# Comparison-Based Automatic Evaluation for Meeting Summarization

*Ziwei Gong*[1,2], *Lin Ai*[1,2], *Harshsaiprasad Deshpande*[1], *Alexander Johnson*[1], *Emmy Phung*[1], *Zehui Wu*[2], *Ahmad Emami*[1], *Julia Hirschberg*[2]

[1]Machine Learning Center of Excellence, JPMorgan Chase & Co., USA
[2]Department of Computer Science, Columbia University, USA
{sara.ziweigong, lin.ai, julia}@cs.columbia.edu

## Abstract

Large Language Models (LLMs) have spurred interest in automatic evaluation methods for summarization, offering a faster, more cost-effective alternative to human evaluation. However, existing methods often fall short when applied to complex tasks like long-context summarizations and dialogue-based meeting summarizations. In this paper, we introduce CREAM (Comparison-based Reference-free Elo-ranked Automatic evaluation for Meeting summarization), a novel framework that addresses the unique challenges of evaluating meeting summaries. CREAM leverages a combination of chain-of-thought reasoning and key facts alignment to assess conciseness and completeness of model-generated summaries without requiring reference. By employing an ELO ranking system, our approach provides a robust mechanism for comparing the quality of different models or prompt configurations.
**Index Terms**: speech recognition, summarization, evaluation

## 1. Introduction

The rapid advancement of Large Language Models (LLMs) has significantly influenced the field of automatic evaluation for text summarization, including meeting summarization derived from automatic speech recognition (ASR) transcripts, making it faster and more cost-effective compared to traditional human evaluation [1, 2]. However, despite progress in automatic evaluation techniques, existing methods primarily target general-purpose summarization tasks, which typically involve shorter, more structured text inputs. These approaches may not be well-suited for evaluating summaries derived from ASR transcripts, which often contain disfluencies, speaker overlaps, and transcription errors [3]. Accurately summarizing lengthy spoken interactions is essential in NLP applications, particularly in industrial and enterprise settings. Transforming spoken discussions into structured, actionable insights enhances project management, ensures compliance, and supports strategic planning.

Long-context meeting transcripts and their summarization pose challenges that current evaluation metrics fail to address [3]. Unlike text-based documents, ASR-generated transcripts introduce additional complexity due to errors in recognition, a lack of punctuation, and difficulties in speaker attribution. Additionally, issues like the "middle curse" [4, 5]—where models neglect middle sections of long transcripts—further highlight the limitations of general-purpose summarization metrics. This raises critical questions about the effectiveness of existing evaluation methods for meeting summarization, a hypothesis we thoroughly examine in this work. Our findings reveal that current LLM-based evaluators struggle to assess meeting summaries reliably, underscoring the need for a more specialized approach tailored to ASR-derived long content.

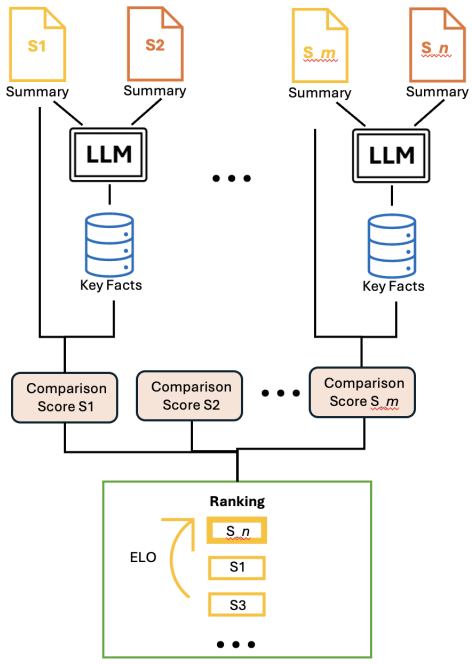


Figure 1: *Illustrations of the CREAM framework. CREAM distills candidate summary pairs into key facts, conducts pair wise comparison of summary pairs, and then uses an Elo rating system to rank summaries.*

