Towards Large Language Model Organization: A Case Study on Abstractive Summarization

Krisztián Boros
Human Resocia
ite001458@athuman.com

Masafumi Oyamada
NEC Corporation
oyamada@nec.com

***Abstract*—In this work we propose ”LLM Organization”, an organizational structure-based LLM workflow for improving the performance of standard abstractive summarization techniques and mitigate unfaithful summary generation. We formulated the organizational structure-based LLM workflow as a directed acyclic graph (DAG), where each node corresponds to an LLM and each edge to a communication protocol. Our workflow is benchmarked on 5 datasets from various domains, using 7 evaluation metrics. The results indicate that LLM Organization could mitigate unfaithfulness and increase the overall performance of abstractive summarization methods.**

***Index Terms*—abstractive summarization, large language models, faithfulness, gpt-3.5-turbo, benchmarking**

I. Introduction

The recent years have witnessed the rapid development of methodologies based on Large Language Models (LLMs), including Question Answering, Sentiment Analysis, Chatbots, and Summarization, among others [1] [2]. While these novel approaches yield state-of-the-art results across various tasks, they exhibit significant shortcomings such as hallucinations, high computational costs, or gender bias [3] [4]. Specifically, in the context of summarization tasks, there is a risk that LLMs may ”hallucinate” facts, i.e. incorporate factually incorrect[1] information in the generated summary [19]. This issue is more pronounced in abstractive summarization, where, in contrast to extractive summarization, the model is expected to create a summary containing rewritten sentences.

In this work, we introduce ”LLM Organization,” an organizational structure-based LLM workflow designed to enhance the performance of standard LLM summarization techniques and mitigate unfaithful summary generation.

To thoroughly evaluate our approach, we benchmark our workflow on 5 datasets from diverse domains, using 7 evaluation metrics. These datasets originate from distinct domains, including patents, news, and dialogues; while the evaluation metrics span various quality dimensions such as similarity, informativeness, and factuality.

Throughout our analysis, we categorize our datasets based on length and communication type, highlighting the strengths and weaknesses of our proposed method.

Our main contributions are as follows:

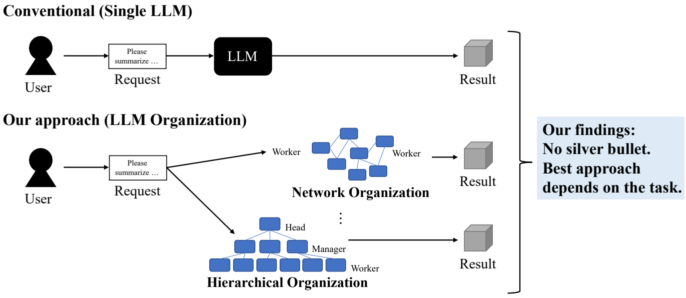

Fig. 1. Overview of the paper.

- We create an Organization-based LLM workflow to enhance the quality of the generated summaries and mitigate unfaithfulness.
- The proposed workflows are benchmarked in a diverse set of benchmark datasets and metrics, ensuring a comprehensive evaluation and a nuanced understanding of our method's performance.
- Our experiments underscore the significance of information extraction in creating more faithful summaries.

II. Methodology

Building on existing work [5], we formulated the organizational structure-based LLM workflow as a directed acyclic graph (DAG), where each node corresponds to an LLM[2] and each edge represents a communication protocol. A node can be either a ”worker” or a ”manager”, depending on its role in the organization. For the communication protocol, we employed basic prompting strategies. In our setup, workers perform information extraction, while managers handle revision and summarization. We benchmarked 2 single-node and 4 Organization-based LLM workflows.

A. Single-node methods

In the single-node workflows there is only one node doing the summarization.

1) Single Ordinary Manager: The Single Ordinary Manager (SOM) workflow is equivalent to a single request of ”Create a concise summary of the below text”.

