大規模言語モデルを用いた二段階要約における hallucination の分析

榎本昌文 [1] 竹岡邦紘 [1] 定政邦彦 [1] 小山田昌史 [1]
Kiril Gashteovski[2] Chia-Chien Hung[2] Wiem Ben Rim[2] Zhao Xu[2] Carolin Lawrence[2]
[1] 日本電気株式会社 データサイエンスラボラトリー
[2]NEC Laboratories Europe
{masafumi-enomoto,k_takeoka,sadamasa,oyamada}@nec.com
{Kiril.Gashteovski,chia-chien.hung,Wiem.Ben-Rim}@neclab.eu
{zhao.xu,carolin.lawrence}@neclab.eu

概要

大規模言語モデル (LLM) は高品質な要約を生成する一方，hallucination と呼ばれる入力文書に含まれない情報を生成する傾向があるため，要約が依拠する情報を利用者が検証できる仕組みが必要である．そのため，入力文書から要約に用いるテキストを先に抽出して，それらを合成することで説明可能性の高い要約を生成する段階的な方式が提案されている．しかし，入力文書の異なる箇所から抽出されたテキストを合成するため，合成を行うモデルが元々の文脈を理解できずに，誤った情報を生成する可能性がある．本論文では二段階要約において LLM でテキスト合成を行う際に発生する hallucination の現象分析を行う．実験結果から，対話文書・長文文書における hallucination の増減の傾向や， 指示学習・周辺文脈の追加による hallucination の抑制など，既存研究では不明であった新しい知見が得られた．

1 はじめに

大規模言語モデル (LLM) の台頭によって，これまでにない高品質な要約が自動的に生成できるようになった．その品質は，人手で書かれた要約と同等，もしくはそれ以上と報告されている [1, 2]．しかし，LLM には hallucination と呼ばれる入力文書に含まれない情報を生成する傾向がある [3]．例えば，現在公開されている強力なオープンソース LLM である Llama-2 [4] であっても，生成されたニュース要約において約 5% 程度の hallucination が含まれることが分かっている [1)]．従って LLM を安全に活用するためには，生成された要約が依拠する情報を，利用者が検証できる仕組みが求められている．

1） https://github.com/vectara/hallucination-leaderboard

表 1 二段階要約における hallucination の分析に関する既存研究の比較表．Koh et al. [5] は指示学習が行われていない要約特化モデルのみで評価を行っている．Zhang et al [6] は単一のブラックボックス LLM のみで評価を行っている．また両研究とも対話要約タスクにおける分析は行っていない

	要約/合成モデル			対象文書	
	OSS LLM	Blackbox LLM	指示学習無し LM	長文	対話
Koh et al. [5]			✓	✓	
Zhang et al [6]		✓		✓	
本研究	✓	✓	✓	✓	✓

要約の説明可能性を向上させる方法として，入力文書から要約に用いるテキストを先に抽出して，それらを合成することで要約を生成する方式が提案されている [7, 8]．この二段階方式を用いることで，抽出されたテキストを要約が依拠する情報として利用者に提示できる．この方式は，特に入力文書が長くて一目で確認できないような使用事例，例えば法律文書や科学論文の要約において有用である．しかし，この方式は入力文書の異なる箇所から抽出されたテキストを合成するため，合成を行う言語モデルが元々の文脈を把握できずに，誤った情報を生成する可能性がある．例えば，入力文書から抽出されたテキストそのものを要約として提示する抽出型要約であっても，連続しない文を結合するため，誤った代名詞の参照関係や文の接続関係が発生することが指摘されている [9]．そのため，二段階方式の合成モデルとして LLM を用いた際の hallucination の現象分析が必要とされている．