In this paper, we address this gap by developing a new evaluation framework tailored specifically for meeting summarization.We propose CREAM (Comparison-based Reference-free Elo-ranked Automatic evaluation for Meeting summarization), a novel system designed to fill the gaps in specialized and customizable evaluation for meeting summaries as illustrated in Figure 1. Our research addresses the following key questions:

1. Do current LLM-based automatic evaluators work effectively for meeting summarization? (We find that they do not.)
2. How can we design an efficient, reference-free, automatic evaluator for meeting summarization? (We propose comparison-based CREAM framework using ELO ranking.)

Our results highlight the limitations of current LLM-based evaluators and demonstrate the effectiveness of our specialized framework, CREAM, with its novel comparison-based Elo ranking method for summarization evaluation. We benchmark various GPT models against our framework and find that GPT-4o excels in completeness, GPT-4 in conciseness, but all struggle to find a balance between completeness and conciseness.

## 2. Related Work

**Reference-based Evaluation Metrics.** Meeting summarization typically relies on reference-based metrics to assess quality. Ngram-based metrics like ROUGE [6] and BLEU [7] evaluate n-gram overlap and precision, respectively, while embedding-based metrics like BERTScore [8] and BARTScore [9] assess semantic similarity and generation likelihood. These similarity-based metrics often overlook key aspects like factuality, conciseness, and completeness, underscoring the need for more advanced evaluation frameworks for meeting summarization.

**LLM-based Evaluation and Metrics.** LLM-based evaluation methods like G-Eval [1] and FineSurE [10] show improved alignment with human judgments. G-Eval leverages GPT-4 with Chain-of-Thought reasoning, while FineSurE offers fine-grained evaluations of faithfulness, completeness, and conciseness. However, their applicability to long-context meeting summarization remains underexplored, highlighting the need for further adaptation.

**Elo Rating in Model Evaluation.** The Elo rating system [11], widely used for ranking competitors in games like chess, has recently been adapted for evaluating natural language generation systems. [12] applied Elo in RAGElo to assess Retrieval-Augmented Generation (RAG) systems, aligning well with expert annotations, while [13] emphasized careful calibration for LLM rankings. [14] introduced Auto-Arena, an Elo-based system showing strong correlation with human preferences, and [15] analyzed Elo among other ranking methods. Our work extends these efforts by applying Elo to rank summarization models on conciseness and completeness, offering a reliable framework for meeting summarizations.

## 3. Limitations of Current Methods

In order to answer the question, *"Do current evaluators work effectively for meeting summarization?"*, we experiment with the state-of-the-art frameworks [1, 10] for automatic summary evaluation and identify several issues. Existing LLM-based methods evaluate summaries by comparing them to reference texts, using human-written summaries as gold standards and producing absolute scores under various metrics via a CoT process. This approach often biases towards self-generated content.

Our experiments reveal that current LLM-based evaluators: 1) struggle with accurate completeness and conciseness scoring for long-context documents, often exhibiting self-bias; 2) frequently generate false positives in complex dialogues, suggesting industrial applications should prioritize completeness and conciseness in evaluation frameworks.

### 3.1. Experiment Setup

**Datasets.** We select both public and private datasets which collectively cover a range of summarization-related tasks: factuality assessment, key-fact [1] extraction, and summarization.

We use both FRANK [17], containing 2,246 summaries with sentence-level faithfulness annotation, and REALSumm [16], containing 2,500 summaries with key facts alignment, for key-fact extraction and factuality evaluation.

QMSum [18] is a query-based, long meeting summarization dataset that includes 232 meetings with average number of 556.8 turns across diverse domains including product, academic, and committee meetings.

[1] A key fact refers to a concise sentence conveying a single key information, comprising at most 2-3 entities [16, 10].

Our Internal Zoom Meeting Summarization (IZMS) data consists of 139 examples from internal Zoom meetings, used to assess factuality and key-fact extraction. This dataset includes transcripts from 10 product meetings and an additional 40 meetings, providing real-world data for evaluation.

