In-context Learning of Large Language Models for Controlled Dialogue Summarization: A Holistic Benchmark and Empirical Analysis

Yuting Tang[†][*], **Ratish Puduppully**[§‡], **Zhengyuan Liu**[§‡], **Nancy F. Chen**[§‡]
[†]Nanyang Technological University, Singapore [§]CNRS@CREATE, Singapore
[‡]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore

Abstract

Large Language Models (LLMs) have shown significant performance in numerous NLP tasks, including summarization and controlled text generation. A notable capability of LLMs is in-context learning (ICL), where the model learns new tasks using input-output pairs in the prompt without any parameter update. However, the performance of LLMs in the context of few-shot abstractive dialogue summarization remains underexplored. This study evaluates various state-of-the-art LLMs on the SAMSum dataset within a few-shot framework. We assess these models in both controlled (entity control, length control, and person-focused planning) and uncontrolled settings, establishing a comprehensive benchmark in few-shot dialogue summarization. Our findings provide insights into summary quality and model controllability, offering a crucial reference for future research in dialogue summarization.

1 Introduction

Abstractive dialogue summarization aims to distill human conversations into natural, concise, and informative text, and is a challenging and interesting task in text summarization (Chen and Yang, 2020; Liu et al., 2021). The major challenges come from several aspects: 1) it lacks large human-annotated datasets unlike document summarization (Feng et al., 2021), and 2) it requires responses to be not only fluent but also factually consistent (Liu and Chen, 2022; Wang et al., 2022). Moreover, in practical use cases, users may impose additional constraints on system outputs, and this task is known as controlled dialogue summarization, which requires models to be capable of coherent and flexible language generation.

In controlled dialogue summarization, users can specify desired attributes (i.e., control signals) to guide the response of language models. Previous works have explored to incorporate control signals during pre-training (Keskar et al., 2019), task-specific fine-tuning (Liu and Chen, 2021), and prompt tuning (Zhang et al., 2022b). Meanwhile, the advancements in LLMs have unveiled new paradigms. For instance, instruction tuning, which enables models to understand users' intent in natural language, is considered to be promising for conditional text generation (Zhang et al., 2023). Additionally, the emergence of in-context learning (ICL) in LLMs has recently gained attention. The ICL ability refers to learning from a few input-output pairs written in the natural language form (also called demonstrations) (Dong et al., 2023). Followed by demonstrations, a query question is appended at the end to form a complete prompt. Compared to the traditional supervised learning, ICL requires no training and only a few annotated samples. Motivated by the paradigm shift with LLMs and the challenges encountered in controlled dialogue summarization, this study answers the following two key questions:

- How is the quality of the dialogue summaries generated by LLMs via ICL?
- How is the controllability of LLMs in dialogue summarization?

We comprehensively evaluate a range of recent Large Language Models (LLMs) on the SAMSum dataset (Gliwa et al., 2019) using a few-shot framework. Our assessment covers several controlled scenarios, including entity control, length control, and person-focused planning, as well as uncontrolled settings. We establish a comprehensive benchmark for few-shot dialogue summarization in Section 2, and elaborate on the findings in Section 3. Specifically, in our experiments, we observe that LLMs can summarize dialogues reasonably given several demonstrations, and LLaMA and Alpaca achieve a factual consistency rate exceeding 90% in the automatic evaluation. Moreover, adding control

[*]Work done during internship at I^{2}R, A*STAR.

Model	Architecture	Instruction-tuned	Training Data
OPT (Zhang et al., 2022a)	Decoder-only		RoBERTa + The Pile + Reddit
OPT-IML (Iyer et al., 2022)	Decoder-only	✓	OPT-IML Bench
mT5 (Xue et al., 2021)	Encoder-Decoder		mC4
CEREBRAS-GPT (Dey et al., 2023)	Decoder-only		The Pile
LLaMA (Touvron et al., 2023)	Decoder-only		CommonCrawl + C4 + Github, etc.
Alpaca (Taori et al., 2023)	Decoder-only	✓	Instruct dataset generated by GPT-3
BLOOM (BigScience Workshop, 2022)	Decoder-only		ROOTS

Table 1: Summary of the experimented LLMs.

Summary of SAMSum Dataset	
Training Set	14,732 samples
Validation Set	818 samples
Testing Set	819 samples
Language	English
Annotation Method	Manual

Table 2: Data details of the SAMSum dataset.

