What's under the hood: Investigating Automatic Metrics on Meeting Summarization

Frederic Kirstein[1,2], Jan Philip Wahle[1], Terry Ruas[1], Bela Gipp[1]
[1]University of Göttingen, Germany
[2]kirstein@gipplab.org

Abstract

Meeting summarization has become a critical task considering the increase in online interactions. Despite new techniques being proposed regularly, the evaluation of meeting summarization techniques relies on metrics not tailored to capture meeting-specific errors, leading to ineffective assessment. This paper explores what established automatic metrics capture and the errors they mask by correlating metric scores with human evaluations across a comprehensive error taxonomy. We start by reviewing the literature on English meeting summarization to identify key challenges, such as speaker dynamics and contextual turn-taking, and error types, including missing information and linguistic inaccuracy, concepts previously loosely defined in the field. We then examine the relationship between these challenges and errors using human-annotated transcripts and summaries from encoder-decoder-based and decoder-only Transformer models on the QMSum dataset. Experiments reveal that different model architectures respond variably to the challenges, resulting in distinct links between challenges and errors. Current established metrics struggle to capture the observable errors, showing weak to moderate correlations, with a third of the correlations indicating error masking. Only a subset of metrics accurately reacts to specific errors, while most correlations show either unresponsiveness or failure to reflect the error's impact on summary quality.

1 Introduction

The rise in remote collaboration has increased the need for effective meeting summarization (Mroz et al., 2018; Pratama et al., 2020), beneficial for participants and non-attendees. However, established evaluation metrics do not fully capture the challenges of meeting transcripts such as contextual turn-taking and discourse structure (Rennard et al., 2023; Kumar and Kabiri, 2022). The commonly used ROUGE metric (Lin, 2004) has limitations in reflecting summary quality (Liu and Liu, 2008; Cohan and Goharian, 2016; Fabbri et al., 2021), and newer metrics like BERTScore (Zhang et al., 2020b) have not been thoroughly tested for meeting summarization (Kirstein et al., 2024). Yet, this assessment would be crucial because meeting summarization comes with unique challenges, such as high abstraction, low extraction rate, and complex reasoning (Gao and Wan, 2022).

This study examines how automatic metrics relate to human annotations for meeting summarization and what they actually measure in their scores. We aim to create a unified understanding of the challenges in meeting summarization and the errors that occur when these challenges are unmet. Using the QMSum dataset (Zhong et al., 2021), we have experts annotate the challenges in the transcripts and errors in automatically generated summaries using various models, including domain-standard encoder-decoder architectures and notable decoder-only models. This setup allows us to find connections between challenges, errors, and eight automatic metrics.

The results show problems with current automatic metrics for evaluating meeting summarization. *Structural disorganization* errors are often penalized, matching human judgments, but *hallucination* errors are sometimes rewarded. ROUGE (Lin, 2004) struggles to distinguish the impact of different errors on quality, even though it is good at penalizing omissions. Surprisingly, about a third of the metrics-error combinations either ignore or reward errors. For example, Perplexity favors *incorrect references*, and Lens favors *structural disorganization*. These observations highlight the need for better evaluation methods in meeting summarization. Our contributions are threefold:

- We conduct a comprehensive literature review to identify and unify the challenges specific

to English meeting summarization and the types of errors that commonly occur in model-generated summaries.

- We build the first direct correlations between the intrinsic challenges and observable errors they induce for encoder-decoder and decoder-only model architectures, providing a framework for tracking their impact.
- We rigorously evaluate the efficacy of nine prevalent automatic metrics, uncovering what they accurately measure, neglect, and their sensitivity to varying error severities, all corroborated by human annotations.

The codebase, annotations, and guidelines are available on GitHub:

https://github.com/FKIRSTE/emnlp2024-Meeting-Sum-Metrics.

2 Methodology

We conducted a focused literature review to identify challenges and observable errors when failing these in English abstractive meeting summarization. We target papers published between the introduction of the BART model (Lewis et al., 2020) in 2019, which serves as the typical backbone model for the field, and 2024. This scope allows us to focus on challenges pertaining to currently used Transformer-based models. We set additional inclusion criteria to prioritize relevance. We consider works that introduce novel methodologies or contribute to defining challenges, with a preference for transformer architectures, and exclude only loosely related studies, multi-lingual, multi-modal approaches, non-abstractive, and non-Transformer-based approaches. We considered publications from 2024, irrespective of their citation index, acknowledging their emergent influence. The survey follows the PRISMA checklist (Page et al., 2021) to prevent bias.

Google Scholar is chosen as the primary database for its comprehensive coverage and advanced query capabilities, yielding more works conforming to our review scope than Web of Science and DBLP.

Our search comprises two stages: The initial phase involves broader queries combining "meeting summarization" with adjunctive terms such as *challenge*, *literature review*, or *survey*, while the subsequent refinement uses challenge-related keywords and synonyms. Papers are ranked based on Google Scholar metrics, considering the top 100 from each query, and serve as primary source for our analysis.

We apply our criteria to the initial pool of 300 papers to select 18 core articles directly addressing challenges and 40 additional papers discussing challenges and errors implicitly through methodological contributions. Each shortlisted paper undergoes detailed data extraction to identify relevant challenges, methodologies, foundational models, and metrics collated in a structured format for cross-validation. This approach allows us to construct a comprehensive view of current challenges and observable errors in English meeting summarization.

3 Definitions

3.1 Challenges

This section summarizes key challenges in meeting summarization identified in the literature and their implications on the summarization process.

Spoken language. Handle colloquialisms, domain-specific terminology, and various forms of linguistic noise, such as false starts, repetitions, and filler words (Koay et al., 2020; Kumar and Kabiri, 2022; Antony et al., 2023), which can appear due to the spoken language nature of meetings. This challenge can affect the accuracy and clarity of the generated summaries.

