CADS: A Systematic Literature Review on the Challenges of Abstractive Dialogue Summarization

Frederic Kirstein KIRSTEIN@GIPPLAB.ORG
Jan Philip Wahle WAHLE@UNI-GOETTINGEN.DE
Bela Gipp GIPP@UNI-GOETTINGEN.DE
Terry Ruas RUAS@UNI-GOETTINGEN.DE
Georg-August University Göttingen, Papendiek 14
Göttingen, 37073, Germany

Abstract

Abstractive dialogue summarization is the task of distilling conversations into informative and concise summaries. Although focused reviews have been conducted on this topic, there is a lack of comprehensive work that details the core challenges of dialogue summarization, unifies the differing understanding of the task, and aligns proposed techniques, datasets, and evaluation metrics with the challenges. This article summarizes the research on Transformer-based abstractive summarization for English dialogues by systematically reviewing 1262 unique research papers published between 2019 and 2024, relying on the Semantic Scholar and DBLP databases. We cover the main challenges present in dialog summarization (i.e., language, structure, comprehension, speaker, salience, and factuality) and link them to corresponding techniques such as graph-based approaches, additional training tasks, and planning strategies, which typically overly rely on BART-based encoder-decoder models. Recent advances in training methods have led to substantial improvements in language-related challenges. However, challenges such as comprehension, factuality, and salience remain difficult and present significant research opportunities. We further investigate how these approaches are typically analyzed, covering the datasets for the subdomains of dialogue (e.g., meeting, customer service, and medical), the established automatic metrics (e.g., ROUGE), and common human evaluation approaches for assigning scores and evaluating annotator agreement. We observe that only a few datasets (i.e., SAMSum, AMI, DialogSum) are widely used. Despite its limitations, the ROUGE metric is the most commonly used, while human evaluation, considered the gold standard, is frequently reported without sufficient detail on the inter-annotator agreement and annotation guidelines. Additionally, we discuss the possible implications of the recently explored large language models and conclude that our described challenge taxonomy remains relevant despite a potential shift in relevance and difficulty.

1. Introduction

Abstractive dialogue summarization, a task within Natural Language Processing (NLP) and text summarization, entails condensing key information from conversations into succinct and coherent summaries (Xu et al., 2022). This sub-field of text summarization is gaining prominence and is relevant for various real-world scenarios, including customer service (e.g., social media, Feigenblat et al., 2021, and e-commerce, Lin et al., 2022), healthcare (Nair et al., 2023), daily life (Chen et al., 2021), meetings (Manuvinakurike et al., 2021), and open-domain conversations (e.g., online-chat, Gliwa et al., 2019). The relevance of this task,

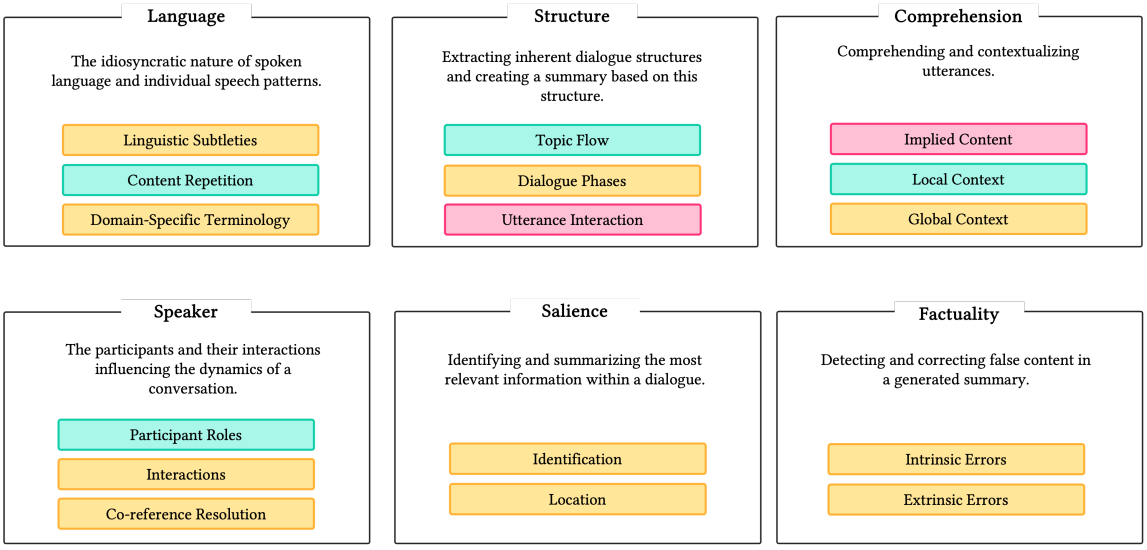

Figure 1: Overview of the six challenges in dialogue summarization, including a brief description of each challenge and an estimation of progress for related sub-challenges. Progress is evaluated based on two factors: (1) the extent to which mitigation strategies have been developed, and (2) the measurable improvement in summary quality as a result of these strategies. Green means mostly mitigated, orange means good progress, and red stands for marked challenges still exist.

which offers insights into discussed concerns or issues, is thereby driven by the increase in digital conversations in these scenarios (Jones et al., 2004). While the common, manual approach to dialogue summarization is considered to be prone to human errors and consumes significant time and effort, automatic dialogue summarization may effectively reduce this overhead on the user side (Mane et al., 2024).