Control Signal Example: Length Control

Summarize the conversation with the defined length:
Kevin: Hi, will you come to the workshop?
Elena: I have to, I will present a paper.
Kevin: Nice, I can't wait!
Summary with the length of 8 words: Elena will present a paper at the workshop.

Summarize the conversation with the defined length:
Jamilla: remember that the audition starts at 19:30.
Kiki: which station?
Jamilla: Antena 3
Yoyo: roger that
Summary with the length of 9 words: <output>

signals in prompts (particularly keywords) can effectively guide models to include key information in generated summaries.

2 Our Experimental Setting of ICL Dialogue Summarization

In this section, we describe how we establish the benchmark of evaluating LLMs' in-context learning for abstractive dialogue summarization.

2.1 Selected Models & Prompt Template

To conduct an extensive comparison, we evaluate various models that differ in architectures, training corpora, and paradigms. Previous work shows when LLMs reach a certain parameter size, their differences in performance on dialogue summarization become relatively small (Wang et al., 2023). Therefore, to balance the performance and inference latency, here we select models that are smaller than a 10B parameter size. Details of the experimented models are shown in Table 1. For a reproducible and fair comparison, consistent prompt templates are employed across all models, as detailed in Appendix A. Moreover, considering the encoder-decoder architecture of mT5, we follow the approach of Puduppully et al. (2023) for prompting bidirectional LLMs, specifically by adding control keywords and infilling text between them.

2.2 Experimental Dataset

All models are evaluated using SAMSum (Gliwa et al., 2019), a human-annotated dataset for abstractive multi-turn dialogue summarization. Table 2 lists some information about the dataset. We use samples from the test set for model evaluation. For the few-shot ICL inference, all demonstrations are randomly sampled from the training set.

2.3 ICL Inference Configuration

In this study, we consider two experiment settings: **uncontrolled** and **controlled** dialogue summarization. An uncontrolled setting is identical to a traditional summarization task without control signals. In contrast, a controlled setting involves user-provided control signals as constraints to LLMs' outputs. Here we focus on three types of control signals that are common and straightforward control aspects to users (He et al., 2022; Liu and Chen, 2021; Wang et al., 2023):

- **Entity control:** Given a set of user-specified keywords or entities, the generated summary should include them.
- **Length control:** In this case, the user determines the desired length for the summaries.
- **Personal named entity planning:** This is a specific form of entity control, where the user provides models with a sequence of personal named entities, indicating person-focused perspectives.

Control Signal Setup: To quantitatively evaluate the controllability of LLMs, we extract oracle control signals from human-annotated references (i.e.,

Model	Size	ROUGE-1	ROUGE-2	ROUGE-L	Perplexity	Factual Consistency(%)
OPT	1.3B	30.7	6.6	22.6	64.7	60.2
OPT-IML	1.3B	**34.6**	**9.9**	**27.8**	264.4	80.9
mT5-XL	3.7B	21.9	7.4	21.5	139.3	48.4
CEREBRAS-GPT	6.7B	31.5	7.4	22.4	**28.0**	66.6
LLaMA	7B	31.0	7.3	22.9	41.1	94.0
Alpaca	7B	32.0	7.1	23.7	90.8	**97.3**
BLOOM	7B	32.1	7.7	23.2	38.2	82.1
GPT3-davinci-003	175B	43.8	17.0	39.4	66.6	-

Table 3: Evaluation results in the **uncontrolled setting**. The ROUGE F-scores are reported. The optimal performance is highlighted in bold. GPT-3 serves as the factual consistency evaluator, so its factual consistency is excluded.

Model	Size	ROUGE-1	ROUGE-2	ROUGE-L	Perplexity	Success Rate(%)
OPT	1.3B	33.2	8.2	24.5	53.2	65.3 (↑ 14.8)
OPT-IML	1.3B	37.8	11.6	30.5	294.1	54.5 (↑ 9.5)
mT5-XL	3.7B	**39.8**	**15.2**	**34.6**	112.6	**100.0**
CEREBRAS-GPT	6.7B	36.0	9.7	26.0	**40.5**	73.0 (↑ 16.9)
LLaMA	7B	34.1	9.3	25.4	52.8	62.5 (↑ 14.0)
Alpaca	7B	35.9	9.6	27.1	111.9	63.4 (↑ 12.3)
BLOOM	7B	36.6	10.2	27.2	60.1	71.1 (↑ 17.0)
GPT3-davinci-003	175B	48.8	22.3	39.1	112.2	94.0 (↑ 18.8)

Table 4: Evaluation results in the **entity control setting** with 3 keywords. The ↑ symbol denotes the change of the appearance likelihood of keywords compared to the uncontrolled setting.

gold summaries), assuming the user provides the appropriate signals (He et al., 2022). For **entity control**, the top k words in every gold summary with the highest TF-IDF scores are extracted as keywords. Considering the shorter lengths of the dialogue summaries, the range of k is set as $\{1, 2, 3\}$. Table 8 shows several generated examples of entity control. For **length control**, the expected length is set equal to the length (number of words) of the gold summary. In **personal named entity planning**, the order of named entities[1] follows their occurrence in the gold summaries. The control signals are included in the prompt, and the prompt templates are shown in Appendix A.