二段階要約における hallucination の発生は要約タスクや合成モデルとして用いる LLM の種類に依存すると考えられるが，既存研究の現象分析は多様な条件の下で実験が行われていない（表 1 を参照されたい）．Koh et al. [5] は，長文要約タスクにおいて抽出済みテキストで追加学習された Pegasus モデル [10] は，テキスト抽出を行わないデータで学習された同モデルよりも，hallucination を生成しやすいと報告している．一方で Zhang et al [6] は，抽出済みテキストを用いて ChatGPT [2] に要約させることで，全文を入力して要約させるより hallucination が減少すると報告している．前者は指示学習済みの LLM における分析ではないし，後者は単一のブラックボックスなモデルでしか分析を行っていない．また両研究とも対話要約タスクを含む多様なタスクにおける分析が不足している．

以上の背景から，本論文では大規模言語モデルを用いた二段階要約における hallucination の現象分析を行った．実験結果から，以下の新たな知見が明らかになった: **(1)** テキスト抽出によって平均的には hallucination が増加する **(2)** 対話要約データセットである DialogSum [11] においては特に増加傾向が顕著である **(3)** 長文要約データセットの MultiNews [12] においては hallucination が減少する．テキスト抽出によって入力長が短くなった結果，モデルのコンテキスト長に収まったことが一因であると分析に基づいて示唆された **(4)** 指示学習を行っていない要約特化モデルの Pegasus [10] では hallucination の増加傾向が顕著である **(5)** 周辺文脈をモデルに入力することでテキスト抽出による hallucination を抑制できる．

2 実験

2.1 実験設定

実験の概要 本実験では抽出されたテキストを合成して要約を生成する二段階要約について，次の問いを検証する: **抽出をしない普通の要約より hallucination は増加するか？** 本実験では既存研究 [7] にならって，要約に含むべき最適なテキストが抽出された前提でテキスト合成を行う．具体的には，入力文書を文単位に分割して，人手で書かれた要約に対する ROUGE スコア [13] が最も高い文の集合を貪欲法で選択する．そして，抽出された各文を合成するように指示が書かれたプロンプトを LLM が受け取って要約を行う．抽出をしない場合は，入力文書の全文を受け取って要約を行う．プロンプト作成に用いたテンプレートは付録の図 3, 4 に示す．

表 2 SummZoo [14] の各データセット．トークン数の列には，入力文書と要約の平均トークン数を掲載．

データセット	要約タスク/ドメイン	トークン数
MultiNews [12]	複数文書/ニュース	2,103/264
XSum [23]	単一文書/ニュース	431/20
ArXiv [24]	単一文書/学術論文	4,938/220
WikiHow [25]	単一文書/ハウツー記事	580/62
Reddit-TIFU [26]	単一文書/ネット掲示板	433/23
SAMSum [27]	対話/雑談	94/28
DialogSum [11]	対話/日常会話	131/24
QMSum [28]	対話・クエリ指向/会議	1,310/65

データセット 要約対象の文書として，SummZoo [14] に含まれる各データセットのテストセットから 100 個のサンプルを無作為に抽出して実験に使用した．各データセットの種類を表 2 に示す．

大規模言語モデル テキスト合成を行うモデルとして，指示への追従能力を測る MT-Bench [3] のスコアが高い以下の指示学習済みモデルを用いた: Llama-2-chat [4], Vicuna [15], Wizardlm [16], Mistral-Instruct [17], TÜLU-2 [18], Mosaic Pretrained Transformer (MPT) [19], Falcon-Instruct [20], Dolly [21], GPT-4 [22]．上記のモデルに加えて，合成モデルが指示への追従能力を持たない場合の挙動を調べるために，要約タスクに特化した言語モデルである Pegasus [10] を用意した．SummZoo の一部データセットの訓練データで追加学習されたモデル [4] を実験に用いた．Pegasus を用いて要約を生成させる際は明示的な指示を与えずに，入力文書もしくは抽出されたテキストの集合を改行してモデルに入力した．

事実整合性の指標 要約に含まれる hallucination の多寡を推定するために，入力文書と要約の事実整合性を評価する最先端の自動指標を 3 個用意した．各指標は 0 から 1 の範囲で正規化されており，要約に書かれている内容が入力文書に包含される度合い—hallucination の少なさ—を示す．(1) **AlignScore**$_{\text{large}}$ [29] は含意関係認識，質問応答，事実検証などの多様なデータセットで RoBERTa$_{\text{large}}$ [5] を追加学習したモデルで評価される指標である．(2) **TrueTeacher** [30] は LLM によって生成・評価された