2) Single Smart Manager: The Single Smart Manager (SSM) workflow also features a single manager node, but the

IEEE International Conference on BigData. Sorrento, Italy. 2023.

[1] Throughout the paper, we use faithfulness and factuality interchangeably

[2] We used OpenAI's `gpt-3.5-turbo-0613` model as a backbone model, source

prompt includes information extraction steps.

B. Organization-based methods

The Organization-based workflows consist of 7 worker nodes responsible for information extraction and 1 to 4 manager nodes, depending on the specific workflow. Each node in the organization has a role and task description in its prompts.

1) Basic LLM Organization: The most basic Organization-based workflow is Basic LLM Organization (LLMOrg), which includes 7 information extraction worker nodes and 1 manager node for summarization. The inner workings of this workflow can be described as follows.
Initially, the input text (source text we want to summarize) is given to the worker nodes separately. Each worker node is responsible for extracting a specific type of information[3] (e.g., keywords). Using the extracted information, the manager node summarizes the input and outputs the generated summary.

2) LLM Organization with Abstract Goals: In the case of LLM Organization with Abstract Goals (LLMOrg-abs), the flow of the input is the same. The only difference between LLMOrg-abs and LLMOrg is the prompt of the worker nodes. For LLMOrg, the workers are only prompted to extract the given number of relevant information (e.g., 5 keywords). In LLMOrg-abs, however, the instruction for workers explicitly states that the goal of the whole information extraction process is to gather useful information for better summarization. This gives an "abstract goal" for the workers to keep in mind.

3) LLM Organization with Specific Goals: The LLM Organization with Specific Goals (LLMOrg-spec) improves upon the abstract goal description of LLMOrg-abs by specifying the quality dimensions that each worker should pay attention to. For example, the task description of a date extraction worker node explicitly specifies that date extraction could help the summary be more faithful and informative. We defined quality dimensions and the corresponding information based on the works of (add references).

4) LLM Organization with Reduction Managers: In the LLM Organization with Reduction Managers (LLMOrg-man), there is an additional supervision/revision stage before summarization. After the worker nodes extract the relevant information from the text, 3 manager nodes revise and rewrite the information based on the input text. After revision, the final manager node summarizes the input text. In this workflow, the worker nodes are prompted as in LLMOrg (basic instructions), and the 3 intermediate manager nodes as in LLMOrg-abs (abstract goals).

III. Datasets and Metrics

A. Datasets

The benchmark datasets contain texts from news, patents, chats, emails, comments, and forum discussions (see Table I). During the dataset selection phase, we prioritized those datasets that contained human-generated reference summaries to assess our models more precisely and avoid low-quality references that may introduce bias [6] [11].

TABLE I
Dataset statistics

Dataset	N	Mean token number	Median token number	Human communication	Length category
Stanford News	109	861	849	no	short
BigPatent	100	5355	3025	no	long
DialogSum	100	178	163	yes	short
SAMsum	100	135	100	yes	short
ConvoSumm	400	1145	938	yes	long

1) "Stanford News" (SF): In their influential paper on news summarization, researchers from Stanford and Columbia University created a news summarization dataset based on the XSUM and CNN/Daily Mail datasets [6]. In this dataset, which we will refer to as "Stanford News," every reference summary is human-generated, providing high-quality benchmark possibilities.

2) DialogSum (DS): A dataset consisting of real-life scenario dialogues with human-generated abstractive summaries [8].

3) SAMSum (SS): The SAMSum dataset is another popular choice for dialogue summarization. The dialogues and reference summaries are created by linguists and can involve multiple interlocutors [9].

4) ConvoSumm (CS): ConvoSumm is a conversational dataset containing four sub-domains: news comments, Reddit discussions, email threads, and StackOverflow question-answering threads. The reference summaries were created by crowdsourced human workers [10].

5) BigPatent (BP): The BigPatent dataset was collected from patent applications. The patent description serves as the source text, and the abstract serves as the summary [7].

