although StableLM is initialized by LLaMA like Vicuna, it is still not well-aligned with human instructions on summarization tasks, with less fluent summaries on many domains.

And although Pegasus has been pre-trained and then fine-tuned with news data, it is still far away from LLMs, even in the news domain. This aligns with the findings of recent studies on the summarization quality of fine-tuned systems and LLMs (Zhang et al., 2023).

Different aspect metrics Besides the overall scores, all summaries in SummQual are evaluated based on different aspects, as shown in Figure 3.

For coherence, we can see that the labelers give high scores to GPT-4, with an average score of over 4.8. Bard and Instruct-GPT get second place on 4 or 5 domains out of 9. Vicuna also gets reasonable scores on different domains. StableLM and Pegasus get the worst performance on all the domains.

The consistency metric has a close relation to hallucination. As shown in Figure 3, GPT-4, Vicuna, and Instruct-GPT have almost perfect consistency scores on Wikipedia, PubMed, and Supreme Court domains. Bard and Pegasus still have space to improve. Overall, we find all models except StableLM can achieve good consistency scores across domains.

The coverage metric indicates that all models have space to improve, including GPT-4. Vicuna is in second place and can even beat GPT-4 on the domains of Wikipedia and Pubmed, where the text is cleaner and easier to understand and summarize. Note that to reduce the challenges of summarizing PubMed, we only use the abstracts as source documents to summarize. The abstracts are already good summaries of the whole papers. However, the

	Coh	Con	Cov	Flu	Overall	Court	Clinic	Finance	News	NewsTV	Wiki	Pubmed	Youtube	Reddit
BartScore	-0.129	0.263	0.037	0.043	-0.034	0.142	0.124	0.102	-0.053	-0.072	-0.141	-0.255	-0.082	-0.306
UniEval	0.356	0.275	0.436	0.077	0.364	0.393	0.429	0.375	0.381	0.376	0.342	0.110	0.377	0.448
G-Eval	0.662	0.710	0.730	0.567	0.776	0.830	**0.835**	**0.816**	**0.760**	0.722	0.678	0.698	0.739	0.829
Vicuna (FT)	**0.749**	**0.725**	**0.765**	**0.620**	**0.785**	**0.861**	0.787	0.809	0.726	**0.736**	**0.754**	**0.772**	**0.765**	**0.880**
					Hal	Court	Clinic	Finance	News	NewsTV	Wiki	Pubmed	Youtube	Reddit
BartScore	-	-	-	-	0.122	0.075	0.246	0.176	0.213	0.197	0.205	0.059	0.132	0.036
UniEval	-	-	-	-	0.264	**0.222**	0.227	0.293	0.366	0.256	0.380	0.211	0.184	0.112
G-Eval	-	-	-	-	0.296	0.216	0.326	0.305	**0.440**	**0.305**	**0.448**	**0.236**	0.183	0.129
Vicuna (FT)	-	-	-	-	**0.308**	0.219	**0.328**	**0.331**	0.310	0.279	0.335	0.207	**0.260**	**0.437**

Table 2: Meta-evaluation on SummQual's test set. The table on the top is the Pearson correlation coefficient between meta-evaluation models and human ratings on the aspects of coherence (Coh), consistency (Con), coverage (Cov), fluency (Flu), and overall. We also list the correlations on different domains regarding the overall rating. The table at the bottom is Kendall's Tal coefficient on whether the summary has any hallucination / inconsistent span or not.

lower absolute scores show that these summarization systems still tend to overlook some important facts in the document.

For the fluency metric, while GPT-4 still leads the performance, the gap between different models is much smaller than other metrics. For the News and News TV domains, all models exhibit similar performances. Notably, for domains with noisier input text, e.g., Youtube and clinical dialogues, the fluency scores are much lower than other domains.

To summarize, the annotations show that these summarization systems can generate summaries with good coherence and consistency, but relatively poor coverage. Overall, GPT-4 performs the best on all metrics. Vicuna, Instruct-GPT, and BART are close to each other, while Pegasus and StableLM trail far behind across domains. All models work better in the news domain, probably due to the more prevalent news corpus.

4.2 Meta-Evaluation

One of the most important usages of a human-annotated summary quality dataset is for meta-evaluation to assess the performance of automatic evaluators by determining how well their outputs align with human judgments. In this section, we introduce the meta-evaluation process using the SummQual dataset. Each evaluator takes a document and a summary as input and outputs a score for each aspect: coherence, consistency, coverage, fluency, hallucination, and overall. To meta-evaluate an evaluator, we compute the Pearson correlations or Kendall's Tau between the scores generated by the evaluator and those provided by human annotators. A higher correlation indicates that the evaluator is more effective at judging the summary quality.

A traditional evaluator in summarization, such as ROUGE (Lin, 2004), requires a reference summary to compare against the system summary being evaluated. In SummQual , we did not request annotators to write a summary because we believe this would be a much harder task than rating the existing summaries on a fine-grained aspect. Therefore, we only choose reference-free evaluators which only require the document and the system summary as input. More specifically, we choose 1) BartScore (Yuan et al., 2021), which computes the perplexity of generating a summary given a document by BART (Lewis et al., 2019), 2) UniEval (Zhong et al., 2022), which evaluates the summary quality by asking boolean questions, 3) G-Eval (Liu et al., 2023), which prompts GPT-4 to score summaries given the source document, 4) Vicuna (FT), i.e., finetuned Vicuna-13b, which adds a multi-layer perceptron (MLP) layer on top of the hidden representation of the last token, and is trained with mean squared error loss on SummQual 's training set. We used the same prompts as in G-Eval (Liu et al., 2023) and trained a unified model for all aspects of evaluation.