2) https://chat.openai.com

3) https://arena.lmsys.org

4) https://huggingface.co/docs/transformers/model_doc/pegasus

5) https://huggingface.co/roberta-large

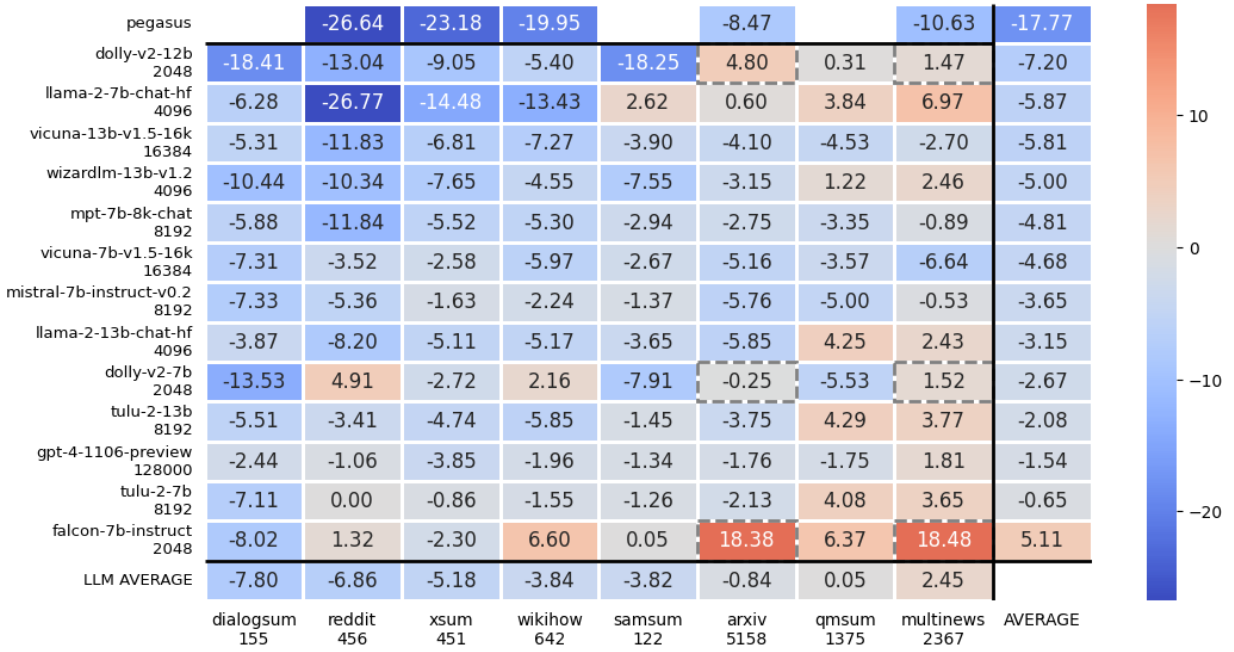

図 1　事実整合性スコアの増減表: 二段階要約のスコアから抽出無し要約のスコアを差し引いた値．負の値はテキスト抽出によって hallucination が増加したことを意味する．各行が合成に使用したモデル，各列がデータセットを表す．モデルには事前学習時の入出力トークン数 (コンテキスト長)，データセットには入力文書と要約を合計したトークン数の平均を併記した．データセットのトークン数がモデルのコンテキスト長を超過するペアを破線で示す．最後の行と列は，Pegasus を除く全 LLM と全データの平均スコアを表す．

事実整合性判定用の合成データで FLAN-T5 XXL [6] を追加学習したモデルで評価される指標である．(3) **G-Eval** [31] は特定の評価用プロンプトを用いて，LLM で評価される指標である．評価用モデルとして OpenAI 社 の gpt-3.5-turbo-1106 [7] を用いた．全ての指標において，抽出の有無によるスコアの増減傾向が大きく変わらなかったため，本論文ではこれらのスコアの平均を提示した．各指標の結果は付録の図 5, 6, 7 に示す．