**Models.** Our experiments use GPT-based models from OpenAI [19], including GPT-4-omni (GPT-4O-2024-05-13), GPT-4-turbo (GPT-4-TURBO-2024-04-09), and GPT-3.5 (GPT-35-TURBO-16K-0613). We clear the history for each evaluation instance, following precedents [10, 20].

**Metrics.** We evaluate LLM reliability on three key metrics using methods consistent with recent research [10]. *Completeness* assesses the extent to which the summarizer includes all key facts from the input text in the output summary. *Conciseness* measures the summarizer's ability to avoid incorporating information outside the key facts in the output, maintaining focus and brevity. *Faithfulness* evaluates whether the summarizer accurately represents the information in the input text.

### 3.2. Experiments Findings

For both **completeness and conciseness**, LLM evaluators show low correlation with human preference and high self-bias in long-context dialogues like meeting summaries, although perform well in evaluating shorter summaries. Current evaluators struggle with accurately identifying **factual errors** in meeting summaries with high false positives.

#### *3.2.1. Completeness*

LLMs evaluators like GPT-4o can accurately extract key facts and compare summaries with high correlation to human evaluations in shorter summarization tasks like those in REALSum, which involve news-style articles with summaries averaging $3-5$ sentences. In our experiments using REALSum, GPT-4o showed a high correlation with human scores when using human-annotated key facts as references (rank correlation of $0.95$, Pearson's r), and also similarly positive correlation when comparing with the extracted key facts from entire articles.

However, these evaluators struggle with accurately evaluating completeness in long-context dialogues, such as meeting summaries from the QMSum dataset (average of $556.8$ turns). As in Table 1, our evaluation revealed that all versions of GPT have a weak correlation (Pearson's r $= 0.5$, ) with gold standard human scores with a strong self-bias. Both GPT-4 and GPT-3.5 tend to give near-perfect scores for summaries generated by LLMs, with GPT-4o performing only slightly better but still showing a bias towards its own generated content.

#### *3.2.2. Conciseness*

Similar to completeness, the evaluators have a weak correlation with gold standard human scores with a strong self-bias when evaluating conciseness. As in Table 1, LLMs are less effective for long-context dialogues, and performance without gold reference tends to result in an overestimation of conciseness.

#### *3.2.3. Factuality*

We evaluated whether current LLM-based evaluators (GPT-4o) are effective on factuality. We reproduce results on the FRANK dataset, which consisted of sentence-level factuality errors. We achieve a balanced accuracy of $61.5\%$ at the sentence level. Summary-level correlations with human judgments were moderate (Pearson$= 0.38$ and Spearman$= 0.35$). At system level, therank correlation is perfect (Spearman'r$= 1.0$).

| | Max# keyfacts | Summarizer Model | Human Summary (GPT4o) | Machine summary (GPT4o) | Transcript (GPT4o) | Transcript (GPT4) | Transcript (GPT3.5) |
|---|---|---|---|---|---|---|---|
| COMPLETENESS | 16 | GPT3.5 | 56.2% | 100.0% | 89.7% | 99.3% | **99.9%** |
| | | GPT4 | 57.9% | 100.0% | 88.5% | 99.0% | 99.8% |
| | | GPT4o | **77.9%** | 100.0% | **98.9%** | **99.9%** | 99.8% |
| | 30 | GPT3.5 | 54.8% | 100.0% | 82.5% | 99.4% | **99.9%** |
| | | GPT4 | 55.6% | 100.0% | 81.2% | 99.4% | 98.9% |
| | | GPT4o | **76.9%** | 100.0% | **98.1%** | **99.9%** | 99.7% |
| CONCISENESS | 16 | GPT3.5 | 58.2% | 93.3% | 91.9% | **99.0%** | 96.7% |
| | | GPT4 | **69.3%** | **95.0%** | **96.0%** | **99.0%** | **97.8%** |
| | | GPT4o | 48.8% | 62.1% | 70.5% | 91.2% | 78.9% |
| | 30 | GPT3.5 | 56.3% | 99.4% | 97.6% | **99.4%** | 97.1% |
| | | GPT4 | **68.1%** | **99.8%** | **98.0%** | 99.2% | **97.5%** |
| | | GPT4o | 49.8% | 92.5% | 92.5% | 96.3% | 79.5% |