B. Metrics

For measuring similarity/n-gram overlap, we used ROUGE-1/2/L, BERTScore (F1, precision, recall), and METEOR. To evaluate faithfulness and informativeness, we used BLANC-Help, SummaQA (F1, confidence), and SummaC. Additionally, we employed GPT-Eval (add reference) (with gpt-3.5-turbo-0613) as a proxy for human evaluation.

1) ROUGE: This popular metric measures the lexical overlap (n-gram) between the source and reference summary. We use the ROUGE-1, ROUGE-2, and ROUGE-L versions [12].

2) METEOR: The METEOR score is similar to ROUGE but includes more semantic information, and adds an additional penalty term [13].

3) BERTScore: BERTScore also measures the similarity between the source and reference summary but uses pretrained models for comparing the two texts [14]. It can capture semantic similarity better than ROUGE or METEOR. We use every piece of information from BERTScore: F1-score, Precision, and Recall.

4) BLANC-Help: BLANC-Help is a variant of the BLANC metric used for assessing informativeness and factual consis-

[3] We consider the following 7 information categories: keywords, topic, dates, events, event relations, entities, entity relations

tency in summarization [15]. It concatenates the summary to the source text and measures ”helpfulness” of the summary during a language understanding task.

5) SummaQA: SummaQA is a Question-Answering (QA)-based method for assessing factual consistency [16]. The metric generates question-answer pairs from the source text by masking named entities, then uses the generated summary to infer the answers.

6) SummaC: SummaC is a Natural Language Inference (NLI)-based method for assessing factual consistency [17]. It splits the source text and generated summary into sentences and checks whether the latter is a logical continuation of the former.

7) GPT-Eval: We use the method proposed by [18] as a proxy for a human-like quality assessment. GPT-Eval assesses the generated summary from four dimensions: consistency, coherence, fluency, and relevance. Following the suggestion of [6], we exclude fluency from the evaluation. We should note that since our workflows also use gpt-3.5-turbo, G-Eval results might be biased.

IV. Experiments

For the experiments we use `gpt-3.5-turbo-0613` model for each node. We run every experiment 5 times to get more reliable estimates. Temperature of the underlying model is held at 0 for more consistent summary generation. For detailed results, see Table II.

V. Results

Our experiments show that organizational structure-based LLM workflows generate more faithful summaries than single-node methods regardless of source text length or communication type. For longer texts, organizational structure-based LLM methods perform better or on par with single-node methods, including similarity-based metrics (BigPatent, ConvoSumm). For datasets consisting of human-communication-based texts, single-node methods perform better or similarly to organizational structure-based workflows in similarity-based metrics (DialogSum, SAMSum).

VI. Conclusion

The results indicate that organizational structure-based workflows could mitigate unfaithfulness and increase the overall performance of abstractive summarization methods. Additionally, our experiments indicate that organization structure-based methods can handle longer texts better than single-node methods. Single-node methods seem to perform better on shorter, human-communication-based texts but still lag behind organizational structure-based methods in factuality. Furthermore, our results highlight the need for the assessment of abstractive summarization systems in more diverse domains and with more quality reference summaries. We hope that our approach can provide useful insights to develop more factually consistent and high-quality models for abstractive summarization.

References

[1] W. X. Zhao et al., “A Survey of Large Language Models.” arXiv, May 07, 2023. http://arxiv.org/abs/2303.18223

[2] T. B. Brown et al., “Language Models are Few-Shot Learners.” arXiv, Jul. 22, 2020. http://arxiv.org/abs/2005.14165

[3] Z. Ji et al., “Survey of Hallucination in Natural Language Generation,” ACM Comput. Surv., vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

[4] A. Abid, M. Farooqi, and J. Zou, “Persistent Anti-Muslim Bias in Large Language Models,” Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 298–306, Jul. 2021, doi: 10.1145/3461702.3462624.