Speaker dynamics. Accurately distinguish and track different speakers, their utterances, and specific roles (e.g., project manager, applicant), particularly when roles are topic-dependent (Khalifa et al., 2021; Gu et al., 2022; Rennard et al., 2023). Failing to do so can introduce biases and result in incomplete or flawed summaries, leaving out noteworthy elements.

Coreference. Manage the resolution of references to other speakers and previous actions to ensure coherent and complete summarization (Liu et al., 2021). Inaccurate handling can lead to ambiguous or incomplete summaries lacking context.

Discourse structure. Understand and track the inherent high-level structure and flow of a meeting (Li et al., 2023) throughout the different meeting phases. These phases may refer to multiple topics (Feng et al., 2022), e.g., during an argumentation, or swiftly shift from one topic to another (Shinde et al., 2022). Failure to consider this hierarchical

structure can result in summaries that are either incomplete or lack coherence.

Contextual turn-taking. Capture the evolving local dynamics of a meeting as it progresses through multiple speaker turns. The task involves accurately processing shifts in discourse complicated by interruptions, repetitions, and redundancies (Ma et al., 2022; Zhang and Zhao, 2021; Shinde et al., 2022). Inadequate capture can result in misleading or shallow summaries.

Implicit context. Account for unspoken or implied context, such as tacit organizational knowledge or prior discussions that are not explicitly referred to. Neglecting this can produce misleading or shallow summaries (Xu et al., 2022).

Low information density. Identify key points when salient information is sparse and unevenly distributed, especially relevant in decision-making scenarios (Rennard et al., 2023; Zhang et al., 2021). This challenge can affect the resulting summary's depth and level of salience.

Data scarcity. The lack of diverse, high-quality, real-world scenario training samples hampers model training (Rennard et al., 2023; Kumar and Kabiri, 2022; Jacquenet et al., 2019).

Long transcripts. Process long transcripts, which can result from longer meetings due to the quadratic computational cost of transformer-based models, leading to efficiency issues with increasing dialogue length (Kumar and Kabiri, 2022; Feng et al., 2022; Zhang et al., 2021). Only processing a sub-part of a meeting might exclude salient information and, therefore, result in an incomplete or incorrect summary.

Heterogeneous meeting formats. Different types of meetings necessitate distinct summarization approaches (Rennard et al., 2023). Failure to adapt can result in summaries that miss key points or include irrelevant information.

3.2 Error Types

Errors in summaries arise when challenges are not correctly addressed. We have collated error types from the literature into six principal categories, observed in the summaries generated during our experiments as detailed in Section 5.

Missing information (MI). This error involves missing information from the meeting, such as significant decisions or actions (Zou et al., 2023; Chen and Yang, 2020). We consider *total omission*, where relevant topics or points discussed are entirely absent from the summary, and *insufficient detail*, where the summary mentions a salient topic but does not capture its depth or the detailed discussion from the original meeting.

Redundancy (Red). The summary contains repeated or redundant information, impacting brevity and clarity (Chen and Yang, 2020). Such repetitions can manifest in different ways: reiterated key points, overuse of individual words, or duplicating entire phrases.

Wrong references (WR). The model misattributes statements, opinions, or actions to incorrect meeting participants or omits their mention altogether (Chen and Yang, 2020).

Incorrect reasoning (IR). The model draws conclusions that are not supported by the discussions in the meeting (Chen and Yang, 2020).

Hallucination (Hal). The model produces inconsistencies, such as incorrect dates, names, or locations, not aligned with the meeting content (Wang et al., 2022; Ji et al., 2023). This encompasses *intrinsic* hallucinations, introducing events not present in the original meeting or contradicting the input, and *extrinsic* hallucinations, misrepresenting actual events.

Incoherence (Inc). The model generates summaries with disjointed logic or flow. This manifests as *intra-sentence* disconnections and *inter-sentence* inconsistencies, such as flawed progression and incorrect, contrastive expressions (Wang et al., 2023).

We have identified two additional error categories during our human annotation process, as detailed in Section 4.3. Notably, our review of the pertinent literature reveals that these categories have not been explicitly defined in existing studies in a way that matches our findings.

Linguistic inaccuracy (LI). The model uses inappropriate, incorrect, or ambiguous language or fails to capture unique linguistic styles. This error type spans issues from adopting unsuitable words from the source to grammar mistakes and employing contextually unclear or ambiguous terms.

Structural disorganization (SD). The model may produce summaries that misrepresent the original

order or logic of the meeting's discourse, misplacing topics or events. This error covers only when the order of stated events is wrong. Including non-existing events or excluding events does not count towards this error.

4 Experimental Framework

4.1 Models

We use the Longformer Encoder Decoder (LED) (Beltagy et al., 2020) as our primary model for its strong summarization performance. LED employs local+global sliding window attention in the encoder and full self-attention in the decoder, efficiently handling lengthy documents. To comprehensively assess challenges and errors in encoder-decoder architectures, we further consider DialogLED (Zhong et al., 2022) and PEGASUS-X (Phang et al., 2022). DialogLED extends LED with dialogue-centric pre-training, while PEGASUS-X enhances PEGASUS (Zhang et al., 2020a) for extended inputs using staggered block-local attention. All models are finetuned on the QMSum training subset (general summaries) for three epochs.

We also explore large language models, using GPT3.5 turbo via ChatGPT and Zephyr-7B-α, a refined version of Mistral-7B-v0.1 [1]. Mistral's adaptations employ sliding window attention, outperforming Llama2 (Touvron et al., 2023). To handle context size limitations, we use a chunk-based approach (Bhaskar et al., 2023) with the prompt "Create a TL;DR of the following meeting chunk," inspired by a Microsoft guideline [2]. LLMs are used with a zero-shot setup.

For the experiments, the models are grouped into encoder-decoder models containing LED, DialogLED, and Pegasus-X, and the decoder-only models, i.e., GPT-3.5 turbo and Zephyr-7B-α. This grouping is motivated by similar architecture and error distribution and frequency patterns as shown in Table 8.