Demonstration Selection: During few-shot inference, the prompt includes several input-output pairs followed by a query dialogue. We limit the number of demonstrations to $\{1, 2, 3\}$ due to computational constraints on the prompt's length. Demonstrations are randomly selected from the training set but are kept consistent across all models. Given the potential variance of ICL (Min et al., 2022), we repeat the generation process in 5 times using different demonstrations and report the average scores. The input-output pairs are concatenated with the query dialogue to compose the prompt.

Evaluation Metrics: Our evaluation has two primary objectives: 1) to assess the quality of the generated summaries, and 2) to measure the controllability of the models.

For assessing text-level quality automatically, we employ the **ROUGE** metric (Lin, 2004), which gauges the correspondence between the generated summaries and the reference (or gold) summaries. Following previous work (Fan et al., 2018), we also calculate the **perplexity** of model generations using GPT-2 (Radford et al., 2019), which serves as a measure of textual fluency.

Factual consistency represents another essential facet of quality. Since GPT-3 has demonstrated robust performance across various evaluation tasks (Luo et al., 2023; Fu et al., 2023; Chia et al., 2023), we utilize it as a binary natural language inference classifier. This classifier assesses **factual consistency** by determining if the generated summary aligns with the underlying dialogue. Further elaborations on this are provided in Appendix B.

Additionally, we perform automatic holistic evaluations of writing quality, focusing on **coherence** and **relevance**. Following previous work (Chia et al., 2023), where GPT-3 is used for automatic evaluations to benchmark instruction-tuned models, we instruct GPT-3 to score the generated summaries on a discrete scale of 1 to 5. To ensure comparability, we adopt the same prompt templates as in Chia et al. (2023). The average scores are pre-

[1] The personal named entities data is acquired from https://github.com/seq-to-mind/planning_dial_summ/tree/main/data (Liu and Chen, 2021).

Model	Size	ROUGE-1	ROUGE-2	ROUGE-L	Perplexity	Length Deviation
OPT	1.3B	30.7	6.5	22.2	54.2	12.4 (↓ 0.9)
OPT-IML	1.3B	**36.0**	**10.4**	**28.8**	252.0	11.7 (↓ 0.5)
mT5-XL	3.7B	21.1	5.6	18.4	102.4	10.7 (↑ 0.7)
CEREBRAS-GPT	6.7B	31.2	6.7	22.5	**31.8**	16.1 (↓ 1.4)
LLaMA	7B	33.7	8.2	24.8	57.8	12.3 (↓ 2.2)
Alpaca	7B	34.7	8.2	26.4	185.3	**7.2** (↓ 4.7)
BLOOM	7B	32.9	7.9	24.3	45.0	13.1 (↓ 1.3)
GPT3-davinci-003	175B	47.8	20.1	38.0	219.0	7.1 (↓ 12.6)

Table 5: Evaluation results in the **length control setting**. The ↑ and ↓ symbols denote the change of length deviations compared to the uncontrolled setting.

Model	Size	ROUGE-1	ROUGE-2	ROUGE-L	Perplexity	Success Rate (%)
OPT	1.3B	30.5	7.2	23.6	57.4	82.7 (↑ 4.1)
OPT-IML	1.3B	**36.5**	**11.1**	**29.5**	239.2	76.1 (↑ 4.2)
mT5-XL	3.7B	28.3	7.9	24.8	50.1	100
CEREBRAS-GPT	6.7B	32.8	8.8	24.2	**31.1**	88.0 (↑ 5.5)
LLaMA	7B	33.3	8.6	25.2	51.1	77.8 (↑ 7.1)
Alpaca	7B	33.8	8.5	25.8	102.4	76.6 (↑ 2.0)
BLOOM	7B	33.4	9.0	25.2	43.3	**89.2** (↑ 6.4)
GPT3-davinci-003	175B	47.3	21.6	36.7	65.7	96.8 (↑ 4.9)

Table 6: Evaluation results in the **person-focused planning setting**. The ↑ and ↓ symbols denote the change of length deviations compared to the uncontrolled setting.