2.2　抽出による事実整合性スコアの増減

全体的な傾向　図 1 に実験結果を示す．各セルの値は抽出済みテキストの合成による要約（二段階要約）のスコアから，抽出を行わない要約のスコアを差し引いたものである．負の値はテキスト抽出による hallucination の増加を意味する．平均的には二段階要約が hallucination を増加させることが分かる．全モデルとデータセットの組みあわせの約 76%でスコアが減少している．また，全データセットのスコアの平均を取ると，falcon-7b-instruct 以外のモデルでスコアが減少していることが分かる．

DialogSum データでの傾向　対話文書を含む DialogSum データセットでは，−7.8% と平均的には最もスコアが減少した．特に dolly-v2-12b では，DialogSum に加えて同じく対話文書を含む SAMsum データセットでもこの傾向が顕著である．この現象は，対話の抽出によって発話内容の把握が難しくなったことが一因だと考えられる．まず対話要約を正しく行うためには，発話と発話者を結びつける必要がある．対話文書は発話者を示すマーカーと発話内容で構成されるが，文の抽出によって発話内容のみが抽出される可能性がある．加えて，対話文書は口語表現や直接話法で書かれているため，内容の理解のためには非対話文書より長い文脈が必要になると考えられる．

長文要約タスクでの傾向　長文のニュース文書を含む MultiNews データセットでは, 2.45% と平均的に最もスコアが増加した．特に falcon-7b-instruct では，MultiNews に加えて長文の学術論文を含む ArXiv データセットでもこの傾向が顕著である．この傾向は，テキスト抽出によって推論時の入出力トークン数が，事前学習時の訓練データのトークン数（以降，コンテキスト長と呼ぶ）に収まることが一因だと考えられる．データセットの入出力トークン数がモデルのコンテキスト長を超過するペア（図 1 にお

6)　https://huggingface.co/google/flan-t5-xxl
7)　https://platform.openai.com/docs/models/gpt-3-5

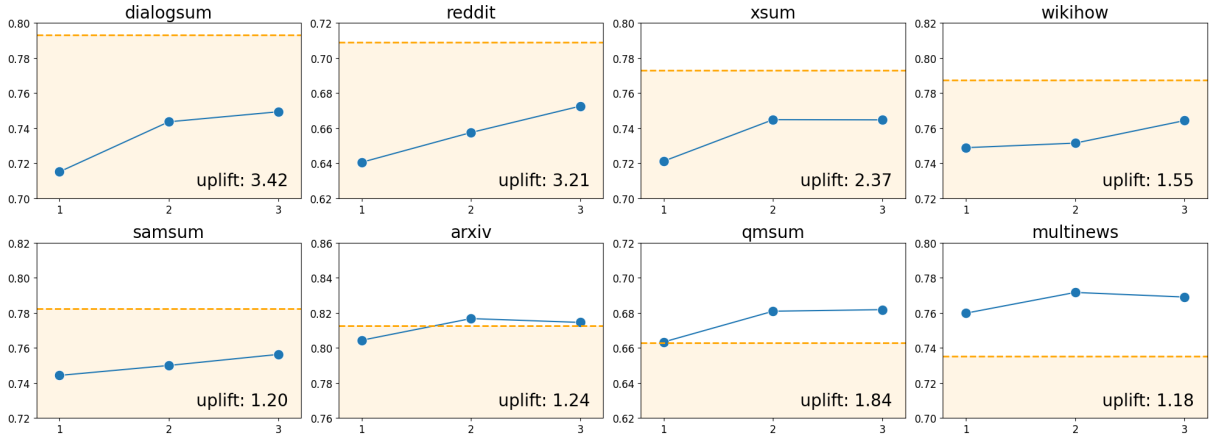

図 2　各データセットにおいて，抽出されたチャンクに含まれる文の数 (横軸) と事実整合性スコア (縦軸) の関係を示した図．Pegasus を除く全モデルの平均スコアを折れ線で示した．破線は抽出無し要約の平均スコアを意味する．各図の右下に，文数の増加に伴うスコアの最大増加量を明記した．