4.2 Datasets

We use as input samples from QMSum (Zhong et al., 2021), an established dataset for query-based multi-domain meeting summarization. It includes transcripts from academic (ICSI), product (AMI), and committee meetings (Welsh and Canadian Parliament: WPCP). ICSI (Janin et al., 2003) offers informal research meetings with linguistic challenges. AMI (Mccowan et al., 2005) provides staged meetings with natural dialogue dynamics. WPCP presents formal, agenda-driven discussions from UK and Canada Parliament committee meetings. Detailed statistics are in Table 1. While there are other datasets out there in the context of meeting summarization, such as the frequently used MeetingBank (Hu et al., 2023) and ELITR (Nedoluzhko et al., 2022) datasets, we do not include these. MeetingBank is conceptually close to the WPCP meetings, therefore not enhancing the diversity of our data selection and potentially introducing an imbalance in meeting type. ELITR aims to produce meeting minutes with bullet points containing key insights and actions discussed. Therefore, it does not fit the abstractive summaries we aim to investigate.

4.3 Metrics

We align human annotations with automatic metrics considering count-based, model-based, and QA-based methodologies, chosen based on their prevalent use in meeting and dialogue evaluation (Gao and Wan, 2022).

Count-based. ROUGE (Lin, 2004) is the default-used metric suite that assesses the overlap between n-grams in generated summaries that appear in the reference. Researchers mainly consider unigrams, bigrams, and the longest common sequence to gauge the relevance and accuracy of generated content. BLEU (Papineni et al., 2002) evaluates how many n-grams from the reference appear in the generated summary. The score is designed to reflect the precision of the generated text, focusing primarily on lexical similarity to the reference. METEOR (Banerjee and Lavie, 2005) builds on BLEU by accounting for synonyms, word stems, and sentence structure, offering a holistic assessment of lexical, syntactic, and semantic alignment in precision and recall.

Model-based. BERTScore (Zhang et al., 2020b) measures the contextual similarity between generated and reference texts using a pre-trained BERT model, reflecting semantic and syntactic similarity [3]. Perplexity (PPL) measures a language model's

[1] https://huggingface.co/mistralai/Mistral-7B-v0.1

[2] https://www.microsoft.com/en-us/microsoft-365-life-hacks/organization/using-chatgpt-creating-meeting-agendas-minutes-notes

[3] We report the rescaled BERTScore-F score: https://github.com/Tiiiger/bert_score/blob/master/journal/rescale_baseline.md

Dataset	# Meetings	# Turns	# Speakers	avg. Len. of Meet.	avg. Len. of Sum.
AMI	137	535.6	4.0	6007.7	70.5
ICSI	59	819	6.3	13317.3	53.7
WPCP	36	207.7	34.1	13761.9	80.5
all (QMSum)	232	556.8	9.2	9069.8	69.6

Table 1: Statistics for the QMSum subsets and the entire QMSum meeting summarization dataset.

	Spoken Language	Speaker Dynamic	Co-reference	Discourse Structure	Contextual Turn-Taking	Implicit Context	Data Scarcity	Low Inf. Density	Overall
detection	0.84	0.77	0.68	0.79	0.73	0.90	0.87	0.78	0.79
frequency	0.84	0.71	0.78	0.70	0.89	0.91	0.90	0.84	0.82

Table 2: Krippendorff's alpha for inter-annotator agreement on challenges and their frequencies, with challenges (abbreviations) in top row ordered as in Section 3.1.

uncertainty in predicting words, evaluating the quality and fluency of utterances. We use GPT-2 (Radford et al., 2019) as the language model. BLANC (Lita et al., 2005) measures how well a generated summary aids a language model in understanding the original document, reflecting informativeness. LENS (Maddela et al., 2023) is a trainable metric that assesses the alignment of generated text with human references in content and style. While this metric is also explored for meeting summarization, it is noteworthy that it is based on RoBERTa, adopting its maximum context length of 512 tokens, making it technically less suitable for assessing meeting summaries considering transcripts.

QA-based. QuestEval (Scialom et al., 2021) combines FEQA (Durmus et al., 2020) and SummaQA (Scialom et al., 2019) scores, using a question-answering model to answer questions formed from the reference text (SummaQA) or the generated summary (FEQA), extracting answers from the opposite source. QuestEval evaluates factuality, coherence, informativeness, and relevance.

Human annotation. We adapt proven methodologies (Zhang et al., 2023) for a thorough annotation process involving four graduate students from diverse backgrounds (i.e., computer science, psychology, communication science), all well-versed in English and familiar with meeting summarization. From the QMSum general-summary test set, we choose 35 general-summary samples, each containing meeting transcripts, gold summaries, and model-generated summaries, resulting in 175 distinct samples for annotation. Annotators identify challenges (Section 3.1) in the transcript and errors Section 3.2 in the generated summaries using yes/no questions such as: "Does the given summary omit crucial information or provide insufficient detail about salient points?" corresponding to the *missing information* error. The annotators further rate the frequency of challenges and the impact of errors for the LED model using Likert scales from rare occurrence or minimal impact (1) to frequent presence or high impact (5).

To ensure reliability and consistency, we assess inter-annotator agreement using Krippendorff's alpha, achieving an average of 0.81 (see Tables 2 and 3). A preliminary pilot test serves as annotator training and refinement of guidelines. Regular review meetings are held to maintain consistency. Annotators also highlight summary segments with errors. Discrepancies are discussed, and an expert annotator is available to discuss complex issues.

A full presentation of the annotation process is stated in Appendix A. Details on annotated labels are shown in Appendix B.

5 Analysis

5.1 Linking challenges and error types

We analyze the relationship between challenges and frequently observed errors for encoder-decoder and decoder-only architectures using Point-biserial correlation based on human annotations from Section 4.3. Table 4 presents the results.