Model	Size	Consistency (%)	Fluency	Coherence	Relevance
OPT	1.3B	60.2	64.7	**3.5**	3.2
OPT-IML	1.3B	80.9	264.4	3.4	3.2
mT5-XL	3.7B	48.4	139.3	3.3	3.0
CEREBRAS-GPT	6.7B	66.6	**28.0**	3.4	3.4
LLaMA	7B	94.0	41.1	3.4	**3.7**
Alpaca	7B	**97.3**	90.8	**3.5**	3.5
BLOOM	7B	82.1	38.2	3.4	3.5

Table 7: Holistic evaluations on the writing quality. For each aspect, the best score is in bold.

sented in Table 7, and a more detailed description is provided in Appendix C.

Regarding the model controllability, the **success rate of entity control** is determined by the proportion of specified keywords present in the generated summaries (He et al., 2022). This success rate metric is also applicable to personal named entity planning. Length controllability is evaluated by the mean absolute **length deviation**, formulated as $\frac{1}{N} \cdot \sum_{i=1}^{N} |l_{\text{generated}} - l_{\text{gold}}|$, where N represents the size of the test set, and l denotes the text length.

3 Our Findings

In this section, we analyze the experimental results by discussing about the two research problems.

3.1 How is the quality of the dialogue summaries generated by LLMs via ICL?

The experimented LLMs can generate reasonable summaries via ICL inference, but perform differently. In the uncontrolled setting, as shown in Table 3, OPT-IML 1.3B exhibits the highest ROUGE scores, despite having the fewest parameters. However, ROUGE scores cannot reflect factual correctness from the semantic view. In contrast, LLaMA and Alpaca tend to produce summaries that are the more factually consistent (i.e., their consistency in GPT3's assessment exceeds 90%). Considering other evaluation aspects, CEREBRAS-GPT consistently generates the most fluent sentences across all settings, and summaries generated by LLaMA obtain the highest score regarding relevance among all models (see Table 7).

In the controlled setting, as shown in Table 4, both entity control and person-focused planning improve ROUGE scores, and the former brings a substantial performance gain. Length control has a minor impact on ROUGE scores, especially for models without instruction-tuning. Moreover, mT5's performance is inferior to other models in non-entity control settings in terms of ROUGE scores and factual consistency. However, it achieves optimal ROUGE scores in entity control, as its architecture ensures the inclusion of keywords.

Gold Summary	Model	Uncontrolled setting	Entity control
The audition starts at 7.30 P.M. in Antena 3.	BLOOM-7b1	Jamilla and Yoyo are going to an audition for a TV show. They are waiting for the station to start broadcasting.	Jamilla and Kiki are going to an audition for a TV show. The station is Antena3.
Adam and Karen are worried that May suffers from depression. Karen will call her friend who is a psychologist and ask for advice.	mT5-XL	Melissa shows Brian what she bought and he finds it awesome.	Adam and karen talk to their friend about her depression. They talk to their friend's psychologist and she agrees to help.
Mike suspects he might have had an allergic reaction to something.	LLaMA-7B	Mike wakes up with a strange rash on his face. He has no clue what it is and asks his girlfriend Emma to help him figure it out.	Mike suspects that he is allergic to something, but he does not know what it is.
Flo cannot get an appointment at the salon until the 6th. Flo worries she's going to be gray. Flo will have to get a touch-up kit at Tesco.	CEREBRAS-6.7B	Gina and Flo are having a conversation about getting their hair done at the same time.	Flo is worried that she will be too late for her appointment at the hair salon. She will get a kit to get her hair touched up.

Table 8: Comparisons of generated summaries between controlled and uncontrolled settings. Underlined keywords are the control signals included in the prompts.

3.2 Can LLMs do controlled dialogue summarization?

Experimental results demonstrate that LLMs are capable of achieving controlled dialogue summarization via ICL. Evaluation details of LLMs in controlled settings are presented in Table 4 for entity control with 3 keywords, Table 5 for length control, and Table 6 for person-focused planning. Table 10 in Appendix provides a comprehensive evaluation with varying numbers of keywords.

In the entity control experiments, keywords are included in the prompts as constraints. All models exhibit improved ROUGE scores, and the likelihood of keywords appearing in the summaries is increased, indicating that LLMs effectively utilize the information from control signals. Notably, mT5 achieves a 100% success rate, benefiting from its bi-directional encoding architecture. Examples presented in Table 8 show how keywords can guide models to generate better summaries. Surprisingly, non-instruction-tuned models like CEREBRAS-GPT and BLOOM demonstrate better controllability than instruction-tuned models like Alpaca and OPT-IML in entity control.

