Benchmarking Generation and Evaluation Capabilities of Large Language Models for Instruction Controllable Summarization

Yixin Liu[*1] **Alexander R. Fabbri**[*2] **Jiawen Chen**[1] **Yilun Zhao**[1] **Simeng Han**[1]
Shafiq Joty[2] **Pengfei Liu**[3] **Dragomir Radev**[1] **Chien-Sheng Wu**[2] **Arman Cohan**[1,4]

[1]Yale University [2]Salesforce AI [3]Shanghai Jiao Tong University [4]Allen Institute for AI

yixin.liu@yale.edu, afabbri@salesforce.com, arman.cohan@yale.edu

Abstract

While large language models (LLMs) can already achieve strong performance on standard generic summarization benchmarks, their performance on more complex summarization task settings is less studied. Therefore, we benchmark LLMs on ***instruction controllable*** text summarization, where the model input consists of both a source article and a natural language requirement for desired summary characteristics. To this end, we curate an evaluation-only dataset for this task setting and conduct human evaluations of five LLM-based systems to assess their instruction-following capabilities in controllable summarization. We then benchmark LLM-based automatic evaluation for this task with 4 different evaluation protocols and 11 LLMs, resulting in 40 evaluation methods. Our study reveals that instruction controllable text summarization remains a challenging task for LLMs, since (1) all LLMs evaluated still make factual and other types of errors in their summaries; (2) no LLM-based evaluation methods can achieve a strong alignment with human annotators when judging the quality of candidate summaries; (3) different LLMs show large performance gaps in summary generation and evaluation capabilities. We make our collected benchmark InstruSum publicly available to facilitate future research in this direction.

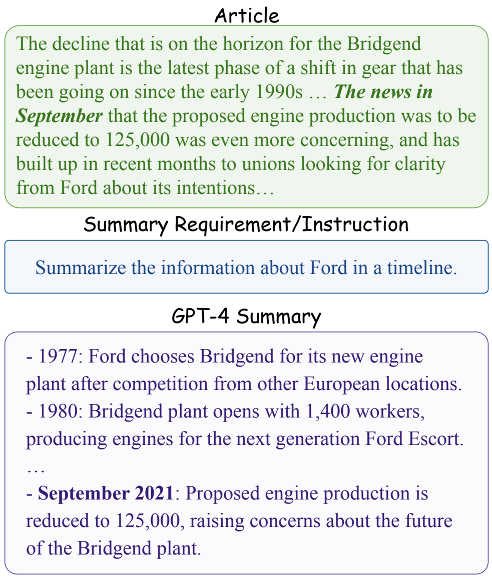

Figure 1: Task example of instruction controllable text summarization. GPT-4 made a factual error in the summarized timeline by including a fabricated date (2021) not mentioned in the source article.

1 Introduction

Recent studies (Goyal et al., 2022; Liu et al., 2023b; Zhang et al., 2024; Pu et al., 2023) have found that large language models (LLMs), e.g., GPT-3.5 (Ouyang et al., 2022), can achieve state-of-the-art, even human-level performance on widely used summarization benchmarks such as the CNN/DailyMail dataset (Nallapati et al., 2016). Moreover, there are signs that LLM performance is saturated on the task of generic summarization, since on these benchmarks (1) LLMs with varying capacity levels, e.g., GPT-3.5 and GPT-4 (OpenAI, 2023), are rated similar under human evaluation (Zhang et al., 2024; Pu et al., 2023); (2) the inter-annotator agreement for comparing strong-performing systems is usually low and significantly influenced by subjective preferences (Goyal et al., 2022; Liu et al., 2023b; Zhang et al., 2024).

We argue that a root cause of this saturation is that traditional summarization settings, such as "*summarize this article in a few sentences,*" can be too simplistic and underconstrained (Kryscinski et al., 2019); without specifying the information need of an intended user, there exist many "good" summaries, but no clear criteria to compare them. Consequently, the summaries generated under these settings may not fully satisfy practical usability criteria, and it remains an open question whether LLMs can perform well in more controlled settings aligned with users' real needs.

[*] Equal contribution

Findings of the Association for Computational Linguistics: NAACL 2024, pages 4481–4501
June 16-21, 2024

Therefore, we aim to study LLMs' capacities in ***instruction controllable***[1] text summarization. To this end, we define a summarization task that takes both a source article and a summary requirement/instruction as input. This task setting can be viewed as an extension of both query-focused summarization (Zhong et al., 2021; Vig et al., 2022) and aspect/attributed-based controllable summarization (He et al., 2022; Zhang et al., 2023). However, the natural language instructions offer greater controllability and flexibility for more complex situations, such as a combination of an information query and a formatting requirement, leveraging the LLMs' instruction-following abilities (Ouyang et al., 2022). We show a task example in Figure 1.

To study our proposed task setting, **we curate a human annotation benchmark that evaluates the performance of several representative LLMs on ins-controllable summarization** (§2). Specifically, we construct task samples by manually selecting articles from the XL-Sum dataset (Hasan et al., 2021) and writing the summary requirements ourselves, aiming to reflect the actual information needs of the users during reading. Then, we collect human annotations of representative LLMs on this task along 4 quality dimensions: (1) overall quality, (2) missing information, (3) irrelevant information, and (4) factual consistency. The evaluation results present a comprehensive view of the current LLMs performance on the ins-controllable summarization task, demonstrating large performance gaps among LLMs with different capacities. Furthermore, we found that even the strongest LLM that we evaluated, i.e., GPT-4, still makes factual and other types of errors, indicating room for future improvement.

During the human annotation collection, we found that as the complexity of the summarization task rises, evaluating the summaries becomes increasingly difficult. Therefore, **we investigate the performance of a variety of LLM-based automatic evaluation methods on our proposed task** (§3). To this end, we compare 40 evaluation methods, each a combination of an evaluation protocol, such as point-wise scoring (e.g., G-Eval (Liu et al., 2023a)), and an LLM as the backbone model. Using the collected human annotations to evaluate these methods, we observe significant performance gaps among different evaluation protocols and different LLMs. Moreover, we found that while several methods we investigate, such as the GPT-4 powered ones, already achieve a strong performance at comparing summarization systems, none of them are well-aligned with the human evaluation when comparing individual summaries.

Having identified the most reliable automatic evaluation methods, e.g., pairwise comparison powered by GPT-4, **we investigate whether these evaluation methods can reliably automate the benchmarking of ins-controllable summarization** (§4). Specifically, we evaluated 11 different LLMs along the quality dimensions we defined. We found that the current LLM-based evaluation methods fail to provide convincing results since they can be biased by confounding factors such as summary lengths.

Our contributions are as follows:

(1) We curated a manually annotated evaluation dataset for ins-controllable text summarization to facilitate the evaluation of LLM-based summarization and summarization evaluation.

(2) We collected a human evaluation benchmark, InstruSum, consisting of multi-dimensional quality annotations of summaries generated by different LLMs on the ins-controllable summarization task, and made InstruSum publicly available.[2]

(3) We benchmarked a series of LLM-based automatic evaluation methods that couple different evaluation protocols with various LLMs using our collected human annotations, and highlighted their limitations in automatic benchmarking for ins-controllable summarization.

2 Human Annotation Collection

We curate a human evaluation benchmark for ins-controllable summarization with two steps: (1) dataset creation and (2) system output evaluation.

2.1 Dataset Creation

Ins-controllable summarization can be defined as

$$S \leftarrow f(D, I), \tag{1}$$

where D is the input document, I is a specific summary requirement, S is the desired summary, and f is a summarization system. To ensure the quality of our following evaluation, we (the authors) manually construct an evaluation-only dataset with the proposed task format. The articles chosen are from the English split of XL-Sum (Hasan et al., 2021) dataset, containing news articles from

[1] For brevity, we will use the term "ins-controllable" to refer to "instruction controllable" throughout this paper.

[2] InstruSum is available at https://github.com/yale-nlp/InstruSum.

Summarize the main factors that led to the conflict between the Ethiopian government and forces in the Tigray region.
Summarize the notable figures from the Prohibition era mentioned in this article.
Summarize the history of the Cononish gold mine in bullet points.
Summarize the views of Democrats and Republicans on trusting information coming from the WHO.
Summarize the experiences of Chum Mey in the 1970s with a timeline.
Summarize the similarities of the definitions of collusion provided by different people in this article.
Summarize the concerns and opposition from the public about the new PNR directive into bullet points for the views of each group.
Summarize the possible explanations for why there hasn't been any firm evidence of aliens' existence, under the assumption that they do exist.
Summarize the people quoted in the article and their identity.
Summarize the efforts of the Brazilian Tourist Board to attract more tourists in three sentences.