いて破線で表示）のほとんどで，スコアが増加している．上記のペアにおいて falcon-7b-instruct のほうが dolly-v2-7b/12b よりスコア増加が著しい傾向は，指示学習で用いたデータに起因すると考えられる．dolly は文書要約データを含む databricks-dolly-15k [8] で指示学習されているため，抽出無し要約でもスコアが相対的に高く，テキスト抽出による恩恵をそれほど強く受けなかったと考えられる．また，入出力トークン数が多い Arxiv/MultiNews/QMsum では，モデルのコンテキスト長が短くなるにつれて，二段階要約から抽出無し要約のスコアを差し引いた差分が大きくなる傾向がある．GPT-4 と Pegasus を省いたモデル群において，コンテキスト長とスコア差分の相関係数はこれらのデータセットにおいて –0.4 以下である．これらの結果から，推論時の入出力トークン数がモデルのコンテキスト長に収まることで hallucination が抑制されることが分かる．

指示学習の有無による影響　全モデルのうち，指示学習が行われていない Pegasus のスコアの減少が最も顕著だった．LLM において平均的にスコアが増加している MultiNews データセットであっても Pegasus ではスコアが大きく減少している．この結果は，指示学習は二段階要約による hallucination を抑制することを示している．

2.3　抽出テキストのサイズが与える影響

二段階要約によって発生する hallucination は，テキスト合成を行う言語モデルが抽出されたテキストの文脈を把握できないことに起因すると考えられる．そこで，抽出される文の周辺にあるテキストをモデルに入力したときの hallucination の増減を確認する．具体的には，入力文書を複数の文で構成されるテキスト（以降，チャンクと呼ぶ）に分割して，そのチャンクを合成して得られる要約の事実整合性スコアを報告する．

図 2 にチャンクに含まれる文の数 (横軸) と事実整合性スコア (縦軸) の関係を示す．破線は抽出無し要約のスコアを示す．まず，全データセットにおいて，文数の増加に伴ってスコアが向上する傾向が見てとれる．全データセットで平均すると，2% スコアが増加した．この結果から，テキスト抽出によって失った文脈を与えることで，hallucination を抑制できることが分かる．ただし，テキスト抽出によるスコアの減少幅が小さい場合，この抑制の効果も小さい．例えば，減少幅が小さい ArXiv/QMsum/MultiNews において，文数の増加に伴うスコアの最大増加量は平均 1.86% であるのに対して，それ以外では 2.35%である．これらのデータセットは相対的に入力トークン数が多いため，テキスト抽出によって推論時の内挿が可能になった好影響が，文脈を失う悪影響を打ち消したと考えられる．

3　まとめ

本論文では，二段階要約において大規模言語モデルでテキスト合成を行う際に発生する hallucination 現象の分析を行った．実験結果から，対話文書・長文文書における hallucination の増減の傾向や，指示学習・周辺文脈の追加による抑制など，既存研究では不明であった新しい知見が得られた．

8）https://huggingface.co/datasets/databricks/databricks-dolly-15k

参考文献

[1] Tianyi Zhang, et al. Benchmarking large language models for news summarization. **CoRR**, Vol. abs/2301.13848, , 2023.

[2] Xiao Pu, et al. Summarization is (almost) dead. **CoRR**, Vol. abs/2309.09558, , 2023.

[3] Ziwei Ji, et al. Survey of hallucination in natural language generation. **ACM Comput. Surv.**, Vol. 55, No. 12, pp. 248:1–248:38, 2023.

[4] Hugo Touvron, et al. Llama 2: Open foundation and fine-tuned chat models. **CoRR**, Vol. abs/2307.09288, , 2023.

[5] Huan Yee Koh, et al. How far are we from robust long abstractive summarization? In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022**, pp. 2682–2698. Association for Computational Linguistics, 2022.