The impact of length signals is relatively minor compared to keyword signals on ROUGE scores. However, the length distribution with length signals is more aligned with the actual length across models, except for mT5. Notably, Alpaca demonstrates the best length controllability. We also find that OPT-IML appears to have lower controllability compared to its foundation model, OPT.

Model	Success Rate (%)
OPT-IML-1.3B	19.0 (↑ 4.2)
LLaMA-7B	10.1 (↑ 4.7)
Alpaca-7B	7.8 (↑ 3.5)
BLOOM-7B	28.3 (↑ 17.0)

Table 9: The success rates of numerical keywords.

3.3 Further Analysis

Numerical keywords (e.g., time and quantity) tend to be left out by LLMs. Preliminary error analysis shows a large portion of the missing keywords in entity control contain numerical information. To verify that, the models are prompted with only numerical keywords (e.g., time, date, quantity, and percent) extracted from gold summaries using SpaCy (Honnibal et al., 2020). The results in Table 9 demonstrate a significant decrease in the success rates across all models. It implies that LLMs have some intrinsic bias toward non-numerical content, potentially causing them to overlook crucial numerical details within dialogues.

4 Conclusion

In this study, we have benchmarked the in-context learning performance of state-of-the-art LLMs in controlled and uncontrolled settings for abstractive dialogue summarization. We assessed their summarization quality, factual consistency, and controllability, while also conducting holistic evaluations and empirical analysis. We hope this study provides insights for the follow-up research about dialogue summarization using LLMs.

Limitations

One limitation of this study is that only LLMs with less than 10B parameters are experimented with due to hardware constraints. To address this issue, we release the evaluation codes, in order to facilitate the follow-up research.

Meanwhile, the control signals in this work are oracle, which means we assume the user provides indicative keywords to be included in the summary. There are automatic methods to extract keywords from dialogues (He et al., 2022), but it is not the focus and therefore not discussed in this study.

Due to time constraints, we adopted GPT-3 to conduct automatic qualitative evaluations. While GPT-based evaluations have proven to be competitive in some evaluation tasks, the necessity for human evaluations remains.

Acknowledgments

This research was supported by the Agency for Science, Technology and Research (A*STAR), Singapore under its Industry Alignment Pre-Positioning Fund (Grant No. H19/01/a0/023 - DCOF), and by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme. We also thank the anonymous reviewers for their precious feedback to help improve and extend this piece of work.

References

BigScience Workshop. 2022. BLOOM (revision 4ab0472).

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*.

Nolan Dey, Gurpreet Gosal, Zhiming, Chen, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, and Joel Hestness. 2023. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. A survey on dialogue summarization: Recent advances and new frontiers. *arXiv preprint arXiv:2107.03175*.

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

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: 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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Dániel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2021. Controllable neural dialogue summarization with personal named entity planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 92–106, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengyuan Liu and Nancy Chen. 2022. Entity-based de-noising modeling for controllable dialogue summarization. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 407–418, Edinburgh, UK. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy Chen. 2021. Coreference-aware dialogue summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 509–519, Singapore and Online. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy F. Chen. 2023. Decomposed prompting for machine translation between related languages using large language models.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Bin Wang, Zhengyuan Liu, and Nancy F Chen. 2023. Instructive dialogue summarization with query aggregations. *arXiv preprint arXiv:2310.10981*.

Bin Wang, Chen Zhang, Yan Zhang, Yiming Chen, and Haizhou Li. 2022. Analyzing and evaluating faithfulness in dialogue summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4897–4908.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022a. Opt: Open pre-trained transformer language models.

Yubo Zhang, Xingxing Zhang, Xun Wang, Si qing Chen, and Furu Wei. 2022b. Latent prompt tuning for text summarization.

A Appendix: Prompt template

This section includes examples of the prompt templates, which remain consistent across models.