Table 1: INSTRUSUM summary requirement examples.

the BBC website. We chose XL-Sum because (1) XL-Sum was already made public by Hasan et al. (2021), which makes it easier for us to release our benchmark; (2) XL-Sum is newer and less commonly studied than other datasets such as CNN/DailyMail (Nallapati et al., 2016), which reduces the concern of data contamination. We collected 100 article-requirement pairs in total with the following steps:

(1) **Searching for challenging articles**. Since not all XL-Sum articles are complicated enough to require specialized summaries, we first select articles with abundant and complex information, of which the requirement-specific summaries can be very beneficial to the readers. Besides, only articles with around 1000-1200 words are selected to ensure they are sufficiently long but not too difficult for human evaluation.

(2) **Writing summary requirements**. After selecting an article, we write one or more summary requirements with different focuses, simulating the real reading experience, where readers may have different informational needs throughout the reading process, as well as structural or formatting preferences. We also used GPT-4 to generate candidate requirements in order to increase the requirement diversity.[3] However, they were not frequently used and were edited by us to ensure their naturalness and correctness. In Table 1, we show a list of summary requirements.

Article	Requirement	Initial Summ.	Hybrid Summ.
1193.4	15.4	115.1	107.7

Table 2: Dataset statistics of INSTRUSUM. The average length (tokens) of the article, the summary requirement, the initial LLM summary, and the hybrid LLM-human summary are reported.

(3) **Obtaining hybrid LLM-human summary**. With the article-requirement pair, we prompt the LLMs to generate a summary using a zero-shot prompt.[4] We then make minimal necessary edits to the LLM summary to obtain a hybrid LLM-human summary. To analyze the effect of the choice of LLM on the human-edited summary, we interchangeably used three OpenAI LLMs to generate the initial summary: `text-davinci-003`, `gpt-3.5-turbo-0301`, and `gpt-4-0314`.[5]

The basic dataset statistics are in Table 2.

2.2 System Output Evaluation

We benchmark the LLMs' performance on the ins-controllable summarization task by collecting human judgments over 4 quality dimensions on the 100 samples from above, resulting in a new benchmark, INSTRUSUM, consisting of 500 summary-level human annotations. The dimensions are:

(1) **Overall Quality**: This rating assesses the overall quality of the summary in relation to the summary requirement.

(2) **Missing Information**: Does the summary omit any crucial information from the article concerning the summary requirement?

(3) **Irrelevant Information**: Does the summary include any information that is not relevant to the summary requirement?

(4) **Factual Consistency**: Is the summary consistent with the facts presented in the article, without contradicting or misrepresenting any information?

We annotate each quality dimension using a ranking protocol, ranking summaries from 1 (best) to 5 (worst). For factual consistency, we ask the annotators to select the span(s) containing a factual inconsistency, and for overall preference, we ask the annotators to explain the reasoning behind their overall rankings. Screenshots of our annotation

[3]The prompt template is in Appendix A.1.

[4]The prompt template is in Appendix A.1.

[5]https://platform.openai.com/docs/models

Mode	Overall	Missing	Irrelevant	Factual
listwise	0.2571	0.2247	0.1925	0.0196
pairwise	0.4428	0.3657	0.2588	0.0721

Table 3: Inter-annotator agreements (Krippendorff's alpha) for INSTRUSUM across various quality dimensions at both *listwise* ranking and *pairwise* comparison levels.

protocol can be found in Appendix B.

We select the following four LLMs for annotation: `text-davinci-002`, `text-davinci-003`, `gpt-3.5-turbo-0301`, `gpt-4-0314`, in addition to the hybrid LLM-human summary. These models are chosen to help study recent LLM development over multiple sizes and training paradigms.[6] For each summary, we collect three annotations.

For this annotation, we recruit annotators on Amazon Mechanical Turk[7] (MTurk). The annotators must pass two rounds of qualification tests. Moreover, to ensure the annotation quality, we maintained ongoing conversations with the annotators to exchange feedback and address their questions. Additionally, we conducted spot checks on each batch of annotations to maintain quality and provide feedback to the crowd annotators.

The inter-annotator agreements are presented in Table 3 for both the original ranking annotation task and the converted pairwise comparison results with the MASI distance (Passonneau, 2006) following Goyal et al. (2022) to enhance comparability. We are able to achieve moderately high agreements on most dimensions, including the overall quality evaluation, which has been shown to be difficult to annotate with high agreements in prior work (Zhang et al., 2024). Regarding factual consistency, we note that low agreement may stem from the sparsity of errors in the dataset. We (the authors) manually verified whether the annotated spans contained factual errors. Our annotation revealed that the errors made often proved to be subtle errors or nuanced different understandings of the article, and we found the accuracy of the crowd annotations to be 88.4%. Factual error examples are provided in Appendix E.

We note the difficulty of our annotation task. Specifically, (1) Earlier iterations of our annotation interface used a Likert scale, but we found that this resulted in a low inter-annotator agreement; (2) Only around 5% of the crowd-workers that participated in the qualification tests achieved acceptable performance to be recruited for our task, although all of them have at least a 99% acceptance rate on MTurk; (3) The average time to complete one annotation task is around 30 minutes. Furthermore, we increased the payment level to enhance annotator retention, as the high cognitive demands of our task tend to discourage annotators from completing more tasks. As a result, we found it challenging to expand the annotation sample size because of both budget constraints and the intensive labor required.

[6]Model details are in Appendix C.

[7]`https://www.mturk.com/`

System	Overall	Missing	Irrelevant	Factual
text-davinci-002	2.344	2.595	3.443	0.640
text-davinci-003	3.239	3.703	3.708	0.710
gpt-3.5-turbo-0301	2.897	3.473	2.958	0.800
gpt-4-0314	**3.970**	**4.067**	4.205	**0.860**
hybrid	3.873	3.947	**4.359**	**0.860**

Table 4: INSTRUSUM: human evaluation results of LLM-generated ins-controllable summaries on 4 quality dimensions. The scores for *overall* quality, *missing* information, and *irrelevant* information dimensions range from 1 to 5. The *factual* score is the ratio of factually consistent summaries. *Hybrid* is the hybrid LLM-human summary.

2.3 Are LLMs Good at Ins-Controllable Summarization?

Results from our human evaluation are found in Table 4. For the *overall quality*, *irrelevant information*, *missing information* dimensions, we convert the human-annotated rankings we obtained in §2.2 into system scores as follows on each data example:

$$s_i = N - \sum_{j=1}^{N} \mathbb{1}_{\{r<r_i\}}(r_j), \tag{2}$$

where s_i is the converted score of the summary of the i-th system, r_i, r_j are the summary rankings and a smaller ranking represents higher quality, $\mathbb{1}$ is the indicator function, and N is the number of ranked summaries. Using this scoring schema, a perfect system would achieve a full score of N (i.e., 5 in our case). For factual consistency, since we found a low agreement in the crowd annotations, we report the ratio of factual summaries according to the human annotations verified by the authors. We note the following findings from Table 4:

(1) **There is a large performance gap among the different LLMs**. Specifically, GPT-4 is significantly better than the GPT-3.5 models,[8] and the supervisedly fine-tuned `text-davinci-002` archives

[8]The p-value is less than 0.01 for all the comparisons, except for `gpt-4-0314` v.s. `gpt-3.5-turbo-0301` on the factual consistency dimension, of which the p-value is 0.058.

Dimension1	Dimension2	Agreement
Overall	Irrelevant	0.412
Overall	Missing	**0.611**
Irrelevant	Missing	0.209

Table 5: The inter-dimension agreements of human annotations on INSTRUSUM. *Overall quality* and *missing information* dimensions have the highest agreements.

worse performance than the LLMs fine-tuned with reinforcement learning from human feedback.[9] In contrast, recent work (Pu et al., 2023) found that the performance of GPT-3.5 and GPT-4 are very similar on the generic news summarization task. This suggests that the ins-controllable summarization task we proposed can be a more suitable benchmarking task for the LLMs.

(2) **All LLMs we evaluated still make a considerable amount of factual errors in their summaries**. For example, the error rates of `text-davinci-002` and `gpt-4-0314` are 36% and 14%, respectively. This is also different from the patterns on the generic news summarization task, on which the factual error rate is only 1-2% for `text-davinci-002` (Zhang et al., 2024).

(3) **The hybrid LLM-human summary can only outperform GPT-4 on the irrelevant information dimension**, suggesting that GPT-4 is close to the human-level performance, especially when the human annotator is asked to only edit the LLM summary. It also indicates that generating summaries without irrelevant information is the most challenging dimension for current LLMs. Interestingly, the following section (§3.3) will show that irrelevant information is also the most challenging dimension for the LLM-based automatic evaluation methods. We provide a fine-grained comparison of the initial LLM and hybrid summaries in Appendix D.[10]

Inter-Dimension Analysis To explore the relationship between human evaluation results across different quality dimensions, we examine the (listwise ranking) agreements between the summary scores for these dimensions in Table 5. The results show that the *missing information* dimension has a higher influence on the *overall quality* dimension than the *irrelevant information* dimension, suggesting that human annotators favor comprehensive over concise summaries, similar to recent work's findings on length bias (Liu et al., 2023b; Singhal et al., 2023; Zheng et al., 2023; Huang et al., 2023; Saito et al., 2023) in human evaluation.