Table 1: *Evaluation of current evaluators on completeness and conciseness for meeting summarization. All models show weak correlation with human scores and high self-bias, often assigning near-perfect scores to their own summaries. Summaries generated by various models are evaluated using different evaluator models (in brackets) with varying reference setups. The "Max # Key Facts" column sets the max number of key facts extracted from reference text. "Human Summary" column is the gold standard baseline using key facts from human summaries.*

Yet, when the same method is applied to our IZMS data, the results are less promising. The summary-level correlations were low (Pearson$= 0.11$ and Spearman$= 0.14$), indicating that the model's ability to find factual errors across different contexts is limited. There are no significant differences in factuality between short and long summaries, with error ratios of $16.3\%$ and $15.2\%$, respectively. The system-level ranking for different summary types also did not show significant differences.

In conclusion, our research shows that while LLM-based evaluators perform well in designed factuality datasets, they often generate false positives and struggle with complex, dialogue-based summaries, making them less reliable for automatic factuality evaluation in meeting summarization tasks. Analysis of the IZMS dataset shows that factuality errors are rare in real-world scenarios, typically arising from ASR errors or GPT-generated out-of-context information, contrasting with their higher frequency in controlled datasets.

**Since factuality errors are rare in real-world data, unlike their prevalence in designed datasets**, we prioritize on completeness and conciseness in our evaluation framework.

# 4. CREAM Framework

To address the limitations of existing long-context summarization evaluation methods, we propose the CREAM framework: Comparison-Based Reference-Free Elo-Ranked Automatic Evaluation for Meeting Summarization. This framework leverages Chain-of-Thought (CoT) reasoning and comparison-based methods to enhance the evaluation process.

Traditional completeness and conciseness metrics struggle due to challenges in key fact extraction, especially in multi-speaker dialogues. CREAM mitigates this by comparing summaries directly rather than relying on reference transcripts. As shown in Figure 1, CREAM evaluates summaries in two steps. First, a CoT-based prompt extracts concise, non-redundant key facts from the concatenated summaries (Section 4.1.2). Then, the model checks how well each summary preserves these facts, identifying supporting sentences (Section 4.1.2).

The results are then parsed for calculation of the summaries' completeness and conciseness, which results in multiple comparison-based scores between models. Detailed results are in Section 4.2.1. Scores from these comparisons inform an Elo ranking system, which assigns ratings based on pairwise evaluations (Section 4.1.1), allowing robust, reference-free model ranking. The final results are in Section 4.2.3.

| **Max Number of Key Facts** | **16** | **20** | **30** | **40** | **50** | **60** | **70** | **100** |
|---|---|---|---|---|---|---|---|---|
| GPT4o | 15.8 | 19.7 | 27.0 | 26.6 | 19.6 | 21.3 | 23.4 | 22.4 |
| GPT4 | 12.8 | 12.9 | 13.0 | 12.9 | 12.3 | 12.8 | 13.3 | 12.0 |
| GPT3.5 | 12.4 | 12.3 | 13.2 | 12.5 | 12.7 | 12.9 | 12.6 | 13.3 |

Table 2: *Average number of key facts extracted from transcripts by different models from one meeting.*

## 4.1. Implementation

### *4.1.1. Comparison-based Ranking Metrics*

CREAM employs a reference-free, comparison-based evaluation to differentiate model performance, particularly in long-dialogue summarization. This method enables fine-grained comparisons across different summary generation prompts.