Uncontrolled Setting

```
Summarize the conversation:
Selby: anybody for indian?
Terri: yuo cooked?
Selby: yessir
Terri: sounds cool
Winslow: gr8. ill be there too
Summary: Selby invites Terri and Winslow for a home-cooked Indian meal.

Summarize the conversation:
Marta: <file_gif>
Marta: Sorry girls, I clicked something by accident :D
Agnieszka: No problem :p
Weronika: Hahaha
Agnieszka: Good thing you didn't send something from your gallery ;)
Summary:
```

(Last line of mT5)

```
Summary: <extra_id_0>
```

Length Control

```
Summarize the conversation with the defined length:
Kevin: Hi, will you come to the workshop?
Elena: I have to, I will present a paper.
Kevin: Nice, I can't wait!
Summary with the length of 8 words: Elena will present a paper at the workshop.

Summarize the conversation with the defined length:
Jamilla: remember that the audition starts at 7.30 P.M.
Kiki: which station?
Jamilla: Antena 3
Yoyo: roger that
Summary with the length of 9 words:
```

(Last line of mT5)

```
Summary with the length of 9 words: <extra_id_0>
```

Entity control (non-mT5)

```
Summarize the conversation with keywords:
Kevin: Hi, will you come to the workshop?
Elena: I have to, I will present a paper.
Kevin: Nice, I can't wait!
Summary with keywords ['Elena', 'workshop']: Elena will present a paper at the workshop.

Summarize the conversation with keywords:
Jamilla: remember that the audition starts at 19:30.
Kiki: which station?
Jamilla: Antena 3
Yoyo: roger that
Summary with keywords ['audition', 'antena']:
```

Entity Control / Person-focused Planning (mT5)

```
Summarize the conversation:
Selby: anybody for indian?
Terri: yuo cooked?
Selby: yessir
Terri: sounds cool
Winslow: gr8. ill be there too
Summary: Selby invites Terri and Winslow for a home-cooked Indian meal. Selby invites Terri and Winslow for a home-cooked Indian meal.</s>

Summarize the conversation:
Marta: <file_gif>
Marta: Sorry girls, I clicked something by accident :D
Agnieszka: No problem :p
Weronika: Hahaha
Agnieszka: Good thing you didn't send something from your gallery ;)
Summary: <extra_id_0> marta <extra_id_1> file <extra_id_2>
```

Person-focused Planning (non mT5)

```
Summarize the conversation with the focus perspectives provided:
Hugo: Im off to Mazagan (Morocco) this weekend
Lacey: Oh so nice!
Hugo: Where are you these days? In Spain already?
Lacey: No. In Poland all the time. I study here
Summary with the focus perspectives ['Hugo', 'Lacey']: Hugo is off to Mazagan in Morocco this weekend. Lacey studies in Poland.

Summarize the conversation with the focus perspectives provided:
Miley: don't want to go to work tomorrow!
Miley: wanna sleep!!!
Miley: need some rest :(
Aaron: know what u feelin
Aaron: really want some holiday already
Miley: :(
Aaron: fighting girl!
Summary with the focus perspectives ['Miley']:
```

B Appendix: Utilizing GPT-3 for factual consistency evaluation

This study employs text-davinci-003 model as the factual consistency evaluator. The prompt template adopts a zero-shot chain-of-thought style because it outperforms direct prompting and exhibits greater stability compared to the few-shot setup, as suggested in Luo et al. (2023).

An illustrative example of the prompt and GPT-3's response is provided below:

Prompt for factual consistency classification

Decide if the following summary is consistent with the corresponding dialogue. Note that consistency means all information in the summary is supported by the dialogue.
Dialogue:
Will: hey babe, what do you want for dinner tonight?
Emma: gah, don't even worry about it tonight
...
Will: Alright, love you.
Emma: love you too.
Summary: Emma doesn't want to cook for Will tonight.
Explain your reasoning step by step then answer (yes or no) the question:

GPT-3's response
Yes, the summary is consistent with the dialogue.
Step 1: Will asks Emma what she wants for dinner.
Step 2: Emma responds that she doesn't want to worry about it, implying that she does not want to cook.
Step 3: Will then asks what time she'll be HOME, and Emma says soon.
Step 4: Will then offers to pick her up to which Emma responds no, she'll tell him when she gets home.
Step 5: Will then tells her he loves her and she responds in kind.
Yes, the summary is consistent with the dialogue.

C Appendix: Holistic Evaluation

The holistic evaluation of writing quality utilizes GPT-3. We follow the evaluation templates as well as the rubrics in Chia et al. (2023). See details of the prompts for writing an evaluation of relevance and coherence on the next page.

Due to the API cost, each model is evaluated on 100 random samples from the test set.

D Appendix: Impact of the number of demonstrations

ICL's performance instability is influenced by the selection and quantity of demonstrations. This study employs a random selection strategy for efficiency and simplicity, with each test sample having five sets of demonstrations. The average values are reported.