3 Are LLMs Good at Ins-Controllable Summary Evaluation?

Human evaluation of our proposed ins-controllable summarization is complex and time-intensive, requiring scalable, reliable automatic evaluation methods. Consequently, we benchmark recent LLM-based evaluation methods, exploring various evaluation protocols and LLM backbones.

3.1 LLM-based Evaluation Protocols

LLM-based automatic evaluation methods can be categorized along two orthogonal dimensions – the ***evaluation protocols*** and the backbone LLMs. We investigate the following evaluation protocols:

(1) **LLMScore**: direct scoring using predicted probability. GPTScore (Fu et al., 2023) proposes a protocol that interprets the LLM-predicted probability of certain token(s) as a quality score.

(2) **LLMEval**: direct scoring by text completion. In Chiang and Lee (2023) and G-Eval (Liu et al., 2023a), the LLM is asked to assign a quality score.

(3) **LLMCompare**: pairwise comparison between two candidate outputs by text completion (Zheng et al., 2023; Wang et al., 2024).

(4) **LLMRank**: listwise ranking by text completion, simultaneously evaluating a list of candidate outputs (Sun et al., 2023; Liu et al., 2024).

Prompt Design. For each of the evaluation protocols, we design *dimension-specific* prompts for the evaluation quality dimensions we defined in §2.2. The prompt templates are in Appendix A.2.

3.2 Evaluation Settings

We benchmark 11 LLMs in total on ins-controllable summarization evaluation over three quality dimensions, *overall quality*, *missing information*, and *irrelevant information*. We did not benchmark factual consistency evaluation since it is a more unique dimension, which we leave for more dedicated future work. For proprietary LLMs, we use different versions of GPT-3.5 and GPT-4 models provided by OpanAI.[11] For open-source LLMs, we benchmark LLama-2-chat (Touvron et al., 2023) 7B, 13B, and 70B models, and the Mistral-Instruct (Jiang et al., 2023) 7B model.[12] The full model list is in Table 6.

[9] Model details are in Appendix C.

[10] We interchangeably used GPT-4 and GPT-3.5 models to generate the initial LLM summary so Table 4 does not provide a direct comparison between the initial and hybrid summaries.

[11] `https://platform.openai.com/docs/models`

[12] Llama-2 models were released in July 2023, and Mistral 7B was released in September 2023.

	System-level Correlations				Summary-level Correlations			
	LLMRank	**LLMCompare**	**LLMEval**	**LLMScore**	**LLMRank**	**LLMCompare**	**LLMEval**	**LLMScore**
	Overall Quality							
gpt-3.5-turbo-0301	0.738	0.400	0.600	-	0.005	0.185	0.223	-
gpt-3.5-turbo-0613	0.600	0.527	0.527	-	-0.012	0.160	0.048	-
gpt-4-0314	0.800	**1.000**	**1.000**	-	0.095	0.361	0.271	-
gpt-4-1106-preview	0.400	0.800	0.800	-	0.047	**0.483**	0.257	-
text-davinci-002	-0.200	0.400	0.738	0.600	-0.044	0.026	0.114	0.062
text-davinci-003	0.400	0.400	0.949	-0.400	-0.034	0.029	0.052	-0.133
gpt-3.5-turbo-instruct	0.400	0.600	0.738	-0.200	0.006	0.212	0.078	-0.058
llama-2-7b-chat	0.200	0.527	0.527	0.000	-0.062	-0.019	0.028	0.063
llama-2-13b-chat	0.105	0.400	**1.000**	-0.400	-0.058	0.096	0.037	-0.032
llama-2-70b-chat	-0.316	0.400	0.949	0.800	-0.006	0.072	0.016	0.116
mistral-instruct	-0.400	0.200	0.447	-0.200	-0.074	0.139	0.021	0.137
	Missing Information							
gpt-3.5-turbo-0301	0.400	0.400	0.600	-	-0.051	0.283	0.175	-
gpt-3.5-turbo-0613	0.316	0.200	0.400	-	-0.083	0.244	0.200	-
gpt-4-0314	0.949	**1.000**	0.949	-	-0.001	0.440	0.233	-
gpt-4-1106-preview	0.738	0.400	**1.000**	-	0.063	**0.443**	0.085	-
text-davinci-002	0.200	0.200	0.200	0.800	-0.034	0.037	-0.001	0.259
text-davinci-003	0.400	0.400	**1.000**	0.400	-0.077	0.141	0.106	0.190
gpt-3.5-turbo-instruct	0.200	0.600	0.738	0.800	-0.038	0.226	0.129	0.140
llama-2-7b-chat	-0.400	0.738	0.105	-0.200	-0.108	0.012	0.016	-0.103
llama-2-13b-chat	0.527	0.400	0.600	0.000	-0.051	0.246	0.085	-0.046
llama-2-70b-chat	0.527	0.400	0.600	-0.600	-0.023	0.119	0.044	-0.173
mistral-instruct	-0.600	0.600	0.400	0.000	-0.120	0.205	0.061	0.036
	Irrelevant Information							
gpt-3.5-turbo-0301	-0.200	-0.200	0.200	-	-0.008	-0.081	0.013	-
gpt-3.5-turbo-0613	0.000	0.000	-0.200	-	-0.007	-0.024	-0.026	-
gpt-4-0314	0.400	0.600	**0.738**	-	0.057	0.208	0.057	-
gpt-4-1106-preview	0.200	0.600	0.600	-	0.180	**0.332**	0.242	-
text-davinci-002	-0.400	-0.400	0.105	0.200	-0.043	-0.053	0.067	-0.062
text-davinci-003	0.000	0.105	0.600	-0.400	-0.019	-0.009	0.139	0.058
gpt-3.5-turbo-instruct	0.200	0.200	0.120	-0.200	0.023	0.006	0.118	0.013
llama-2-7b-chat	0.000	0.200	0.000	-0.600	-0.010	0.037	-0.029	-0.064
llama-2-13b-chat	0.600	0.000	0.400	0.200	-0.012	-0.102	-0.004	-0.011
llama-2-70b-chat	-0.105	-0.200	0.400	-0.800	-0.042	-0.035	0.062	0.130
mistral-instruct	-0.527	0.000	0.200	-0.200	-0.052	-0.095	0.046	-0.095

Table 6: Kendall rank correlations at both the system and summary levels between human evaluation and LLM-based evaluation over three quality dimensions on INSTRUSUM. The LLM-based evaluation performance of different combinations of backbone LLMs (e.g., `gpt-4-0314`) and evaluation protocols (e.g., LLMRank) is reported. The best performance on each quality dimension at the system level or the summary level is highlighted.

To compare LLM evaluation against human evaluation, we calculate the correlations between their evaluation scores at both system and summary levels. The *system-level* correlations measure how good the LLMs are at comparing summarization system performance, while the *summary-level* correlations measure how good the LLMs are at comparing summary quality on individual data samples. Since we adopted a ranking-based evaluation protocol for our human evaluation collection (§2.2), we use the Kendall rank correlation coefficient as the correlation measurement, and we transform the evaluation results of different evaluation protocols into rankings. More evaluation setting details including the formal definitions of the correlation measurement are in Appendix F.

3.3 Result Analysis

The evaluation results are reported in Table 6. For each of the backbone LLMs we chose to benchmark (§3.2), we evaluate its performance with different evaluation protocols when applicable[13] so in total we evaluate 40 LLM-based evaluation methods. We make the following observations:

(1) **Different LLMs have significantly different performance at evaluating ins-controllable sum-**

[13]We could not use a few OpenAI models with LLMScore since the log-likelihood is not provided by the APIs.

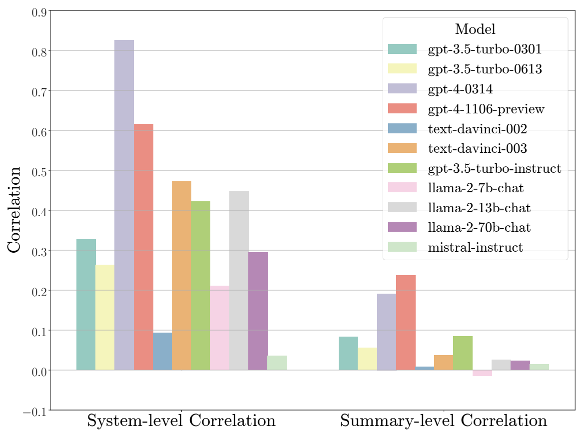

Figure 2: Average LLM performance of ins-controllable summary evaluation across 3 quality dimensions with 3 evaluation protocols on INSTRUSUM.

maries. In Figure 2, we report the average LLM performance across 3 quality dimensions on 3 protocols except LLMScore. In particular, GPT-4 shows a consistent advantage over other LLMs.