Encoder-decoder models exhibit strong links between *incoherence*, *structural disorganization*, and *redundancy* errors and challenges like *spoken language* and *speaker dynamics*, suggesting struggles with maintaining coherence. Conversely, *wrong references*, *linguistic inaccuracy*, and *hallucination*

		MI	Red	WR	IR	Hal	Inc	LI	SD	overall
detection	Encoder-Decorder	0.74	0.875	0.83	0.87	0.79	0.61	0.86	0.71	0.78
Decoder-only	0.76	0.87	0.85	0.79	0.77	0.90	0.90	0.71	0.82	
error impact	(LED)	0.82	0.92	0.85	0.73	0.80	0.89	0.87	0.79	0.83

Table 3: Krippendorff's alpha for inter-annotator agreement on errors for encoder-decoder (i.e., LED, DialogLED) and decoder-only models (i.e., GPT-3.5, Zephyr-7B-α). For the LED model, the agreement on error impact is also reported. Errors (abbreviations) in the top row are ordered as in Section 3.2.

Errors	**Encoder-Decoder**	**Decoder-only**
missing information	discourse structure*, implicit context	coreference*
redundancy	speaker dynamics, contextual turn-taking* implicit context*, decision dynamics*	spoken language*, speaker dynamics contextual turn-taking, low info density
wrong references	(none)	spoken language**, contextual turn-taking* low information density**
incorrect reasoning	speaker dynamics, implicit context low information density	(none)
hallucination	contextual items, implicit context*	coreference*, contextual turns* implicit context
incoherence	spoken language*, coreference* contextual turn-takin*, decision dynamics*	coreference
linguistic inaccuracy	coreference, low information density	spoken language*, speaker dynamics** decision dynamics*, low information density
structural disorganization	spoken language**, speaker dynamics** contextual turns*, decision dynamics low information density	coreference*, contextual turns implicit context*, decision dynamics

Table 4: The linkage between challenges and errors. No asterisk indicates low correlation, * signifies mid correlation ($p \leq 0.05$), and ** denotes high correlation ($p \leq 0.01$).

errors show limited associations with the examined challenges. Specifically, the limited association of *hallucination* reinforces current understandings, suggesting that the underlying causes remain elusive (Maynez et al., 2020). Notably, the *implicit context* challenge works as a proxy for *hallucination*. The minimal correlation of the *wrong reference* error, which rarely occurs due to the sentence structure in generated summaries, and the *linguistic inaccuracy* error indicate that challenges such as *coreference* are well-handled by these models.

Regarding specific challenges, the *discourse structure* challenge correlates slightly with the *missing information* error, indicating occasionally missed details within distinct phases. The *low information density* challenge weakly correlates with the *incorrect reasoning* error, revealing difficulties in extracting salient details. The *contextual turn-taking* challenge aligns with *redundancy* and *incoherence* errors, suggesting issues in capturing dynamics on a granular level within the different meeting phases.

Decoder-only models show different patterns, with *incorrect reasoning* being the rarest and weakest correlated error. *Wrong references*, *redundancy*, and *linguistic inaccuracy* errors are most prevalent, aligning with *low information* and *spoken language* challenges. For LLMs, redundancies manifest as repetitive introductory sentences, while linguistic inaccuracies emerge as contextually ambiguous terms. *Missing information* and *structural disorganization* errors are equally prominent and strongly correlated, linking to the *coreference* challenge and emphasizing models' tendencies to list topics without proper context. Some negative correlations suggest specific challenges might decrease the likelihood of certain errors, such as for *low information density* challenge and *missing information* error, warranting further exploration.

5.2 Correlation of automatic metrics and human annotation

Though numerous and frequently applied, existing automatic metrics are not tailored to the intricacies of meeting summarization. We analyze their reactions to the diverse error types observable in meeting summaries by answering the following three research questions.

RQ1: How do automatic metrics correlate with human assessments? Table 5 shows Point-

	MI	Red	WR	IR	Hal	Inc	LI	SD
ROUGE-1	**-0.40***	0.17	-0.07	-0.18	0.05	**-0.30**	**-0.12**	**-0.41***
ROUGE-2	-0.20	0.10	0.07	**-0.29**	**0.10**	**-0.22**	**-0.10**	-0.35*
ROUGE-L	**-0.29**	0.08	-0.04	-0.21	0.02	-0.21	-0.09	**-0.46****
BLEU	-0.20	0.08	-0.09	-0.08	**0.32**	-0.13	0.05	-0.26
METEOR	**-0.26**	**0.27**	**0.16**	**-0.23**	0.08	-0.20	0.02	-0.38*
BERTScore (F)	-0.16	**0.22**	0.09	-0.06	**0.10**	-0.02	-0.01	-0.26
PPL	-0.10	-0.10	**0.44****	**-0.32**	-0.08	0.09	**0.24**	-0.17
BLANC (TS)	-0.19	0.05	-0.13	-0.13	-0.13	**-0.42***	-0.09	-0.36*
LENS	0.01	**-0.38***	**-0.17**	0.20	0.01	0.03	-0.03	**0.45****
QuestEval (F)	0.10	-0.05	**0.16**	0.02	0.26	0.00	-0.03	-0.28

Table 5: Point-biserial correlation between automatic metrics and annotated errors on samples generated by LED. * denotes significance at $p \leq 0.05$ and ** at $p \leq 0.01$. The top row shows abbreviated error types from Section 3.2 in order. A negative correlation indicates worsening metric scores with increasing error occurrence. The three highest absolute values are highlighted in **bold**. We present F-scores (F) for BERTScore and QuestEval, with BERTScores being further rescaled.

biserial correlations between automatic metrics and expert-annotated errors for LED model-generated summaries. Our analysis reveals that no metric consistently correlates highly with all error types, underscoring the complexity of meeting summarization and the absence of a universal metric that captures human judgment well (Gao and Wan, 2022). However, some metrics show trends aligning with human judgment through significant negative scores. Though not designed for structural coherence, several metrics show significant negative correlations with *structural disorganization* errors, indicating that temporal and logical disorganization breaks n-gram sequences and semantic flow. ROUGE-1 exhibits a more significant score than ROUGE-2 and ROUGE-L, aligning with observations in dialogue summarization (Gao and Wan, 2022). In particular, the gaps, disjoint narratives, and incorrect statements from errors such as *missing information*, *incoherence*, and *incorrect reasoning* influence metric scores by not aligning with reference n-grams or shifting meaning (Lin, 2004). *Redundancy*, *wrong references*, and *hallucination* errors remain less frequently detected, as summaries with these errors can still closely match the reference in terms of n-gram overlap and semantic similarity.