Table 10 and Figure 1 show the metric trends for k demonstrations, indicating that an increase in demonstrations may degrade performance and increase instability in some models (e.g., Cerebras-GPT-6.7B, OPT-1.3B), possibly due to their small parameter sizes. Table 3 presents the optimal performance for k values in the set 1,2,3.

Writing evaluation on relevance

Text: Eric, Bella and Eric were talking about their boss and how he appreciated their decision of dismissing a potential client. Eric and Bella were discussing the reasons why they dismissed the client, while Eric was asking Bella about her reaction to his boss' reaction.

Prompt: Summarize the following dialogue:
Eric: Hey Bella, What happened today in boss's room?? Was he angry??
Bella: NO NO!!! He wasn't angry at all.. He actually appreciated on our brave deccision to dismiss the request of client..
Eric: REALLY!! He appreciated this decision.. Bella: Yeah he really did.. I too was astounded by his reaction...
Eric: What could possibly lead to this?? I mean , they were potential clients...
Bella: What he told me was that he was looking forward to bring in new clients which were our current client's competitor..
Eric: Oh that could possibly be the reason.Well anyways you got appreciation xD congo
Bella: hahaha Blessing in disguise xD

How relevant is the text to the prompt? Select a suitable option number between 1 and 5 based on the options below.

1. Inadequate: The text fails to provide any relevant information or insights related to the given prompt.
2. Limited: The text may contain some relevant information, but significant gaps exist, and key aspects of the prompt are not adequately covered.
3. Satisfactory: The text covers the main aspects of the prompt and provides relevant information, but it lacks depth and may not explore the topic in great detail.
4. Proficient: The text provides a comprehensive response by addressing the key aspects of the prompt, offering relevant and well-supported information or arguments.
5. Excellent: The text thoroughly and thoughtfully addresses the prompt, demonstrating a comprehensive understanding of the topic. It offers insightful and original ideas, supported by relevant arguments and information.