We rank models using the Elo rating system, where scores update dynamically based on comparison outcomes:

$$R_{new} = R_{old} + K \times (S - E) \tag{1}$$

where $R_{new}$ is the new Elo rating after the match; $R_{old}$ is the old Elo rating before the match. $K$ is the K-factor, which determines the maximum possible adjustment per match. $S$ is the actual score (1 for a win, 0.5 for a draw, 0 for a loss). $E$ is the expected score, calculated as:

$$E = \frac{1}{1 + 10^{(R_{opponent} - R_{old})/400}} \tag{2}$$

where $R_{opponent}$ is the Elo rating of the opponent.

Using Elo rankings instead of raw scores allows robust, reference-free evaluation, mitigating biases from missing content or imperfect gold annotations.

### *4.1.2. Prompting Instruction*

We employ CoT zero-shot prompting methods to calculate both conciseness and completeness. The prompt instructs the model to perform the following tasks:

1. Key Fact Extraction: reads the provided paragraph and breaks it down into a list of key facts.

2. Key Fact Alignment: then, compares each fact with the sentences in the summary. For each key fact, the model determines whether it is supported by the summary and, if so, identifies the relevant summary sentences by their line numbers.

| | | Completeness | | | Conciseness | | |
|---|---|---|---|---|---|---|---|
| kf | VS | GPT3.5 | GPT4 | GPT4o | GPT3.5 | GPT4 | GPT4o |
| 16 | GPT3.5 | - | **78.8%** | 84.2% | - | **90.0%** | **87.5%** |
| | GPT4 | 76.6% | **-** | 86.6% | **93.6%** | **-** | **92.1%** |
| | GPT4o | **92.1%** | **94.0%** | - | **67.3%** | **68.4%** | **-** |
| 30 | GPT3.5 | - | **80.7%** | 82.2% | - | 99.1% | **96.4%** |
| | GPT4 | 80.0% | - | 86.3% | **99.4%** | **-** | **95.9%** |
| | GPT4o | **94.9%** | **95.5%** | - | 89.7% | 88.6% | - |

Table 3: *Raw pair-wise score for completeness and conciseness on QMSum, kf denotes max number of key facts set. Pair-wise winner in **bold**.*

| | | Completeness | | Conciseness | |
|---|---|---|---|---|---|
| kf | VS | short | long | short | long |
| 16 | short | - | 95.3 | - | 75.6 |
| | long | 98.4 | **-** | 53.5 | - |
| 30 | short | - | 97.7 | - | 92.7 |
| | long | 97.3 | **-** | 80.7 | - |

Table 4: *Raw pair-wise score for completeness and conciseness on IZMS.*

The design of the prompt is enhanced by several key techniques, as outlined below:

**Chain-of-Thought Reasoning** We incorporate a reasoning step to guide the model through the task, ensuring that the model evaluates the key facts systematically and accurately.

**One-Shot Learning for Key Fact Extraction** We experiment with using out-of-domain versus in-domain examples for key fact extraction. Interestingly, out-of-domain examples outperform in-domain ones, reducing confusion in fact recognition.

**Maximum Number of Key Facts** We test the model's performance using different limits on the number of key facts. As shown in Table 2, GPT-4o is highly sensitive to the number of key facts. Based on experiments with the IZMS datasets, we decided to use 16 key facts (following precedent literature) and 30 key facts (the maximum effective number for GPT-4o), with GPT-4o as evaluator model, in all subsequent experiments.

### 4.2. Experiment Results

We use the same setups as in Section 3.1, focusing on evaluting completeness and conciseness.

#### *4.2.1. Pair-wise Raw Scores*

These raw scores of model comparisons show significant differences that are often difficult to discern when using absolute values to compare LLM-generated summaries of long-context dialogues. Table 3 and Table 4 present the raw results of the model comparison scores for completeness and conciseness, on two meeting summarization dataset QMSum and IZMS respectively. For instance, row GPT-4o and column GPT-3.5 means that when GPT-4o and GPT-3.5 are compared pairwise, GPT-3.5 is getting a completeness score of 92.1%, and GPT-4o is getting a completeness score of 84.2% (row GPT-3.5 and column GPT-4o).