All of our experiments on meta-evaluation are shown in Table 2. BartScore is widely used to compare the similarity between the summary and the reference. However, we don't have references in SummQual, which leads to the poor performance of BartScore. UniEval is a reference-free evaluation method that can achieve significantly better performance than BartScore. However, it is still not good at recognizing the fluency of the summaries. G-Eval aligns with human overall rating better than UniEval. As reported by Liu et al. (2023), GPT-4 based G-Eval prefers the summaries generated by GPT-4, and we also find that G-Eval gets worse performance on the domains of Wikipedia and

Document and Summaries	Challenges
YouTube Transcript: you will not believe what just happened; to me a guy on the subway just called me; a biggledy boo what's a bigly boo is it; 18th century word for dark skinned more; I've learned the word black in every; language just so I know when to be; offended Russian; Korean dolphin; well I'm sure it was just an isolated; incident nah I'm telling you.com old; school racism is bad how can racism be; back when we elected a black president; ... **GPT-4 summary**: [A black writer quits his job after accusing his boss of affirmative action and his co-worker of racism.]; His boss discovers that she also benefited from programs that favored women and minorities in her education and career.; [She feels guilty and tries to dance away her troubles.] Annotation: Coherence: 3, Consistency: 4, Coverage: 4, Fluency: 4, Overall: 3	1.Oral language 2.Speech recognition error 3.Punctuation error 4.Long document
Microsoft Annual Report: FINANCIAL REVIEW ... On September 20, 2016, our Board of Directors approved a share repurchase program authorizing up to $40.0 billion in share repurchases... We repurchased the following shares of common stock under the share repurchase programs: (In millions) Shares Amount Shares Amount Shares Amount Year Ended June 30, 2021 2020 2019 First Quarter 25 $ 5,270 29 $ 4,000 24 $ 2,600 Second Quarter 27 5,750 32 4,600 57 6,100 Third Quarter 25 5,750 37 6,000 36 3,899 Fourth Quarter 24 6,200 28 5,088 33 4,200 Total 101 $ 22,970 126 $ 19,688 150 $ 16,799 **Vicuna summary**: 1. [Apple's Board of Directors] approved a share repurchase program in September 2016, which was completed in February 2020. 2. In September 2019, another share repurchase program was approved, which commenced in February 2020 and has no expiration date. As of June 30, 2021, $8.7 billion remained of this program... Annotation: Coherence: 5, Consistency: 4, Coverage: 4, Fluency: 5, Overall: 4	1.Structured data in table 2.Hallucination due to background missing

Table 3: A case study on the challenges of summarization on multi-domain documents. Text in brackets is the human-annotated inconsistent spans to the source document.

Pubmed where Vicuna is comparable to GPT-4. For the finetuned Vicuna model, it is the only supervised evaluation model and achieves the best overall performance. Moreover, it doesn't have a preference for some specific models and achieves much better performance than G-Eval in all aspects and most domains. When we look at the correlation to whether the summary has an inconsistent span, the trend is the same, but with lower absolute performance. This shows that hallucination detection is still a hard task to be solved.

4.3 Challenges

Different domains pose different challenges for natural language processing, especially for summarization. Domain bias may affect the model training and performance, even with large-scale pretraining. In this paper, we present a case study of two under-explored domains and their difficulties for summarization, as shown in Table 3.

The first domain is YouTube transcripts, which are often long, informal, and poorly punctuated. They are generated from speech recognition, which may introduce errors or misunderstandings, such as the sentence "I'm telling you.com old". These noisy inputs make it hard for summarization systems to extract the main points and avoid hallucinations, as shown by the red text in the output.

The second domain is annual reports, which contain many numbers and structured data, such as tables. These require summarization systems to handle numerical and logical reasoning, as well as background knowledge. For example, the system needs to know that the document is about Microsoft, not Apple. However, the system may confuse the context with other similar documents and generate inaccurate or misleading summaries, as shown by the red text in the output.

We believe that these domains, and others, deserve more attention and evaluation to test and improve the capabilities of large language models for summarization.

5 Conclusion

In this paper, we present a comprehensive, human-annotated summarization evaluation dataset, *SummQual* , consisting of approximately 6,000 documents. Each document is accompanied by summaries generated from 6 state-of-the-art summarization models, along with multi-aspect human judgement on the summary quality. This is the first dataset that evaluates LLMs for summarization across a wide range of domains and the largest dataset for multi-domain summarization meta-evaluation. Based on this dataset, we provide a detailed comparison of summarization systems and a meta-evaluation of existing evaluators. We believe that our work can offer valuable insights and benchmarks for future research and development of LLMs in the field of summarization.

Limitations

Our dataset doesn't provide a golden reference for each document. Budget constraint is one reason. Training a summarization expert is time-consuming and costly. Asking labelers to write summary is much more expensive than ranking summaries.

Although our labelers come from professional data annotation companies and we have been training the labelers on the annotation, we still miss some expert reviews on ranking the summaries, especially domain experts or users who care about the summarization on the corresponding domains.

References

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. 2023. Do language models know when they're hallucinating references? *arXiv preprint arXiv:2305.18248*.

Manik Bhandari, Pranav Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100 (EMNLP)*.

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 (NeurIPS)*.

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*.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics (TACL)*.

Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *the Association for Computational Linguistics (ACL)*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

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 (NeurIPS)*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems (NeurIPS)*.

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 (ICML)*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv: 2306.05685*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.