As expected, count-based metrics are responsive to *missing information* (Lin, 2004), with ROUGE-1 correlating significantly with information omissions (Gao and Wan, 2022) and BLEU identifying summaries with key content omissions. Count-based metrics significantly correlate with *structural disorganization* errors though not designed to detect this error, possibly because of disrupted n-gram sequences due to reorganized content (Banerjee and Lavie, 2005). ROUGE and Meteor are potential indicators of incorrect reasoning and incoherence.

Among model-based metrics, BERTScore exhibits sensitivity to *missing information*, reflecting semantic and contextual differences between candidate and gold summaries (Zhang et al., 2020b). However, its correlation with *structural disorganization* is less direct, hinting at the influence of sentence alignment and structural coherence. BLANC, designed for evaluating coherence and fluency, correlates as expected negatively with incoherence. LENS effectively captures redundancy errors. Model-based metrics present milder correlations, possibly due to discrepancies between training data and meeting contexts (Gao and Wan, 2022). While these metrics seem to not align as well with human judgment as count-based metrics, they offer insights into errors not captured by count-based metrics.

The correlation table shows that no single metric predominantly captures all error types, with most correlations being weak to moderate. Examining Table 5, a combination of metrics like ROUGE, BLANC, and LENS could serve as a proxy for various errors, but this approach warrants additional evaluation and score weighting.

RQ2: Do automatic metrics mask errors? We categorize "masking" as indifference to an error (near-zero correlation) or positive reinforcement of a mistake (positive correlation). Count-based metrics predominantly show near-zero or negative correlations with errors, indicating they might not sufficiently penalize specific mistakes. This behav-

	MI	Red	WR	IR	Hal	Inc	LI	SD
ROUGE-1	0.07	**0.24**	-0.02	-0.17	0.10	**-0.30**	-0.09	**-0.42***
ROUGE-2	-0.05	0.15	0.06	**-0.29**	0.13	**-0.23**	**-0.11**	-0.36*
ROUGE-L	0.03	0.15	-0.05	-0.22	0.10	-0.21	-0.08	**-0.44****
BLEU	-0.11	0.08	-0.06	-0.09	**0.35***	-0.14	0.03	-0.31
METEOR	-0.06	**0.21**	0.16	**-0.27**	0.13	-0.21	-0.05	-0.36*
BERTScore (F)	**-0.25**	0.19	0.08	-0.10	**0.15**	-0.02	-0.04	-0.31
PPL	-0.02	-0.13	**0.42**	**-0.29**	0.02	0.00	**0.23**	-0.13
BLANC	-0.03	0.04	-0.07	-0.15	-0.11	**-0.44**	**-0.13**	-0.36*
LENS	**-0.14**	**-0.42***	**-0.26**	0.22	-0.07	0.00	-0.06	**0.40**
QuestEval (F)	**0.13**	0.04	**0.18**	0.05	**0.34***	-0.03	-0.07	-0.24

Table 6: Spearman correlation between metrics and annotated error impacts, using summaries generated by LED as base. * denotes significance at $p \leq 0.05$ and ** at $p \leq 0.01$. The top row lists abbreviated error types from Section 3.2 in sequence. A negative correlation implies declining metric scores with rising error instances. The three most pronounced absolute values are emphasized in **bold**. F-scores (F) are given for BERTScore and QuestEval, with BERTScores being further rescaled.

ior is expected (Saadany and Orasan, 2021; Akter et al., 2022), especially for errors these metrics were not explicitly designed to detect or for which they can only act as a proxy.

From Table 5 using summaries generated by LED as a base, we observe that some metrics show their known struggle with specific error types through near-zero scores (e.g., BERTScore and linguistic inaccuracy (Hanna and Bojar, 2021)), model-based metrics occasionally manifest positive correlations. Perplexity favors incorrect references if they preserve linguistic coherence and fluency, aligning with the language model's expectations and yielding a lower (better) score. LENS appears to struggle with capturing broader logical and temporal structures. Despite being designed for text simplification, it significantly correlates with the *structural disorganization* error and primarily emphasizes word-level and intra-sentence simplification (Maddela et al., 2023). While QuestEval and BETScore do not exhibit significant correlations, they show noteworthy trends: QuestEval may reward errors like *missing information* and *hallucination*, a counterintuitive outcome given its focus on factual accuracy. The finding could indicate that hallucinated details likely bypass metric detection due to generated questions lacking the rigor to spot them, which can be linked to the gap between pre-training and meeting data (Gao and Wan, 2022). BERTScore, emphasizing semantic and syntactic similarities, may reward redundant yet correct details (Hanna and Bojar, 2021).

RQ3: How do the metric scores vary with the severity of the error? To identify the relationship between metric scores and error severity, we analyze how these scores fluctuate by the impact of errors. Table 6 shows the Spearman correlation trends, using summaries generated by LED as a base. Across different models and error severities, most correlations are negligible to weak, and many lack statistical significance. This observation emphasizes the limitations of current metrics in discerning error severity in meeting summarization.

BLEU and QuestEval display significant positive correlations with *hallucination* errors, underscoring their vulnerability to hallucinated content. While metrics such as BERTScore and LENS show sensitivity to *missing information* in RQ1, their precision in assessing error severity is limited. This limitation may stem from the gap between training data and meeting transcripts for the metrics utilizing language models, as the meeting transcript was provided as part of the input (Gao and Wan, 2022). LENS, closely aligned with text simplification, penalizes *redundancy*. These observations underscore that some metrics are intrinsically responsive to specific errors, even if not explicitly designed for them. Perplexity does not exhibit significant correlations.