[6] Haopeng Zhang, et al. Extractive summarization via chatgpt for faithful summary generation. In **Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023**, pp. 3270–3278. Association for Computational Linguistics, 2023.

[7] Haoran Li, et al. EASE: Extractive-abstractive summarization end-to-end using the information bottleneck principle. In **Proceedings of the Third Workshop on New Frontiers in Summarization**, pp. 85–95, Online and in Dominican Republic, November 2021. Association for Computational Linguistics.

[8] Aviv Slobodkin, et al. Summhelper: Collaborative human-computer summarization. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023**, pp. 554–565. Association for Computational Linguistics, 2023.

[9] Shiyue Zhang, et al. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023**, pp. 2153–2174. Association for Computational Linguistics, 2023.

[10] Jingqing Zhang, et al. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In **Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event**, Vol. 119 of **Proceedings of Machine Learning Research**, pp. 11328–11339. PMLR, 2020.

[11] Yulong Chen, et al. DialogSum: A real-life scenario dialogue summarization dataset. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 5062–5074, Online, August 2021. Association for Computational Linguistics.

[12] Alexander R. Fabbri, et al. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In **Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers**, pp. 1074–1084. Association for Computational Linguistics, 2019.

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

[14] Yulong Chen, et al. Unisumm and summzoo: Unified model and diverse benchmark for few-shot summarization. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023**, pp. 12833–12855. Association for Computational Linguistics, 2023.

[15] Lianmin Zheng, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. **CoRR**, Vol. abs/2306.05685, , 2023.

[16] Can Xu, et al. Wizardlm: Empowering large language models to follow complex instructions. **CoRR**, Vol. abs/2304.12244, , 2023.

[17] Albert Q. Jiang, et al. Mistral 7b. **CoRR**, Vol. abs/2310.06825, , 2023.

[18] Hamish Ivison, et al. Camels in a changing climate: Enhancing LM adaptation with tulu 2. **CoRR**, Vol. abs/2311.10702, , 2023.

[19] MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. Accessed: 2023-05-05.

[20] Ebtesam Almazrouei, et al. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

[21] Mike Conover, et al. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023.

[22] OpenAI. GPT-4 technical report. **CoRR**, Vol. abs/2303.08774, , 2023.

[23] Shashi Narayan, et al. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[24] Arman Cohan, et al. A discourse-aware attention model for abstractive summarization of long documents. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)**, pp. 615–621, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[25] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. **CoRR**, Vol. abs/1810.09305, , 2018.

[26] Byeongchang Kim, et al. Abstractive summarization of Reddit posts with multi-level memory networks. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**, pp. 2519–2531, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[27] Bogdan Gliwa, et al. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In **Proceedings of the 2nd Workshop on New Frontiers in Summarization**, pp. 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics.

[28] Ming Zhong, et al. QMSum: A new benchmark for query-based multi-domain meeting summarization. In **Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 5905–5921, Online, June 2021. Association for Computational Linguistics.

[29] Yuheng Zha, et al. Alignscore: Evaluating factual consistency with A unified alignment function. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023**, pp. 11328–11348. Association for Computational Linguistics, 2023.

[30] Zorik Gekhman, et al. Trueteacher: Learning factual consistency evaluation with large language models. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023**, pp. 2053–2070. Association for Computational Linguistics, 2023.

[31] Yang Liu, et al. G-eval: NLG evaluation using gpt-4 with better human alignment. In **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023**, pp. 2511–2522. Association for Computational Linguistics, 2023.