#### *4.2.2. Interpretable Key Facts*

The intermediate results, key facts lists, provide **interpretable** insights on the qualitative differences between two summaries. These lists show which key facts are captured or missed, offering insights into how models condense content. For example, in meeting summaries, Summary A captures casual chit-chat but misses key decisions, while Summary B skips chit-chat to focus on critical outcomes. Comparison-based scores reveal that while Summary A is more complete, Summary B's conciseness makes it more useful for quickly grasping meeting outcomes.

| Summary Model | GPT 4o | GPT 4 | GPT 3.5 |
|---|---|---|---|
| Completeness (Gold) | **1st**$_{(77.9\%)}$ | 2nd$_{(57.9\%)}$ | 3rd$_{(54.8\%)}$ |
| Completeness (CREAM) | **1st** | 2nd | 3rd |
| Conciseness (Gold) | 3rd$_{(48.8\%)}$ | **1st**$_{(69.3\%)}$ | 2nd$_{(58.2\%)}$ |
| Conciseness (CREAM) | 3rd | **1st** | 2nd |

Table 5: *CREAM Results. Completeness and conciseness ranking from gold human ranking (human scores in brackets) with reference to gold summary, and ranking from CREAM framework (when set to 16 key fact).*

#### *4.2.3. Elo-ranked Results*

As shown in 5 Table CREAM outperforms prior baselines in ranking correlation, especially on meeting summarization data, improving ranking correlation from 0.5 to 1.0 (Pearson's r) in both completeness and conciseness. We demonstrate the effectiveness of CREAM in evaluating across different models and identifying optimal prompts for meeting summarization tasks.

We applied the described Elo rating system (Section 4) to raw scores from system comparisons above. The results align closely with the human evaluation results based on gold summaries, as shown in Table 5. We see a perfect correlation in ranking of model preference (rank correlation of 1.0, Pearson's r), even when the gold human-preference scores are close such as the 2nd and 3rd place for completeness score. This consistency suggests that the Elo rating system is an effective method for ranking models in meeting summarization tasks, offering a reliable way to compare their relative performance.

#### *4.2.4. The Trade-off Between Metrics*

A significant trade-off exists between completeness and conciseness. This trade-off is inherently subjective, depending on the specific needs or preferences of the user. Our approach to key fact extraction can address this trade-off more effectively. If added additional selection steps during key-fact comparison, users can customize their summary preferences to focus on the most relevant information while refining verbose outputs into concise yet complete versions.

## 5. Conclusions and Limitations

In conclusion, we first examine the effectiveness of LLM-based evaluators for meeting summarization, revealing that general-purpose LLM-based evaluators do not adapt well to the unique challenges posed by long-context dialogue summarization. To address these challenges, we introduced a novel automatic evaluation framework, CREAM, providing a more accurate and adaptable evaluation process using Elo-based comparison algorithms (improving rank correlation of 0.5, Pearson's r to 1.0). This framework enables faster and cheaper analysis of models, which can significantly speed up the progress of the field by allowing new methods to be benchmarked more efficiently. However, CREAM is designed for meeting summarization and may need adaptation for other tasks. Its effectiveness also depends on the LLMs, which can introduce biases or inconsistencies.

The boarder impact of this work goes beyond evaluation techniques, addressing the need for robust frameworks like CREAM as models surpass human evaluators. By democratizing evaluation, fostering innovation, and enabling integration into reinforcement learning pipelines, CREAM offers scalable, reliable solutions for summarization quality. Our scalable evaluation framework could integrate into reinforcement learning pipelines, providing automatic rewards based on summarization quality—a crucial feature given RL's growing popularity.

## 6. Acknowledgements

The authors would like to thank Yanda Chen for the valuable feedback and insight to this paper. This research is supported in part by the National Science Foundation via NSF DBI-2229929 (The NSF AI Institute for Artificial and Natural Intelligence).