6 Related Work

Research on meeting summarization typically addresses challenges and error types in isolation, leaving their interrelation unexplored. For dialogue summarization, challenges are collected (Chen and Yang, 2020) but lack the detail required for meeting contexts. Previous works by Tang et al. (2022); Wang et al. (2022) lay the groundwork for our error

typologies, enhanced with insights from diverse studies and annotator feedback. Chen and Yang (2020) correlate challenges and errors in dialogue summarization, but their findings do not entirely transfer to English abstractive meeting summarization, though their annotation approach informs ours. Automatic metrics for meeting summaries are underexplored, but similar studies exist for text summarization using the CNN/Dailymail dataset (Fabbri et al., 2021) and dialogue summarization via the SAMSum dataset (Gao and Wan, 2022). We build upon these works by selectively adopting their metrics and augmenting them with measures commonly reported in recent meetings and dialogue summarization studies.

7 Conclusion

We developed a resource suite for meeting summarization evaluation, covering domain-specific challenges, linked error types, predictive correlations for model architectures with known challenges, and expert annotations of the QMSum subset. Our analysis highlighted the limitations and misalignments of current metrics with human judgment in discerning error nuances. Metrics that work well for other summarization tasks either did not react to errors or cannot reflect the impact on quality in their scores. A composite metric may be more effective, but its formulation requires further research. Recent advancements highlighted the potential of LLM-based metrics. Utilizing an LLM prompted with our detailed annotator guidelines and supplemented by examples presents a promising approach for detecting complex errors like structural disorganization and incorrect reasoning. Given the current evolution of techniques, we plan to extend the work as a (dialogue) summarization benchmarking pipeline to capture better how upcoming techniques handle challenges and how well new metrics capture their performance. We encourage the research community to contribute model outputs and introduce new metrics to this initiative.

Acknowledgements

This work was supported by the Lower Saxony Ministry of Science and Culture and the VW Foundation. Frederic Kirstein was supported by the Mercedes-Benz AG Research and Development.

Limitations

Our study, while offering a comprehensive analysis of challenges and errors in meeting summarization, primarily focuses on English-language summaries. This linguistic focus may lead to variations in challenges and errors across languages. Some challenges, especially those subtle to human perception, might be overlooked, creating potential gaps in our annotations. Our reliance on Google Scholar, despite its broad coverage, has its criticisms, as it tends to favor citation counts (Fagan, 2017) and may include less reputable sources (Beall, 2016).

The QMSum dataset, with its 35 samples, provides statistical significance but represents only a fraction of potential meeting types. Thus, findings may vary across different meeting contexts.

We include a total of 175 samples, comparable to the original QMSum dataset (i.e., 232 samples), one of the most used datasets in meeting summarization. The considered dataset maintains a diverse representation of meeting types (academic, business, parliament) within the dataset. We acknowledge the limitation that QMSum might not cover all meeting types in existence. However, we focus on creating a unified understanding of the challenges in meeting summarization and the errors that occur when these challenges are unmet. For that, we used QMSum as a proxy.

The challenges we associate with QMSum are inferred from human annotations rather than directly tested, and our model selection, while reflective of the current landscape, does not capture every variant. Our choice of encoder-decoder models is comprehensive, but we miss out on architectures like hierarchical models due to accessibility issues, which hamper the comparability of the models. The array of large language models is continually expanding, and our selection is based on the standings on the Huggingface LLM Leaderboard at the time of writing. Our method of linking challenges to errors is holistic, yet it might dilute strong connections between them, as we did not test the linking with isolated challenges. Specific metrics, like LENS and QuestEval, may have biased scores since they use the meeting transcript as input and are not domain-trained. Lastly, our findings, rooted in the nuances of meeting summarization, might not seamlessly apply to broader summarization domains like dialogue summarization, given each domain's distinctive traits.

Ethics Statement and Broader Impact

Our research abides by ethical guidelines for AI research and is committed to privacy, confidentiality, and intellectual property rights. We have ensured that the datasets in our study, which are publicly available, do not house sensitive or personal details. While our study leverages existing resources and generative models, it is important to note that these models can possess biases and may occasionally generate summaries with distortions, biases, or inappropriate content (Gooding, 2022). We have configured our models to omit potentially harmful or unsafe content to counteract this. While our research aims to enhance meeting summarization to benefit communication and productivity across sectors, we are acutely aware of the ethical challenges posed by AI in this domain. Meeting summarization models must be wielded with respect to privacy and consent, especially when processing sensitive or confidential material. It's paramount that these models neither violate privacy nor perpetuate harmful biases. As the field evolves, we stress the importance of maintaining these ethical considerations and encourage fellow researchers to uphold them, ensuring that AI advancements in meeting summarization are both beneficial and ethically grounded. An integral aspect of our ethical commitment is reflected in our approach to annotator recruitment and management. The team of annotators, consisting of interns, student assistants, and doctoral students, was meticulously selected through internal channels. This strategy was chosen to uphold a high standard of annotation quality—a quality we found challenging to guarantee through external platforms such as Amazon Mechanical Turk. Ensuring fair compensation, these annotators were reimbursed following institutional guidelines for their respective positions. Further, flexibility in the annotation process was also a priority. Annotators were free to choose their working times and environments to prevent fatigue from affecting their judgment.

References

Mousumi Akter, Naman Bansal, and Shubhra Kanti Karmaker. 2022. Revisiting Automatic Evaluation of Extractive Summarization Task: Can We Do Better than ROUGE? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1547–1560, Dublin, Ireland. Association for Computational Linguistics.

Dinu Antony, Sumit Abhishek, Sujata Singh, Siddu Kodagali, Narayana Darapaneni, Mukesh Rao, Anwesh Reddy Paduri, and Sudha BG. 2023. A Survey of Advanced Methods for Efficient Text Summarization. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0962–0968.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

J Beall. 2016. Best practices for scholarly authors in the age of predatory journals. *The Annals of The Royal College of Surgeons of England*, 98(2):77–79. PMID: 26829665.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted Opinion Summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics.

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.