(2) **The choice of evaluation protocols has a large impact on the evaluation method performance**. For example, the pairwise comparison protocol (LLMCompare) is (almost) always better than the listwise protocol (LLMRank). Besides, the most suitable protocol for each backbone LLM can be different. For instance, `gpt-4-0314` works better with LLMCompare while `gpt-3.5-turbo-0301` tends to work better with LLMEval.

(3) In general, LLM-based evaluation methods have much higher system-level correlations than summary-level correlations, which means these methods are better at evaluating which system is better on average, but **struggle at ranking different summaries of individual data examples**. Notably, the strongest evaluation method we identified, i.e., pairwise comparison using `gpt-4-0314`, can only achieve an agreement value of 0.277 with human evaluation on the *overall quality* dimension in pairwise comparison, lower than the human inter-annotator agreement (0.4428).

(4) **The performance of the LLM-based evaluation methods differs on different quality dimensions**. In particular, *Irrelevant Information* is a more challenging dimension than *Missing Information*, suggesting that these methods are better at recall-based than precision-based evaluation.

Evaluation Consistency A reliable evaluator must yield consistent results across different evaluation protocols. To check this consistency requirement, we examine the LLMs by calculating the correlations of its evaluation results on the LLMCompare and LLMEval protocols over the three quality dimensions, since these two protocols are most reliable. Table 7 indicates low summary-level consistency among all evaluated LLMs. However, `gpt-4-0314` demonstrates a high system-level consistency, indicating it is the most reliable evaluator.

System	System-Level	Summary-Level
gpt-3.5-turbo-0301	0.600	0.149
gpt-3.5-turbo-0613	0.681	0.135
gpt-4-0314	**0.966**	0.227
gpt-4-1106-preview	0.800	**0.262**
text-davinci-002	0.418	0.049
text-davinci-003	0.485	0.089
gpt-3.5-turbo-instruct	0.461	0.114
llama-2-7b-chat	0.111	-0.006
llama-2-13b-chat	0.200	0.072
llama-2-70b-chat	0.442	0.021
mistral-instruct	0.416	0.051

Table 7: LLM evaluation consistency between the LLMCompare and LLMEval protocols. System-level and summary-level Kendall rank correlations are reported.

System	Generator	Re-ranker (Evaluator)	
		LLMCompare	LLMEval
text-davinci-002	2.344	3.335	3.383
text-davinci-003	3.239	3.189	3.374
gpt-3.5-turbo-0301	2.897	3.357	3.504
gpt-4-0314	3.970	3.533	3.561

Table 8: Performance comparison of LLMs as the summary generator and the summary re-ranker with different evaluation protocols. The human-annotated scores on the overall quality dimension are reported. A random-reranking oracle can achieve a score of 3.260.

Generator-Evaluator Consistency Recent work has found that the behavior and performance of LLMs can differ when they are used as a generator or an evaluator on the same task (Li et al., 2024b; West et al., 2024). Thus, we analyze the performance consistency of LLMs on ins-controllable summary generation and evaluation. To this end, we treat the LLM evaluator as a *re-ranker* to make its performance more comparable to the LLM generator. In Table 8, we report the generation and re-ranking performance of the 4 human-evaluated LLMs in §2.2 on the overall quality dimension, where the re-ranker can select its output from the human-evaluated candidate summaries. We note:

(1) The generation performance does not always align with the evaluation performance. For exam-

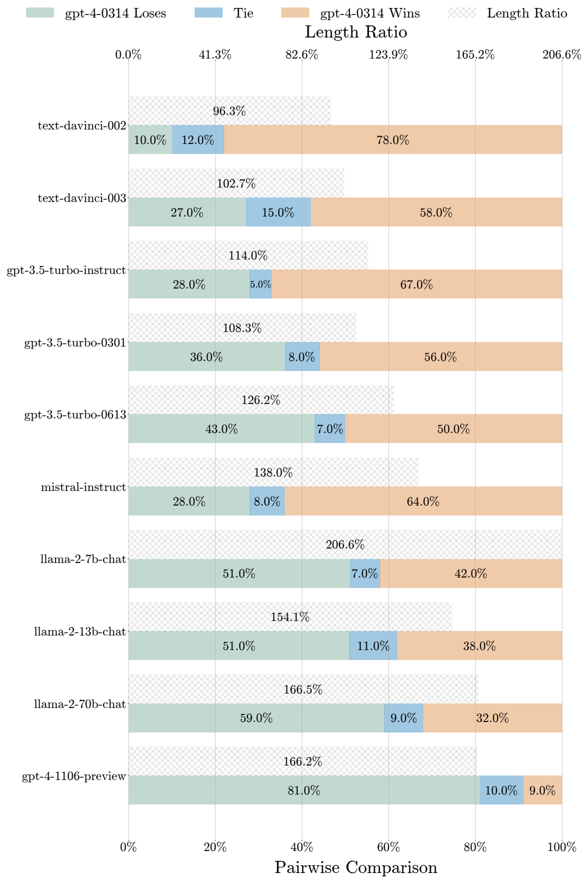

Figure 3: Automatic benchmarking results using `gpt-4-0314` as the evaluator. LLM summaries are compared against `gpt-4-0314`'s summaries. The model evaluation result (bottom) is reported, as well as the summary length ratios (top) relative to `gpt-4-0314`.

ple, `text-davinci-003` has better generation performance than `gpt-3.5-turbo-0301`, but worse evaluation performance.
(2) `gpt-4-0314` fails to outperform its generation performance under the re-ranking task setting. It suggests that despite its promising ability to generate acceptable summaries, GPT-4 might still lack a more in-depth understanding of the task.

4 Can We Automate Ins-Controllable Summarization Benchmarking?

After evaluating LLMs as summary evaluators, we explore their potential for automating ins-controllable summarization benchmarking.

4.1 Evaluation Settings

Since GPT-4 coupled with LLMCompare is the best evaluation method we identified in §3.3, we use it for the automatic benchmarking. To avoid the prohibitive cost, we treat GPT-4 (`gpt-4-0314`) as a baseline and evaluate the other systems by comparing them against GPT-4 only, following recent practices in automatic LLM benchmarking (Dubois et al., 2023; Zheng et al., 2023).[14] We evaluated 11 LLMs over the 100 data examples we used in human evaluation (§2.1). The same prompt template for summary generation, as shown in Appendix A.1, is used for different LLMs.

Pair	Overall	Missing	Irrelevant
Human v.s. Oracle	0.112	-0.268	0.302
GPT-4 v.s. Oracle	0.304	0.098	0.253
GPT-4 v.s. Human	0.277	0.147	0.376

Table 9: Agreements among human evaluation, LLM-based evaluation (`gpt-4-0314`) and the length oracle.

4.2 Result Analysis

The evaluation results are in Figure 3. We found that Llama-2 models show a strong performance under GPT-4's evaluation, even outperforming `gpt-4-0314` in the pairwise comparison setting. However, since we did not observe Llama-2 models achieving performance as strong as GPT-4 in evaluating the ins-controllable summaries (§3.3), we suspect GPT-4 has overestimated Llama-2 models' performance in this summary generation task. The reason is likely that the summaries generated by Llama-2 models are much longer than the `gpt-4-0314` summaries, which tend to be favored by GPT-4 as shown below.

In Table 9, we compare the annotator agreement in pairwise comparison among human evaluation, LLM-based evaluation using `gpt-4-0314` and LLMCompare, and a length oracle that always prefers longer summaries. The results indicate that human evaluation has a positive correlation with the length oracle in the *irrelevant information* dimension and a negative correlation in the *irrelevant information* dimension. Conversely, `gpt-4-0314` has a positive correlation with the length oracle across all quality dimensions. These findings suggest that LLM-based evaluation is more prone to bias from summary length compared to human evaluation. Furthermore, our case study in Appendix G shows that when the length difference is controlled, none of the LLMs we evaluated have a clear advantage over `gpt-4-0314`. Therefore, we find current LLM-based evaluation methods unreliable for automatic ins-controllable summarization benchmarking, and we call for future work in this direction.

[14] We randomly shuffled the summary pairs and did not observe a significant positional bias in the evaluation results.

5 Related Work

Summarization Benchmarks Recent work in summarization benchmarks has focused on aggregating model outputs and annotating them according to specific quality dimensions (Huang et al., 2020; Bhandari et al., 2020; Stiennon et al., 2020; Zhang and Bansal, 2021; Fabbri et al., 2022; Gao and Wan, 2022). In the context of LLMs, Laban et al. (2023) incorporated LLMs into the benchmark-construction process while Maynez et al. (2023) benchmarked LLMs on conditional text generation tasks including summarization. A few recent studies (Goyal et al., 2022; Liu et al., 2023b; Zhang et al., 2024; Pu et al., 2023) point to the strength of LLMs with respect to human-written (reference) summaries on generic news summarization. In this work, we present a benchmark task that poses challenges for current LLMs and allows for further development and model comparison.