A 付録

```
Composite the extracted texts and summarize them in {n_words} words

text_0 {text_0}
...
text_n {text_n}
```

図 3 二段階要約おいてプロンプトを作成するためのテンプレート．“n_words” と “text_*” には人手で書かれた要約の単語数と抽出されたテキストがそれぞれ代入される．

```
Summarize the document in {n_words} words

{document}
```

図 4 抽出無し要約においてプロンプトを作成するためのテンプレート．“n_words” と “document” には人手で書かれた要約の単語数と入力文書の全文がそれぞれ代入される．

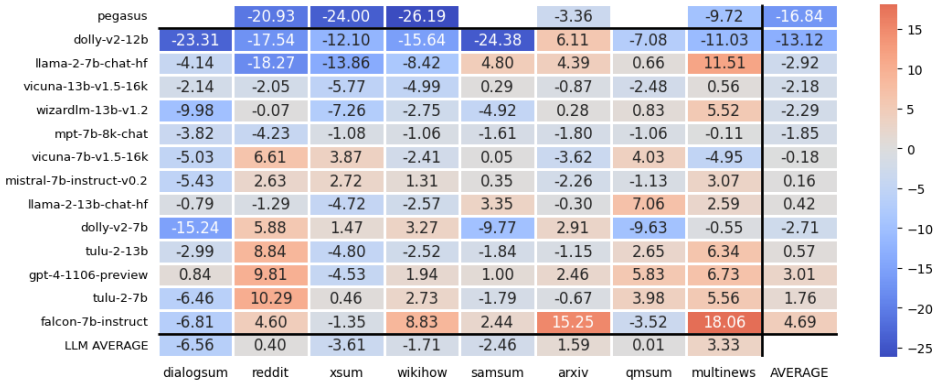

図 5 $\text{AlignScore}_{\text{large}}$ [29] における事実整合性スコアの増減表．図の読み方は図 1 を参照されたい．

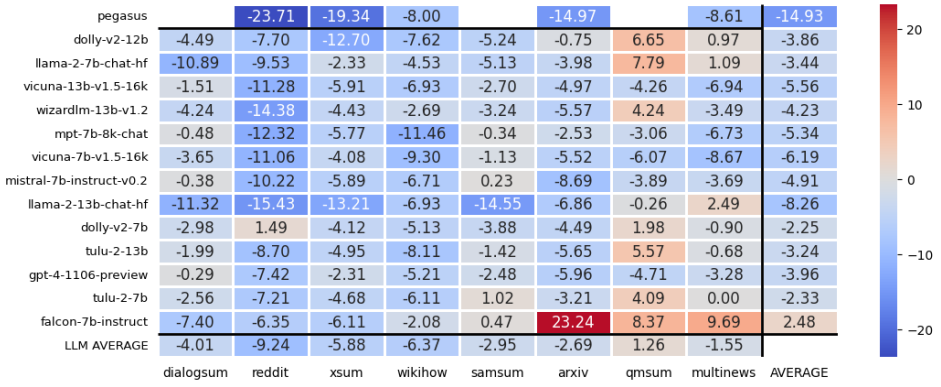

図 6 G-Eval [31] における事実整合性スコアの増減表．図の読み方は図 1 を参照されたい．

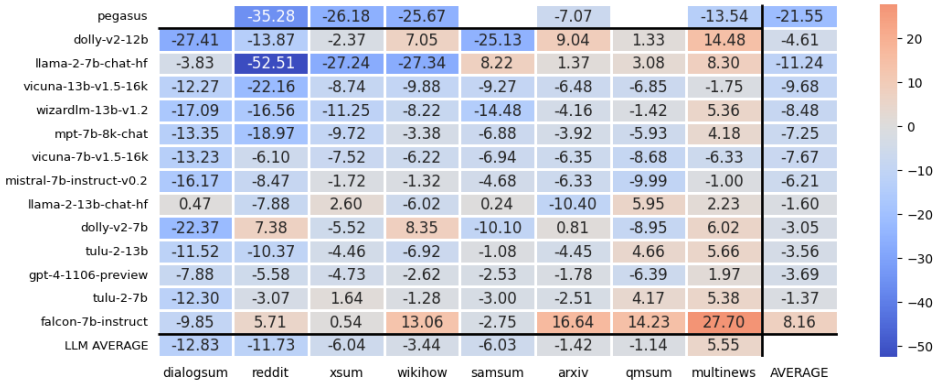

図 7 TrueTeacher [30] における事実整合性スコアの増減表．図の読み方は図 1 を参照されたい．