Arman Cohan and Nazli Goharian. 2016. Revisiting Summarization Evaluation for Scientific Articles. pages 806–813.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A Question Answering Evaluation Framework for Faithfulness Assessment in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Jody Condit Fagan. 2017. An evidence-based review of academic web search engines, 2014-2016: Implications for librarians' practice and research agenda. *Information Technology and Libraries*, 36(2):7–47.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. A Survey on Dialogue Summarization: Recent Advances and New Frontiers.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting Summarization Evaluation for Dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–

5709, Seattle, United States. Association for Computational Linguistics.

Sian Gooding. 2022. On the Ethical Considerations of Text Simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.

Jia-Chen Gu, Chongyang Tao, and Zhen-Hua Ling. 2022. Who Says What to Whom: A Survey of Multi-Party Conversations. In *Thirty-First International Joint Conference on Artificial Intelligence*, volume 6, pages 5486–5493.

Michael Hanna and Ondřej Bojar. 2021. A Fine-Grained Analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A Benchmark Dataset for Meeting Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. Association for Computational Linguistics.

Maor Ivgi, Uri Shaham, and Jonathan Berant. 2022. Efficient Long-Text Understanding with Short-Text Models.

Francois Jacquenet, Marc Bernard, and Christine Largeron. 2019. Meeting Summarization, A Challenge for Deep Learning. In Ignacio Rojas, Gonzalo Joya, and Andreu Catala, editors, *Advances in Computational Intelligence*, volume 11506, pages 644–655. Springer International Publishing, Cham.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, Nathaniel Morgan, B. Peskin, Thilo Pfau, Elizabeth Shriberg, A. Stolcke, and Chuck Wooters. 2003. The ICSI meeting corpus. pages I–364.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):248:1–248:38.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown. 2021. A Bag of Tricks for Dialogue Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8014–8022, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Frederic Kirstein, Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2024. CADS: A Systematic Literature Review on the Challenges of Abstractive Dialogue Summarization.

Jia Jin Koay, Alexander Roustai, Xiaojin Dai, Dillon Burns, Alec Kerrigan, and Fei Liu. 2020. How Domain Terminology Affects Meeting Summarization Performance. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5689–5695, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lakshmi Prasanna Kumar and Arman Kabiri. 2022. Meeting Summarization: A Survey of the State of the Art.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chuyuan Li, Patrick Huber, Wen Xiao, Maxime Amblard, Chloe Braud, and Giuseppe Carenini. 2023. Discourse Structure Extraction from Pre-Trained and Fine-Tuned Language Models in Dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2562–2579, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Lucian Vlad Lita, Monica Rogati, and Alon Lavie. 2005. BLANC: learning evaluation metrics for MT. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, pages 740–747, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Feifan Liu and Yang Liu. 2008. Correlation between ROUGE and Human Evaluation of Extractive Meeting Summaries. In *Proceedings of ACL-08: HLT, Short Papers*, pages 201–204, Columbus, Ohio. Association for Computational Linguistics.

Zhengyuan Liu, Ke Shi, and Nancy F. Chen. 2021. Coreference-Aware Dialogue Summarization. ArXiv:2106.08556 [cs].

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Comput. Surv.*, 55(5).

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A Learnable Evaluation Metric for Text Simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Iain Mccowan, J Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, V Karaiskos, M Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Wilfried Post, Dennis Reidsma, and P Wellner. 2005. The AMI meeting corpus.

Joseph Mroz, Joseph Allen, Dana Verhoeven, and Marissa Shuffler. 2018. Do We Really Need Another Meeting? The Science of Workplace Meetings. 27:096372141877630.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR Minuting Corpus: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3174–3182, Marseille, France. European Language Resources Association.

Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372:n71.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jason Phang, Yao Zhao, and Peter J. Liu. 2022. Investigating Efficiently Extending Transformers for Long Input Summarization. ArXiv:2208.04347 [cs].

Hendri Pratama, Mohamed Nor Azhari Azman, G. Kassymova, and Shakizat Duisenbayeva. 2020. The Trend in Using Online Meeting Applications for Learning During the Period of Pandemic COVID-19: A Literature Review. 1:58–68.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2023. Abstractive Meeting Summarization: A Survey.

Hadeel Saadany and Constantin Orasan. 2021. BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-Oriented Text. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 48–56, Held Online. INCOMA Ltd.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization Asks for Fact-based Evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers Unite! Unsupervised Metrics for Reinforced Summarization Models.

Kartik Shinde, Tirthankar Ghosal, Muskaan Singh, and Ondrej Bojar. 2022. Automatic Minuting: A Pipeline Method for Generating Minutes from Multi-Party Meeting Proceedings. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 691–702, Manila, Philippines. De La Salle University.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. CONFIT: Toward Faithful Dialogue Summarization with Linguistically-Informed Contrastive Fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5657–5668, Seattle, United States. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pancheng Wang, Shasha Li, Shenling Liu, Jintao Tang, and Ting Wang. 2023. Plan and generate: Explicit and implicit variational augmentation for multi-document

summarization of scientific articles. *Information Processing & Management*, 60(4):103409.

Jiabao Xu, Peijie Huang, Youming Peng, Jiande Ding, Boxi Huang, and Simin Huang. 2022. Adjacency Pairs-Aware Hierarchical Attention Networks for Dialogue Intent Classification. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7622–7626. ISSN: 2379-190X.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, pages 11328–11339. JMLR.org.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating Text Generation with BERT.

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

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. An Exploratory Study on Long Dialogue Summarization: What Works and What's Next.

Zhuosheng Zhang and Hai Zhao. 2021. Advances in Multi-turn Dialogue Comprehension: A Survey.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Yicheng Zou, Kaitao Song, Xu Tan, Zhongkai Fu, Qi Zhang, Dongsheng Li, and Tao Gui. 2023. Towards Understanding Omission in Dialogue Summarization.

A Annotation process details

Following, we describe the details of our annotation process.