## 7. References

[1] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: NLG evaluation using gpt-4 with better human alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2511–2522. [Online]. Available: https://aclanthology.org/2023.emnlp-main.153

[2] J. Wang, Y. Liang, F. Meng, Z. Sun, H. Shi, Z. Li, J. Xu, J. Qu, and J. Zhou, "Is ChatGPT a good NLG evaluator? a preliminary study," in *Proceedings of the 4th New Frontiers in Summarization Workshop*, Y. Dong, W. Xiao, L. Wang, F. Liu, and G. Carenini, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1–11. [Online]. Available: https://aclanthology.org/2023.newsum-1.1

[3] Y. Chang, K. Lo, T. Goyal, and M. Iyyer, "Booookscore: A systematic exploration of book-length summarization in the era of LLMs," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://arxiv.org/pdf/2310.00785.pdf

[4] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024. [Online]. Available: https://aclanthology.org/2024.tacl-1.9

[5] M. Ravaut, A. Sun, N. Chen, and S. Joty, "On context utilization in summarization with large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 2764–2781. [Online]. Available: https://aclanthology.org/2024.acl-long.153

[6] 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. [Online]. Available: https://aclanthology.org/W04-1013

[7] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135

[8] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[9] W. Yuan, G. Neubig, and P. Liu, "Bartscore: Evaluating generated text as text generation," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 27 263–27 277. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf

[10] H. Song, H. Su, I. Shalyminov, J. Cai, and S. Mansour, "Finesure: Fine-grained summarization evaluation using llms," in *ACL*, 2024.

[11] A. E. Elo, "The proposed uscf rating system, its development, theory, and applications," in *Chess Life*, vol. 22(8), 1967, p. 242–247.

[12] Z. Rackauckas, A. Câmara, and J. Zavrel, "Evaluating rag-fusion with ragelo: an automated elo-based framework," 2024. [Online]. Available: https://arxiv.org/abs/2406.14783

[13] M. Boubdir, E. Kim, B. Ermis, S. Hooker, and M. Fadaee, "Elo uncovered: Robustness and best practices in language model evaluation," in *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, S. Gehrmann, A. Wang, J. Sedoc, E. Clark, K. Dhole, K. R. Chandu, E. Santus, and H. Sedghamiz, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 339–352. [Online]. Available: https://aclanthology.org/2023.gem-1.28

[14] R. Zhao, W. Zhang, Y. K. Chia, D. Zhao, and L. Bing, "Auto arena of llms: Automating llm evaluations with agent peer-battles and committee discussions," 2024. [Online]. Available: https://arxiv.org/abs/2405.20267

[15] Y. Zhang, M. Zhang, H. Yuan, S. Liu, Y. Shi, T. Gui, Q. Zhang, and X. Huang, "Llmeval: A preliminary study on how to evaluate large language models," 2023. [Online]. Available: https://arxiv.org/abs/2312.07398

[16] M. Bhandari, P. N. Gour, A. Ashfaq, P. Liu, and G. Neubig, "Re-evaluating evaluation in text summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 9347–9359. [Online]. Available: https://aclanthology.org/2020.emnlp-main.751

[17] A. Pagnoni, V. Balachandran, and Y. Tsvetkov, "Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 4812–4829. [Online]. Available: https://aclanthology.org/2021.naacl-main.383

[18] M. Zhong, D. Yin, T. Yu, A. Zaidi, M. Mutuma, R. Jha, A. H. Awadallah, A. Celikyilmaz, Y. Liu, X. Qiu, and D. Radev, "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*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Online: Association for Computational Linguistics, Jun. 2021, pp. 5905–5921. [Online]. Available: https://aclanthology.org/2021.naacl-main.472

[19] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, and S. J. et al., "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[20] C. Shen, L. Cheng, X.-P. Nguyen, Y. You, and L. Bing, "Large language models are not yet human-level evaluators for abstractive summarization," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 4215–4233. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.278