Instruction-Following Evaluation Ouyang et al. (2022) introduce InstructGPT, which learns to follow instructions by aligning to human preference feedback and builds on earlier alignment work in summarization (Stiennon et al., 2020). Following Ouyang et al. (2022), a line of work (Wang et al., 2023d; Zhou et al., 2023; Zeng et al., 2024) has investigated methods of improving and benchmarking the instruction-following capabilities of LLMs. Regarding text summarization, instruction-following text summarization expands upon work in query-focused summarization (Zhong et al., 2021; Vig et al., 2022; Yang et al., 2023a; Pagnoni et al., 2023), aspect-based summarization (Zhang et al., 2024; Pu et al., 2023; Yang et al., 2023b), and controllable summarization more broadly (Dou et al., 2021; He et al., 2022; Zhang et al., 2023; Bao et al., 2023; Ribeiro et al., 2023; Ravaut et al., 2023; Adams et al., 2023b,a; Narayan et al., 2023; Pagnoni et al., 2023; Pu and Demberg, 2023). Closely related to work on query-focused summarization is the task of long-form question answering (Fan et al., 2019), and recent work has benchmarked current models and metrics with a focus on completeness and factuality (Xu et al., 2023). In this work, we explore controllability in the context of instructions. Compared with query-focused summarization, our task format allows for more complex use cases where the information queries can be combined with other user request categories such as the output format. Wang et al. (2023a) extends query-focused summarization and curates an instructive dialog summarization dataset. The most relevant work to our study is Skopek et al. (2023), which develops a dataset consisting of human annotations on instruction-summary pairs. However, their evaluation focuses only on the instruction-following capacities, while our human evaluation is multi-dimensional and puts more focus on the models' text summarization capabilities.

LLM-based Automatic Evaluation and Its Meta-Evaluation A series of recent work has investigated leveraging LLMs for automatic evaluation (Fu et al., 2023; Chiang and Lee, 2023; Liu et al., 2023a; Zheng et al., 2023; Wang et al., 2024; Sun et al., 2023; Gao et al., 2023; Wang et al., 2023b; Li et al., 2024a). While these studies have demonstrated LLMs' promising performance on various evaluation tasks such as summarization evaluation, other recent work has highlighted the limitations of LLM-based automatic evaluation methods. Specifically, LLMs can have various biases in their evaluation results (Koo et al., 2023; Wang et al., 2023c), and they fail to align with human evaluation when evaluating close-performing systems (Shen et al., 2023) or adversarial examples (Zeng et al., 2024). Our work provides a thorough meta-evaluation of LLM-based methods, focusing on diverse protocols and backbone LLMs for ins-controllable summarization evaluation.

6 Conclusions

In this work, we benchmarked large language models for instruction controllable summary generation and evaluation, and presented a new benchmark dataset, INSTRUSUM. We found that several LLMs have already shown promising performance in generating ins-controllable summaries. However, they lack robust holistic capabilities for this task since they still make a considerable amount of errors in their summaries and they can not reliability evaluate different candidate summaries. Furthermore, we notice large gaps between the performance of different generations of LLMs on both ins-controllable summary generation and evaluation. As we believe our proposed ins-controllable summarization setting is more realistic and can provide better usability, we call for future work along this direction to make the text summarization systems more beneficial to the actual users.

7 Limitations

Our analysis is limited to 100 examples for which we collected human annotations of ins-controllable summaries generated by different LLMs. While more statistically significant conclusions could be drawn from a larger evaluation set, as noted above a much larger time and budget allocation would be required, and we encourage the community to apply our protocol to expand our evaluation set.

Due to sparsity and subtleties of factuality errors generated by current LLMs on our benchmark, we did not perform a meta-evaluation of LLMs as factuality evaluators, since it would require a larger collection to observe significant error patterns. We leave a larger evaluation of the factual consistency of current models and error types for future work.

Acknowledgements

We express our gratitude to Tanya Goyal for her insightful suggestions, and we thank the anonymous reviewers for their constructive comments. We are grateful for the compute support provided by the Google TRC program.

References

Griffin Adams, Alex Fabbri, Faisal Ladhak, Noémie Elhadad, and Kathleen McKeown. 2023a. Generating EDU extracts for plan-guided summary re-ranking. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2680–2697, Toronto, Canada. Association for Computational Linguistics.

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023b. From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Hybrid. Association for Computational Linguistics.

Guangsheng Bao, Zebin Ou, and Yue Zhang. 2023. GEMINI: Controlling the sentence-level summary style in abstractive text summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 831–842, Singapore. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Alexander Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2022. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9(0):391–409.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of ACL 2019*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRL-sum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics.

Kung-Hsiang Huang, Philippe Laban, Alexander R Fabbri, Prafulla Kumar Choubey, Shafiq Joty, Caiming Xiong, and Chien-Sheng Wu. 2023. Embrace divergence for richer insights: A multi-document summarization benchmark and a case study on summarizing diverse information from news articles. *arXiv preprint arXiv:2309.09369*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander R Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. LLMs as factual reasoners: Insights from existing benchmarks and beyond. *arXiv preprint arXiv:2305.14540*.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. 2024a. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. 2024b. Benchmarking and improving generator-validator consistency of language models. In *The Twelfth International Conference on Learning Representations*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Yixin Liu, Kejian Shi, Katherine S He, Longtian Ye, Alexander R. Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. 2024. On learning to summarize with large language models as references. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Joshua Maynez, Priyanka Agrawal, and Sebastian Gehrmann. 2023. Benchmarking large language model capabilities for conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9194–9213, Toronto, Canada. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, 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, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Artidoro Pagnoni, Alex Fabbri, Wojciech Kryscinski, and Chien-Sheng Wu. 2023. Socratic pretraining:

Question-driven pretraining for controllable summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12737–12755, Toronto, Canada. Association for Computational Linguistics.

Rebecca 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)*, Genoa, Italy. European Language Resources Association (ELRA).

Dongqi Pu and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Mathieu Ravaut, Hailin Chen, Ruochen Zhao, Chengwei Qin, Shafiq Joty, and Nancy Chen. 2023. Promptsum: Parameter-efficient controllable abstractive summarization. *arXiv preprint arXiv:2308.03117*.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in RLHF. *arXiv preprint arXiv:2310.03716*.

Ondrej Skopek, Rahul Aralikatte, Sian Gooding, and Victor Carbune. 2023. Towards better evaluation of instruction-following: A case-study in summarization. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 221–237, Singapore. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jesse Vig, Alexander Fabbri, Wojciech Kryscinski, Chien-Sheng Wu, and Wenhao Liu. 2022. Exploring neural models for query-focused summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1455–1468, Seattle, United States. Association for Computational Linguistics.

Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023a. Instructive dialogue summarization with query aggregations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7630–7653, Singapore. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023b. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023c. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Wenjin Yao, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023d. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. 2024. The generative AI paradox: "what it can create, it may not understand". In *The Twelfth International Conference on Learning Representations*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023a. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Xianjun Yang, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Xiaoman Pan, Linda Petzold, and Dong Yu. 2023b. OASum: Large-scale open domain aspect-based summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4381–4401, Toronto, Canada. Association for Computational Linguistics.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6617–6632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yusen Zhang, Yang Liu, Ziyi Yang, Yuwei Fang, Yulong Chen, Dragomir Radev, Chenguang Zhu, Michael Zeng, and Rui Zhang. 2023. MACSum: Controllable Summarization with Mixed Attributes. *Transactions of the Association for Computational Linguistics*, 11:787–803.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *North American Association for Computational Linguistics (NAACL)*.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42602–42613. PMLR.

A Prompt Templates

Here we provide the prompt templates we used throughout this work.

A.1 Prompts in Human Annotation Collection

Prompt for Summary Requirement Recommendation In §2.1, we used GPT-4 to generate candidate summary requirements to help the human annotation. The prompt template is as follows:

> Please generate a list of specific summary requirements for a given article.
>
> Here are some requirement examples based on different articles: 1. Summarize the possible explanations for why there hasn't been any firm evidence of aliens' existence, under the assumption that they do exist. 2. Summarize the experience of Chum Mey in 1970s with a timeline. 3. Summarize why Shanghai and Hong Kong seem to outperform Beijing in education. 4. Summarize all people and their identities in the article. 5. Summarize the negative outcomes of the lockdown. 6. Summarize the conclusion of the fraud case. 7. Summarize the opinions of Ronan Barry in the article. 8. Summarize the events of Margaret's debit card fraud in a timeline. 9. Summarize the aftermath of sexual harassment on Meena in one sentence. 10. Summarize the difficulties faced by Uber and Lyft now.
>
> Here's an article: {{article}}

Please generate a list of specific summary requirements for this article.

Prompt for Generating Requirement-Specific Summaries We used the following template to prompt the LLMs to generate the requirement-specific summaries.

Summarize the following article based on the specific requirement.