Annotator selection: Our annotation team comprises four graduate students, officially employed as interns or doctoral candidates through standardized contracts. We select them from a pool of volunteers based on their availability to complete the task without time pressure and their English proficiency (native speakers or C1-C2 certified). This ensures they can comprehend meeting transcripts, human-written gold summaries from QMSum, and model-generated summaries. We aimed for gender balance (two male, two female) and diverse backgrounds, resulting in a team of two computer science students, one psychology student, and one communication science student, aged 24-28.

Preparation: We prepare a comprehensive handbook for our annotators, detailing the project context and defining challenges and error types. Each definition includes two examples: one with minimal impact (e.g., slight information redundancy) and one with high impact (e.g., repeated information throughout). The handbook explains the two-part rating system: a binary yes/no for the existence of a challenge or error, followed by a 1-5 Likert scale impact/frequency rating if a characteristic is observed as existing. This impact/frequency scoring is only used for the challenges (frequency, 1 low - 5 high) and errors (impact, 1 low - 5 high) produced by the LED model. Annotators are further tasked to provide reasoning for each decision. The handbook does not specify an order for processing errors or challenges. We provide the handbook in English and the annotators' native languages, using professional translations. The handbook could be used throughout the whole annotation process as a reference.

We set up a five-week timeline for the annotation process, preceded by a one-week onboarding period. The first two weeks feature twice-weekly check-ins with annotators, which are reduced to weekly meetings for the following three weeks. Separate quality checks without the annotators are scheduled weekly. (Note: week refers to a regular working week)

Onboarding: The onboarding week is dedicated to getting to know the project and familiarizing with the definitions and data. We begin with a kick-off meeting to introduce the project and explain the handbook, mainly focusing on each definition. We note initial questions to revise the handbook potentially. Annotators are provided with 35 samples generated by SLED+BART (Ivgi et al., 2022), chosen for their balance of identifiable errors and good-quality summaries while capable of processing the whole meeting. After the first 15 samples, we hold individual meetings to clarify any confu-

sion and update the guidelines accordingly. The remaining 20 samples are then annotated using these updated guidelines. A second group meeting this week addresses any new definitional issues. After the group meeting, we meet individually with annotators to review their work, ensuring quality and understanding of the task and samples. All four annotators demonstrate reliable performance and good comprehension of the task and definitions, judging from the reasoning they provided for each decision and annotation. We computed an inter-annotator agreement score using Krippendorff's alpha, achieving 0.81, indicating sufficiently high overlap.

Annotation Process: We continue the annotation process similarly. Each week, we distribute all 35 samples generated by one model to one of the annotators. Consequently, one annotator works through all samples of one model in one week. After five weeks, all samples have been processed by all annotators. Annotators are unaware of the summary-generating model and are given a week to complete their set at their own pace and break times. Quiet working rooms were provided if needed for concentration. To mitigate position bias, the sample order is randomized for each annotator. Annotators can choose their annotation order for each sample and are allowed to revisit previous samples to adapt ratings. To simplify the process, we frame each error type as a question, such as "Does the summary omit crucial information or provide insufficient detail about salient points?" for the missing information error.

Regular meetings are held to address questions on definitions or emerging issues. During the quality checks the authors perform, we look for incomplete annotations, missing explanations, and signs of misunderstanding judging from the provided reasoning. If we find such a lack of quality, the respective annotator will be notified to re-do the annotation. The quality checks are not used to bias annotators in their ratings but to ensure a complete and consistent dataset. After the five weeks, we compute inter-annotator agreement scores (shown in Tables 2 and 3). In case we observe a significant difference across annotators at this point, we have planned a dedicated meeting to discuss such cases with all annotators and a senior annotator to ensure that an understanding issue of the task or definition did not lead to the different ratings.

Annotators spend 43 minutes per sample, completing about seven samples daily.

Handling of scheduling conflicts: Given that our annotators have other commitments, we anticipate potential scheduling conflicts. We allow flexibility for annotators to complete their samples beyond the week limit if needed, reserving a sixth week as a buffer. Despite these provisions, all annotators completed their assigned samples within the original weekly timeframes.

B Statistics on Annotation Labels

In Tables 7 and 8, we report statistics of the annotated dataset, showing how many are showing the respective challenge and error type.

	Spoken Language	Speaker Dynamic	Co-reference	Discourse Structure	Contextual Turn-Taking	Implicit Context	Data Scarcity	Low Inf. Density
# detections	32 (91%)	35 (100%)	35 (100%)	34 (97%)	33 (94%)	5 (14%)	30 (86%)	35 (100%)

Table 7: Statistics on the occurance of different challenges in the QMSum samples used as input. We report total number and corresponding percentage.

	MI	Red	WR	IR	Hal	Inc	LI	SD
LED	31 (89%)	14 (40%)	4 (11%)	7 (20%)	9 (26%)	13 (37%)	3 (9%)	20 (57%)
DialogLED	33 (94%)	18 (51%)	4 (11%)	8 (23%)	8 (23%)	18 (51%)	3 (9%)	22 (63%)
Pegasus-X	34 (97%)	27 (77%)	14 (40%)	16 (46%)	13 (37%)	23 (66%)	10 (29%)	28 (80%)
GPT3.5	34 (97%)	7 (20%)	3 (9%)	1 (3%)	5 (14%)	9 (26%)	9 (26%)	22 (63%)
Zephyr	34 (97%)	7 (20%)	3 (9%)	1 (3%)	8 (23%)	11 (31%)	11 (31%)	22 (63%)

Table 8: Statistics on the occurance of different error types in the model-generated summaries. We report total number and corresponding percentage.

Citation for this Paper

```
@inproceedings{kirstein-etal-2024-evaluation,
 author={Kirstein, Frederic and Wahle, Jan Philip and Ruas, Terry and Gipp, Bela},
 title={What's under the hood: Investigating Automatic Metrics on Meeting Summarization},
 booktitle={Findings of the Association for Computational Linguistics: EMNLP 2024},
 publisher={Association for Computational Linguistics},
 year={2024},
 month={11}
}
```