[5] D. Dohan et al., “Language Model Cascades,” 2022, doi: 10.48550/ARXIV.2207.10342.

[6] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, “Benchmarking Large Language Models for News Summarization,” 2023, doi: 10.48550/ARXIV.2301.13848.

[7] E. Sharma, C. Li, and L. Wang, “BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization.” arXiv, Jun. 09, 2019. http://arxiv.org/abs/1906.03741

[8] Y. Chen, Y. Liu, L. Chen, and Y. Zhang, “DialogSum: A Real-Life Scenario Dialogue Summarization Dataset,” Association for Computational Linguistics, 2021, pp. 5062–5074. doi: 10.18653/v1/2021.findings-acl.449.

[9] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer, “SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization,” in Proceedings of the 2nd Workshop on New Frontiers in Summarization, 2019, pp. 70–79. doi: 10.18653/v1/D19-5409.

[10] A. Fabbri et al., “ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining,” Association for Computational Linguistics, Aug. 2021, pp. 6866–6880. doi: 10.18653/v1/2021.acl-long.535.

[11] E. Clark et al., “SEAHORSE: A Multilingual, Multifaceted Dataset for Summarization Evaluation,” May 2023.

[12] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in Text Summarization Branches Out, Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. https://aclanthology.org/W04-1013

[13] S. Banerjee and A. Lavie, “METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments,” presented at the IEEvaluation@ACL, Jun. 2005.

[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” Apr. 2019. https://arxiv.org/abs/1904.09675

[15] O. Vasilyev, V. Dharnidharka, and J. Bohannon, “Fill in the BLANC: Human-free quality estimation of document summaries,” in Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems, Online: Association for Computational Linguistics, 2020, pp. 11–20. doi: 10.18653/v1/2020.eval4nlp-1.2.

[16] T. Scialom, S. Lamprier, B. Piwowarski, and J. Staiano, “Answers Unite! Unsupervised Metrics for Reinforced Summarization Models,” 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), Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3246–3256. doi: 10.18653/v1/D19-1320.

[17] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, “SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization,” Transactions of the Association for Computational Linguistics, vol. 10, pp. 163–177, Feb. 2022, doi: 10.1162/tacl_a_00453.

[18] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, “G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment,” 2023, doi: 10.48550/ARXIV.2303.16634.

[19] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On Faithfulness and Factuality in Abstractive Summarization,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1906–1919, 2020, doi: 10.18653/v1/2020.acl-main.173.