Article: {{article}}

Requirement: {{requirement}}

Summary:

A.2 Prompts for LLM-based Automatic Evaluation

In §3, we analyze the performance of different LLM-based automatic evaluation methods. We designed prompt templates for each evaluation protocol and each evaluation dimension, and slightly fine-tuned templates for several LLMs to ensure that they are able to follow the instructions as much as they can. To enhance the LLM evaluation performance, for LLMCompare and LLMRank, we design chain-of-thought (Wei et al., 2022) style prompts – before the LLM gives the actual answer, it is prompted to first generate an explanation of the answer, mimicking the thinking process of human evaluators. We show the following prompt templates for all the evaluation protocols on the *overall quality* dimension, and all the templates can be found in our code release.

(1) **Prompt template for LLMRank**.

In this task, you will be provided with a news article, a specific summary requirement, and a list of summaries numbered as follows: 1. Summary 1, 2. Summary 2, and so on.

The summaries are crafted to meet a specific summary requirement. Note that there may be identical summaries within the list.

Your task is to evaluate and rank the summaries in ascending order of their overall quality concerning the summary requirement. First, you will explain your ranking, and then you will provide the ranking of each summary. The ranking should be a number between 1 and 5, where 1 is the best and 5 is the worst.

Note: In case of a tie, do not skip a rank. For example, if Summary 1 has ranking 1 and Summary 2 and 3 both have ranking 2, then Summary 4 should be assigned a ranking of 3, not 4.

Please refer to the example below for the format of your response.

Example Response: Explanation: "Your explanation of the ranking." Ranking: "The ranking, e.g., 1, 2, 2, 3, 4."

Here are the actual article, the summary requirement, and the summaries:

Article:

```
{{Article}}
```

Summary Requirement:

```
{{Requirement}}
```

Summaries:

1. Summary 1:

```
{{Summary 1}}
```

2. Summary 2:

```
{{Summary 2}}
```

3. Summary 3:

```
{{Summary 3}}
```

4. Summary 4:

```
{{Summary 4}}
```

5. Summary 5:

```
{{Summary 5}}
```

(2) **Prompt template for LLMCompare**.

In this task, you will be provided with a news article, a specific summary requirement, and two summaries.

The summaries are crafted to meet a specific summary requirement. Note that there may be identical summaries.

Your task is to compare the overall quality of these two summaries concerning the summary requirement and pick the one that is better (there can be a tie). First you will give an explanation of your decision then you will provide your decision in the format of 1 or 2 or tie.

Please refer to the example below for the format of your response.

Example Response:

Explanation: "Your explanation here".

Decision: 1 or 2 or tie.

Here are the actual article, the summary requirement, and two summaries:

Article:

`{{Article}}`

Summary Requirement:

`{{Requirement}}`

Summary 1:

`{{Summary 1}}`

Summary 2:

`{{Summary 2}}`

Please provide your response.

(3) **Prompt template for LLMEval**.

In this task, you will be provided with a news article, a specific summary requirement, and a summary.

Your task is to rate the overall quality of the summary with a score from 1 to 5 concerning the summary requirement, where 1 is the lowest and 5 is the highest.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Example Response:

Evaluation Form (scores ONLY):

- Overall Quality (1-5): 3

Here are the actual article, the summary requirement, and the summary:

Article:

`{{Article}}`

Summary Requirement:

`{{Requirement}}`

Summary:

`{{SUMMARY}}`

Evaluation Form (scores ONLY):

- Overall Quality (1-5):

(4) **Prompt template for LLMScore**.

Answer the question based on the following article, a specific summary requirement, and a summary.

Question: Is the summary of good overall quality in relation to both the article and the summary requirement? (a). Yes. (b). No.

Article:

`{{Article}}`

Summary Requirement:

`{{Requirement}}`

Summary:

`{{SUMMARY}}`

Answer: Yes

B Crowd-Annotation Details

We provide screenshots of the human annotation interface we used for crowd-sourced summary evaluation (§2.2) in Figure 4, 5, and 6. We recruit MTurk annotators who are located in the US or the UK. We set a competitive payment rate for better annotator retention, and the average hourly salary is around 20 US dollars.

Instructions

In this task, you will evaluate the quality of 5 summaries of a news article. To accurately complete this task, follow these steps:

1. **Read the news article carefully.**
2. **Review the summaries.**
3. For each summary, you will need to evaluate its quality on 4 dimensions:
    1. **Factual Consistency**: Is the summary consistent with the facts presented in the article, without contradicting or misrepresenting any information?
    2. **Missing Information**: Does the summary omit any crucial information from the article concerning the *summary requirement*?
    3. **Irrelevant Information**: Does the summary include any information that is not relevant to the *summary requirement*?
    4. **Overall Quality**: This rating assesses the overall quality of the summary in relation to the *summary requirement*.
4. For ***factual consistency***, please **select any span of text that contains an error or inconsistency** and ***right-click*** to add it to the error list below the summary. **You can select and add multiple errors in one summary.** If needed, you can delete an item from the error list by clicking the 'Delete' button next to it. If you find a summary to be factually consistent and without any errors, you do not need to add any items to the error list for that summary.
5. For ***Factual Consistency***, ***Missing Information***, ***Irrelevant Information***, and ***Overall Quality***, after evaluating the ***individual*** summaries, **rank them** from **1** (best) to **5** (worst) based on their quality. Same rankings are allowed if you think two summaries are of the same quality. Please note that there can be identical summaries.
   After assigning the ***Overall Quality*** rankings, please provide a **short explanation** for your ranking.

Warning: Annotations will be checked for quality against control labels. Low-quality work will be rejected.

Factual Error Examples

Source text
US technology firm Apple has offered to refund Australian customers who felt misled about the 4G capabilities of the new iPad. The country's consumer watchdog has taken Apple to court for false advertising because the tablet computer does not work on Australia's 4G network. Apple's lawyers said they were willing to publish a clarification. [...] At a preliminary hearing, Apple lawyer Paul Anastassiou said Apple had never claimed the device would work fully on the current 4G network operated by Telstra. Apple says the new iPad works on what is globally accepted to be a 4G network. The matter will go to a full trial on 2 May.

Summary 1: **Apple lawyer never claimed** that the device would work on the current 4G network operated by Telstra.
This summary contains a **factual error**: in the source text, it is **Apple** not **Apple lawyer** that never made the claim.

Summary 2: Apple has been accused of misleading customers in Australia over its new iPad **3.0 version**.
This summary contains a **factual error**: the version of the new iPad is not mentioned in the source text.

Irrelevant Information Examples

Requirement 1: Summarize what Dr. Mazen Kewara has said.
Summary 1: Dr. Mazen Kewara, Turkey director of the Syrian American Medical Society (SAMS), said that "We cannot use the buildings anymore. Maybe for hours. Maybe until tomorrow. I don't know." He also reported that there were many collapsed buildings in Gaziantep, and that SAMS had four of its hospitals damaged severely by the quake. He added that they had already evacuated two of them. ***Kewara was originally from Damascus, Syria.***
Explanation: The last sentence is irrelevant becuase it does not provide information that is **specifically about the requirement**.

Requirement 2: Summarize the current drawbacks of COP24.
Summary 2: ***COP24 was a success in terms of implementing the 2015 Paris Climate Accord and ensuring that all countries face similar regulations.*** However, many felt that acknowledging the science was only a token effort, and the agreement does not go far enough to tackle climate change. Moreover, the international spirit of co-operation needed to tackle climate change is threatened by the rise of nationalism in many countries and the election of Jair Bolsonaro as Brazil's president.
Explanation: The first sentence is irrelevant becuase **it is not specifically about the requirement**.

Missing Information Examples

Requirement 1: Summarize the companies mentioned in this article.
Summary 1: The article mentions several companies involved in the development of artificial intelligence technology. These include C3.ai, BigBear.ai and SoundHound AI, all of which have seen their share prices surge this year. C3.ai is run by tech veteran Tom Siebel and SoundHound AI is striving to make conversational AI better than humans. BigBear.ai provides AI solutions to US intelligence agencies and has seen its shares surge 700%. All three companies are unprofitable and are not expected to make money this year or in 2024, while BigBear.ai recently raised $25 million in a private placement of stock.
Explanation: The summary failed to mention Google, Microsoft, Baidu, etc.

Requirement 2: Summarize the information in the article in a timeline.
Summary 2: - Early Americans consumed an average of 5.8 gallons of pure alcohol a year in 1790.
- In 1830, alcohol consumption peaked at 7.1 gallons a year, and drinking became a moral issue.
- In 1862, the US Navy abolished the daily rum ration for sailors.
- On January 16, 1919, the 18th Amendment, which set Prohibition into law, became part of the Constitution.
Explanation: One event is missed: "The first arrest for driving under the influence of alcohol was in 1897."

Figure 4: Annotation Interface Part 1: Instructions.