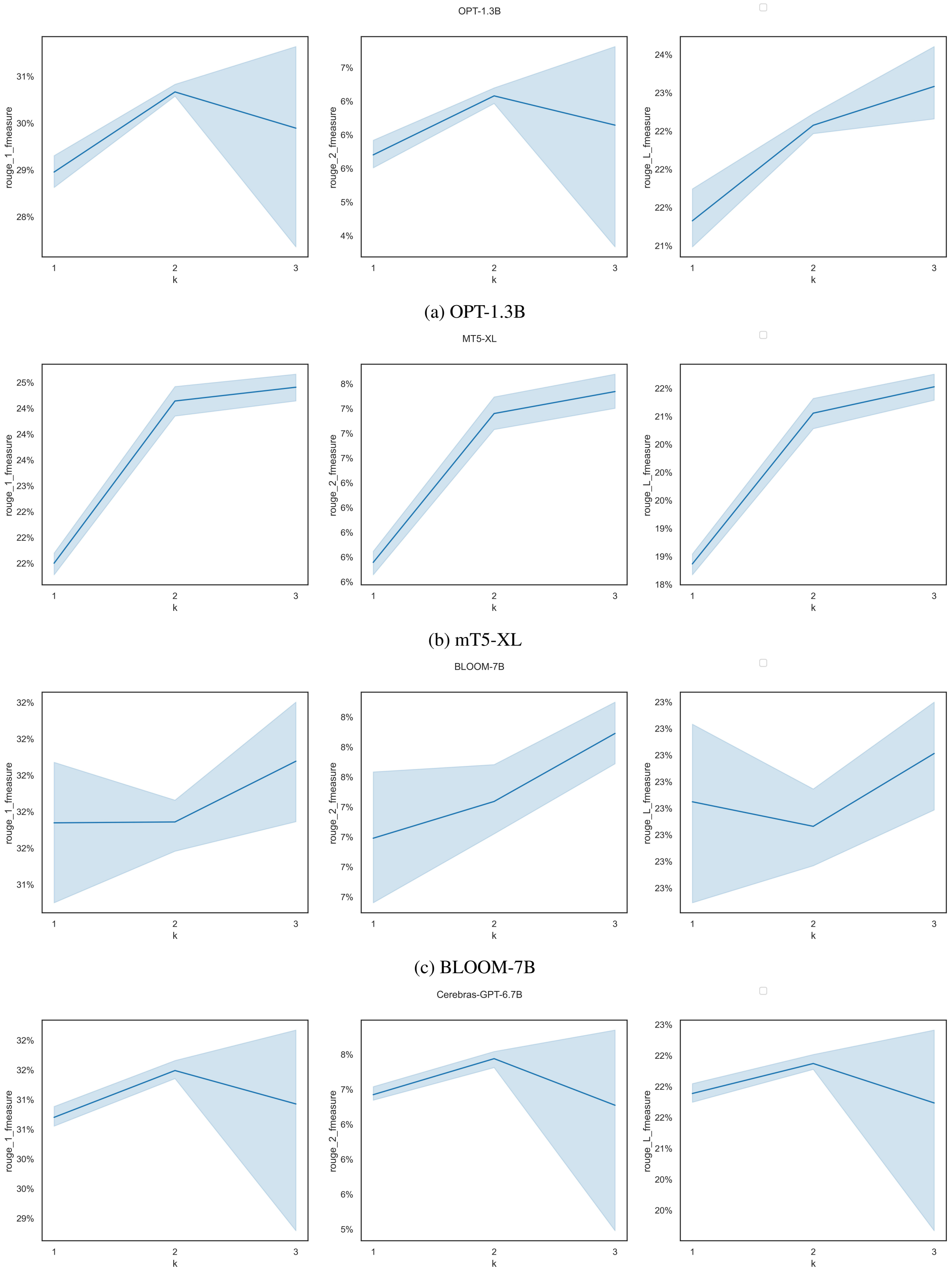
Writing evaluation on coherence

Text: Eric, Bella and Eric were talking about their boss and how he appreciated their decision of dismissing a potential client. Eric and Bella were discussing the reasons why they dismissed the client, while Eric was asking Bella about her reaction to his boss' reaction.
How coherent is the text? Select a suitable option number between 1 and 5 based on the options below.

1. Inadequate: The text lacks logical organization, making it difficult to follow. Ideas are disjointed and phrased awkwardly, requiring significant effort to understand.
2. Limited: The text demonstrates some attempt at organization, but there are significant gaps in coherence. Ideas may be loosely connected, and the arguments lack clarity.
3. Satisfactory: The text generally follows a logical organization, but occasional disruptions or awkward phrasing may occur. There is an acceptable level of readability and understanding.
4. Proficient: The text is clearly organized and easy to understand. Ideas and arguments flow smoothly, contributing to easy comprehension and a pleasant reading experience.
5. Excellent: The text presents exceptionally coherent writing with a fluent and engaging flow of ideas, ensuring effortless comprehension and a delightful reading experience.

Model	***k***	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**Succee Rate (%)**
OPT-1.3B	1	30.0	6.4	22.1	71.7 (↑ 10.7)
	2	32.1	7.7	23.7	68.7 (↑ 14.1)
	3	33.2	8.2	24.5	65.3 (↑ 14.8)
OPT-IML-1.3B	1	36.5	11.0	30.0	61.9 (↑ 7.5)
	2	36.9	11.1	29.6	57.4 (↑ 8.7)
	3	37.8	11.6	30.5	54.5 (↑ 9.5)
mT5-XL	1	32.3	11.0	27.5	100.0
	2	36.3	13.0	31.4	100.0
	3	39.8	15.2	34.6	100.0
Cerebras-GPT-6.7B	1	32.6	7.7	23.2	79.6 (↑ 13.9)
	2	33.9	8.6	24.5	74.9 (↑ 15.1)
	3	36.0	9.7	26.0	73.0 (↑ 16.9)
LLaMA-7B	1	32.0	7.8	23.4	69.1 (↑ 12.6)
	2	33.6	8.8	24.8	65.2 (↑ 13.7)
	3	34.1	9.3	25.4	62.5 (↑ 14.0)
Alpaca-7B	1	33.3	7.6	24.7	67.6 (↑ 8.5)
	2	35.2	8.9	26.3	65.4 (↑ 11.5)
	3	35.9	9.6	27.1	63.4 (↑ 12.3)
BLOOM-7B	1	32.2	7.6	23.1	77.1 (↑ 12.2)
	2	34.9	9.2	25.7	73.0 (↑ 14.8)
	3	36.6	10.2	27.2	71.1 (↑ 17.0)

Table 10: Evaluation results in the entity control setting with k keywords.



OPT-1.3B

(a) OPT-1.3B

MT5-XL

(b) mT5-XL

BLOOM-7B

(c) BLOOM-7B

Cerebras-GPT-6.7B

(d) CEREBRAS-GPT-6.7B

Figure 1: The line plots of evaluation metrics given k demonstrations in the uncontrolled setting. 95% confidence interval is highlighted within the plots.