TABLE II

Experiment results

		Similarity Metrics							G-Eval			Factuality Metrics			
Method	Dataset	ROUGE (1/2/L)			BERTScore (F1/Prec/Rec)			METEOR	Coh.	Con.	Rel.	BLANC	SummaQA (Conf./F1)		SummaC
SOM	SF	**0.39**	**0.15**	**0.26**	**0.83**	**0.82**	0.83	0.30	4.05	4.24	4.24	0.13	0.05	0.05	0.45
	DS	**0.37**	0.15	0.29	**0.85**	**0.83**	**0.87**	0.28	**4.19**	4.37	4.37	0.29	0.13	0.03	0.27
	BP	0.33	0.08	0.20	0.80	0.81	0.79	0.27	3.81	4.16	4.15	0.08	0.02	0.00	0.59
	SS	0.37	0.15	0.29	0.83	0.80	0.86	0.26	4.04	4.30	4.28	0.21	0.10	0.13	0.30
	CS	0.29	0.07	0.18	0.80	0.80	0.79	0.25	3.90	4.18	4.14	0.09	0.04	0.02	0.39
SSM	SF	0.37	0.14	0.24	0.82	0.81	0.83	0.28	4.07	4.23	4.26	0.13	0.06	0.05	0.47
	DS	0.35	0.14	0.28	0.84	0.82	0.86	0.27	4.18	4.39	4.37	0.28	0.14	**0.04**	0.28
	BP	0.32	0.08	0.20	0.80	0.80	0.79	0.27	3.83	4.18	**4.19**	0.08	0.03	0.00	0.59
	SS	**0.40**	0.16	**0.31**	0.83	**0.81**	0.85	**0.31**	3.87	4.29	4.21	0.18	0.09	0.12	0.28
	CS	0.28	0.06	0.18	0.79	0.79	0.78	0.23	3.95	4.19	4.19	0.09	0.04	0.02	0.40
LLMOrg	SF	0.38	0.14	0.25	0.82	0.81	0.84	0.27	4.07	4.26	**4.26**	**0.14**	**0.06**	0.05	0.46
	DS	0.34	0.13	0.28	0.83	0.82	0.85	0.28	4.15	4.37	4.34	0.26	0.13	0.03	0.27
	BP	0.34	0.10	**0.21**	0.80	0.80	0.80	0.29	3.78	4.16	4.15	0.10	0.03	**0.01**	0.61
	SS	0.36	0.15	0.28	0.83	0.80	**0.87**	0.26	4.01	4.32	4.28	0.22	**0.12**	**0.16**	0.29
	CS	0.28	0.06	0.18	0.79	0.79	0.78	0.23	3.95	4.19	4.19	0.09	0.04	0.02	0.40
LLMOrg-abs	SF	0.38	0.15	0.25	0.82	0.81	0.84	0.28	4.07	4.25	4.25	0.13	0.06	0.05	0.46
	DS	0.35	0.14	0.28	0.83	0.82	0.85	0.28	4.14	4.36	4.35	0.26	0.13	0.03	0.27
	BP	0.35	0.09	0.21	0.80	0.80	0.80	0.29	3.80	4.18	4.16	0.10	**0.03**	**0.01**	0.62
	SS	0.36	0.14	0.27	0.83	0.80	0.87	0.25	4.01	4.33	4.27	**0.23**	0.11	0.13	0.29
	CS	0.29	0.07	0.18	0.79	0.79	0.79	0.24	3.92	4.17	4.16	0.11	0.05	0.03	0.41
LLMOrg-spec	SF	0.38	0.14	0.25	0.82	0.81	0.83	0.31	4.06	4.25	4.25	0.13	0.06	0.05	**0.47**
	DS	0.35	0.14	0.28	0.84	0.82	0.86	0.29	4.18	4.35	4.35	**0.31**	0.14	**0.04**	**0.29**
	BP	**0.35**	**0.10**	**0.21**	**0.80**	0.81	**0.80**	**0.30**	3.80	4.18	4.16	**0.10**	**0.03**	0.01	0.61
	SS	0.36	0.15	0.28	0.83	0.79	0.87	0.25	4.02	4.33	4.29	0.22	0.12	0.14	0.29
	CS	0.29	0.07	0.18	0.79	0.78	0.79	0.24	**3.98**	4.19	**4.20**	**0.11**	**0.05**	0.03	**0.41**
LLMOrg-man	SF	0.37	0.14	0.24	0.81	0.81	0.83	0.30	4.05	4.26	4.24	0.13	0.06	0.05	0.45
	DS	0.35	0.14	0.28	0.83	0.82	0.85	**0.31**	4.14	4.35	4.35	0.29	0.13	0.04	0.27
	BP	0.35	0.10	**0.21**	0.80	0.81	0.80	0.29	3.79	4.15	4.16	0.10	**0.03**	0.01	0.59
	SS	0.37	0.16	0.28	0.83	0.80	**0.87**	0.26	4.01	4.31	4.28	0.22	0.11	0.14	0.27
	CS	0.29	0.07	0.18	0.78	0.78	0.79	0.24	3.95	4.16	4.19	0.11	0.05	0.03	0.40

Best result for each metric per dataset is underlined. Significantly best result is also **bolded** ($p < 0.05$).