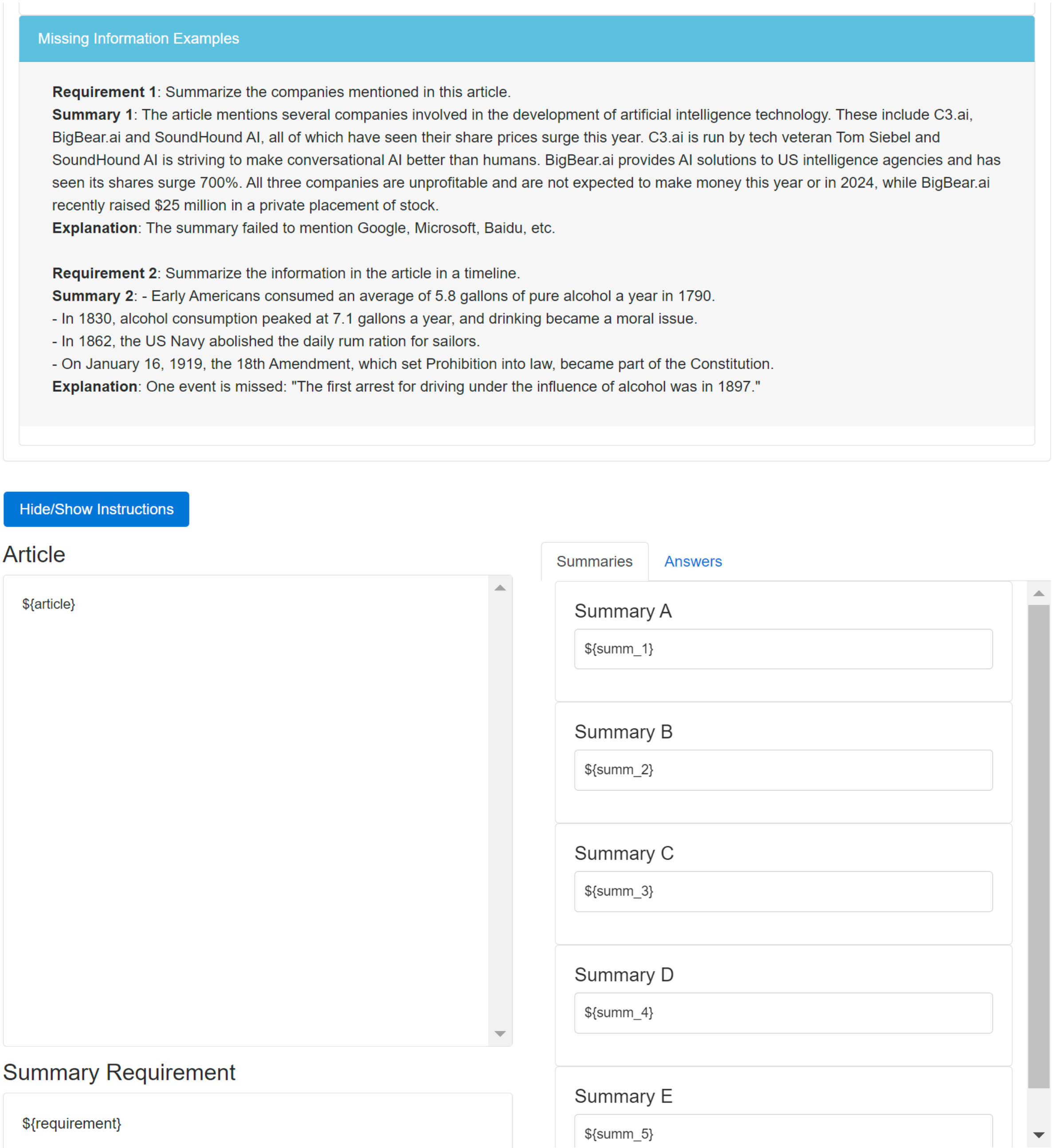

Figure 5: Annotation Interface Part 2: Data Input.

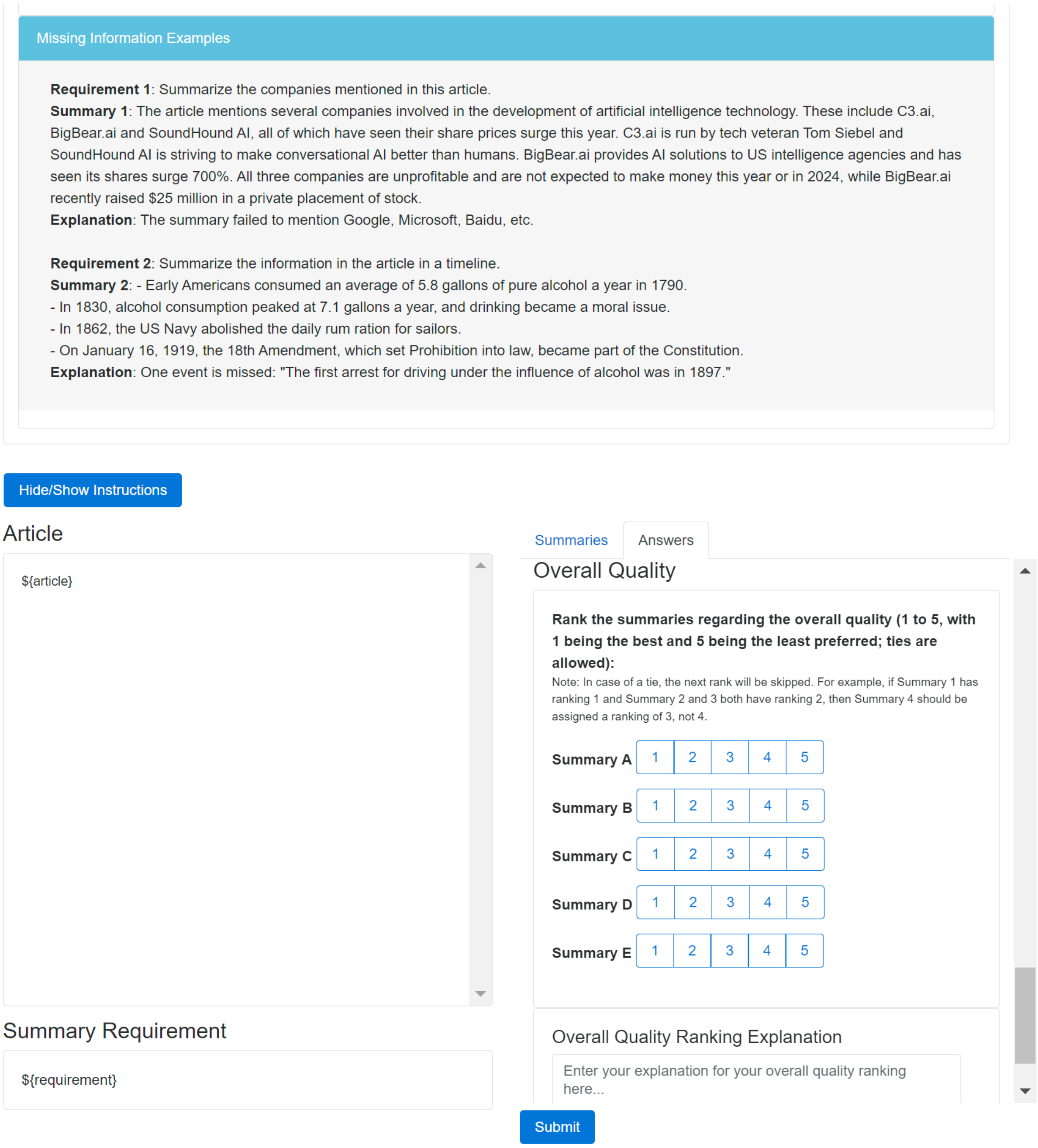

Figure 6: Annotation Interface Part 3: Result Collection.

	Overall	Missing	Irrelevant	Factual
Tie	0.45	0.47	0.50	0.90
Initial	**0.30**	**0.29**	0.20	0.03
Hybrid	0.25	0.24	**0.30**	**0.07**

Table 10: Pairwise comparison between the initial LLM summary and the hybrid summary. Winning rates of both summaries are reported. 37% of summaries are identical because no edits are made.

C OpenAI's Model Index

Here we describe the training methods of OpenAI models we benchmarked for ins-controllable summarization using human evaluation (§2.2). The following information was obtained from a blog post on the OpenAI's website, "Model index for researchers."[15]

`text-davinci-002`: Supervised fine-tuning (FeedME) on human-written demonstrations and on model samples rated 7/7 by human labelers on an overall quality score.

`text-davinci-003`: Reinforcement learning (PPO) with reward models trained from comparisons by humans.

The information about newer models can be found in https://platform.openai.com/docs/models/.

D Fine-grained Analysis of Hybrid LLM-Human Summaries

In §2.3, we found that the hybrid LLM-human summaries can not outperform the GPT-4 summaries on the *overall quality* and *missing information* dimensions. To better understand the performance of the hybrid LLM-human summary, we use the obtained human annotations to perform a pairwise comparison between the initial LLM summary and the hybrid summary. Results in Table 10 show that the hybrid summaries are better at the irrelevant information and factual consistency dimensions while worse at the overall quality and missing information dimensions. We believe this is mainly because there are more "delete" than "add" editing operations in the hybrid summaries since we found that the initial LLM summaries are more likely to include irrelevant information than missing relevant information. As a result, the reduced length of hybrid summaries may make them less favorable than the original summaries on the overall quality and missing information dimensions.

[15]The original page is no longer accessible as of April 1st, 2024. An old snapshot of the page is available at https://archive.ph/n5xMq.

E Factual Error Examples

We found that a considerable portion of the factual errors flagged by the crowd annotators is quite nuanced (§2.2). Below we present a few examples.

Example 1

- **Article (part)**: ... However, in 2007, Australian-listed firm Scotgold Resources entered the scene and revived the mine. It has been a rollercoaster ride ever since. **By 2013**, Scotgold had obtained planning permission and put a funding plan in place, only for the gold price to collapse, making the project less palatable for potential investors ...
- **Summary Error Span (with context)**: Planning permission was obtained **in 2013**.

Example 2

- **Article (part)**: ... Their efforts to attract backers were also undermined by the volatility of the precious metals market, which often saw gold prices slump. **By 2006**, the mine had changed hands several times and was up for sale once more. ...
- **Summary Error Span (with context)**: **2006**: Mine changed hands several times and was up for sale again.

Example 3

- **Article (part)**: ... There was also a divide when it came to trusting information on the virus coming from the WHO. About **one-third** of Republicans said they trusted WHO information versus 80% of Democrats. ...
- **Summary Error Span (with context)**: 80% of Democrats trust information from the WHO, while only **33%** of Republicans do.

Example 4

- **Article (part)**: ... Prof Suzuki says his team will need to carry on their work for many more years to be sure that the children of Fukushima are in the clear. But he and other experts now say they think there will be **very few, or even zero, extra childhood cancers** because of Fukushima. ...

- **Summary Error Span (with context)**: Professor Suzuki believes that the cases of thyroid cancer in Fukushima are not related to the nuclear disaster, and that the children of Fukushima **are not at risk** of developing cancer from the exposure to radiation.

Example 5

- **Article (part)**: ... Chum Mey says he was tortured, as his interrogators **tried to make him confess to spying for the US and Russia**. ... Eventually he said he **confessed to anything** so that the torture would be over. In his confession Chum Mey wrote that **he was working for the CIA** and had recruited dozens of agents in Cambodia. ...

- **Summary Error Span (with context)**: Chum Mey was tortured until he eventually confessed to spying for the US and **Russia**.

F Detailed Evaluation Settings of LLM-based Evaluation Methods

In §3, we benchmark different LLM-based evaluation methods. To this end, we use both system-level and summary-level correlations to evaluate the alignment between human evaluation and LLM-based evaluation. Specifically, given a correlation measurement function $\mathcal{M}$, e.g., the Kendall rank correlation coefficient (Kendall, 1938), and two lists of average system scores $\bar{S}^{(f)}$ and $\bar{S}^{(h)}$ assigned by two evaluation methods f and h, e.g., human evaluation and LLM evaluation, the system level correlation $\mathcal{C}_{\text{sys}}$ between f and h is

$$\mathcal{C}_{\text{sys}} = \mathcal{M}(\bar{S}^{(f)}, \bar{S}^{(h)}). \quad (3)$$

Similarly, the summary-level correlation $\mathcal{C}_{\text{summ}}$ is an average of the correlation between two lists of scores, $S_i^{(f)}$ and $S_i^{(h)}$, assigned by the evaluation methods f and h for the summaries generated by different systems on each data example:

$$\mathcal{C}_{\text{summ}} = \sum_{i=0}^{N-1} \frac{\mathcal{M}(S_i^{(f)}, S_i^{(h)})}{N}, \quad (4)$$

where N is the size of the evaluation dataset.

Since we adopted a ranking-based evaluation protocol for our human evaluation collection (§2.2), we use the Kendall rank correlation coefficient as the correlation measurement. Furthermore, apart from LLMRank, which directly generates a similar ranking, we convert the evaluation results of the other protocols to a ranking of different systems. For LLMScore and LLMEval that perform direct scoring of summaries, we simply convert the scores into a ranking (ties are allowed). For LLMCompare, we use the following scoring mechanism: (1) the winner system in a pairwise comparison receives 2 points, while the lost system receives 0 points; (2) if there is a tie between two systems, each of them receives 1 point; (3) the points are aggregated into a system ranking.

To remove the potential positional biases of the LLM-based evaluation methods (Wang et al., 2023c; Koo et al., 2023), we randomly shuffled the summary order when LLMRank or LLMCompare is used as the evaluation protocol.

G Length Bias in LLM-based Evaluation

As a further investigation of §3.3, we conducted a case study by only comparing summary pairs with similar lengths. Specifically, for each pair of systems, we keep only those pairs of summaries where the difference in lengths falls within the 20th percentile. The results in Table 11 indicate that when the length difference is controlled, none of the LLMs we compared can outperform GPT-4 on the overall quality dimension, and Llama-2 models no longer have a clear advantage over GPT-4. We note that since different examples are used for the comparison of different system pairs, the results in Table 11 are no longer directly comparable.

H Comparing Generic and Requirement-Specific Summaries

As a case study, we use `gpt-3.5-turbo-0301` and `gpt-4-0314` to generate *generic* summaries without specific requirements, and compare them with the requirement-specific summaries using `gpt-4-0314` as the evaluator with the LLMCompare protocol. The evaluation results show that the requirement-specific summaries generated by `gpt-4-0314` have a winning rate of 97% over the generic summaries on the *overall quality* dimension, while those with `gpt-3.5-turbo-0301` have a winning rate of 96%. We also evaluate the similarity between the generic and requirement-specific summaries in Table 12, as well as the summary length. In addition, in Table 13, we report the similarity between greedy-decoded and sampled (with a temperate of 1.0) requirement-specific summaries. Results in Table 12 and Table 13 sug-

	Overall			Missing			Irrelevant			Length	
	Win	Tie	Loss	Win	Tie	Loss	Win	Tie	Loss	System	GPT4
gpt-3.5-turbo-0301	0.23	0.18	0.59	0.18	0.27	0.55	0.09	0.32	0.59	118.0	117.5
gpt-3.5-turbo-0613	0.15	0.10	0.75	0.00	0.45	0.55	0.00	0.60	0.40	110.7	111.8
gpt-3.5-turbo-instruct	0.10	0.05	0.85	0.05	0.40	0.55	0.10	0.55	0.35	109.4	110.2
text-davinci-002	0.14	0.14	0.71	0.14	0.29	0.57	0.05	0.29	0.67	87.8	92.1
text-davinci-003	0.29	0.29	0.43	0.14	0.76	0.10	0.00	0.76	0.24	99.4	98.2
llama-2-7b-chat	0.38	0.10	0.52	0.24	0.29	0.48	0.24	0.33	0.43	132.1	115.1
llama-2-13b-chat	0.29	0.19	0.52	0.14	0.52	0.33	0.10	0.48	0.43	113.6	110.2
llama-2-70b-chat	0.30	0.20	0.50	0.25	0.55	0.20	0.25	0.50	0.25	110.8	102.5
mistral-instruct	0.20	0.15	0.65	0.20	0.20	0.60	0.10	0.25	0.65	91.3	90.0
gpt-4-1106-preview	0.50	0.20	0.30	0.45	0.40	0.15	0.30	0.60	0.10	98.75	91.4

Table 11: Automatic benchmarking results on summary pairs with similar lengths. The summaries of different LLMs are compared against summaries generated by GPT-4 (`gpt-4-0314`) using the LLMCompare protocol powered by `gpt-4-0314`. The number of wins, ties, and losses is reported as well as the average summary length.

Model	R1	R2	Specific	Generic
`gpt-3.5-turbo-0301`	47.21	26.30	127.3	144.0
`gpt-4-0314`	43.00	19.37	117.1	123.7

Table 12: The similarities between the generic and requirement-specific summaries as measured in ROUGE-1/2 (R1/R2). The average summary length is also reported, denoted by `Specific` and `Generic` respectively.

Model	R1	R2	Greedy	Sampled
`gpt-3.5-turbo-0301`	61.63	38.11	127.3	125.48
`gpt-4-0314`	66.16	43.20	117.1	117.7

Table 13: The similarities between the greedy-decoded and sampled requirement-specific summaries as measured in ROUGE-1/2 (R1/R2). The average summary length is denoted by `Greedy` and `Sampled` respectively.

gest that the similarity between the generic and requirement-specific summaries is relatively low, and the generic summaries are not preferred by `gpt-4-0314` despite its longer average length.