Diverging from traditional text summarization, which typically focuses on formal and linear content such as news articles (Ravaut et al., 2022) or scientific publications (Altmami & Menai, 2022), dialogue summarization demonstrates unique challenges (Kryscinski et al., 2019): conversational text is inherently dynamic, interactive, and non-linear (Sacks et al., 1974), marked by verbosity, repetition, and informality. Salient information is often distributed across multiple speakers, mixed with casual and off-topic remarks (Jia et al., 2022b). Additionally, the use of elliptical and fragmented sentences, characterized by, e.g., incomplete thoughts and context-specific abbreviations, further hinders the summarization process. These complexities make traditional text summarization approaches less effective, as they are not trained to bridge the gap between edited text and dialogue and struggle to transfer learned patterns through few-shot learning (Feng et al., 2022).

Despite advancements achieved by leveraging and adapting general-purpose pre-trained neural language models (Lewis et al., 2020; Raffel et al., 2020), and the recent trend to explore Large Language Models (LLMs) for this task (Laskar et al., 2023), there remains a notable gap in the field's foundational understanding of the challenges when processing conversation transcripts. While research papers introduce techniques addressing similar challenges, their definitions and interpretations of these challenges can vary markedly. For

instance, while some describe the informal and ungrammatical nature of spoken language as a language challenge (Feng et al., 2022; Rennard et al., 2023), others frame the language challenge as different individual styles and unstructured expressions (Lee et al., 2021d; Fang et al., 2022). Both viewpoints are thereby valid but, on their own, focus just on a subpart of the language challenge and blur the complete understanding of it. Current surveys leave out the challenges in dialogue summarization and their variations in definitions, focusing instead on details on datasets, metrics, and techniques. Consequently, there is no clear understanding of which aspects of dialogue summarization are well-addressed and where significant research potential lies. Our systematic review consolidates the knowledge researched in the field considering Transformer-based models, aiming to provide a comprehensive overview of the challenges and their progress. We further consider the interest in LLMs and discuss their impact on the relevance of the dialogue summarization challenges.

This article focuses on four key areas, namely challenges, dataset overviews, techniques, and evaluation methods. In *challenges*, we categorize known hurdles for Transformer-based models found when dealing with dialogue transcripts into six broader challenge blocks, i.e., language, structure, comprehension, speaker, salience, and factuality. Our proposed taxonomy is displayed in Figure 1, including the six challenge blocks, a short description, and exemplary sub-challenges of the challenge blocks. In *datasets overview*, our review displays used datasets focused on dialogue summarization, organized by their subdomain, and further an overview of techniques to generate datasets to cope with the existing data scarcity artificially and how to optimize data usage. In *techniques of dialogue summarization*, we then match 97 techniques proposed since 2019 to the six challenge blocks. Furthermore, in *evaluation methods*, we provide an overview of typical evaluation metrics used, spanning count-based metrics (e.g., ROUGE, Lin, 2004), model-based (e.g., BERTScore, Zhang et al., 2020), QA-based (e.g., QuestEval, Scialom et al., 2021), and human evaluation metrics for performance and annotator agreement. This systematic review contributes to a better understanding of the dialogue summarization problem, unifies the inherent definitions, overviews established techniques, and demonstrates the scarcity of datasets and fitting evaluation metrics.

Organization of the Literature Review. The remainder of this review is structured as follows. We introduce the current state of dialogue summarization in Section 1.1 and show our methodology in Section 2. In Section 3, we provide the general problem definition and our challenges taxonomy consisting of six challenge blocks inherent in conversation scripts, further linking proposed approaches to handle the challenges. Section 4 overviews prominent datasets, while the display of evaluation approaches in Section 5 helps readers choose suitable indices to evaluate the effectiveness of a model. Finally, Section 6 discusses future research directions followed by conclusions in Section 7. All resources for our review are publicly available.[1]

1.1 Existing Reviews in Dialogue Summarization

Comprehensive literature reviews on dialogue summarization remain scarce, with notable contributions by Feng et al. (2022) and Jia et al. (2022b). Existing surveys mainly explore the subdomains, their datasets, techniques, and metrics within dialogue summariza-

1. https://github.com/FKIRSTE/LitRev-DialogueSum

tion. They provide a high-level overview and categorization of prominent subdomains like meeting, email, chat, medical, and customer service but often neglect niche areas such as e-commerce and debates (Feng et al., 2022). On the datasets, Tuggener et al. (2021) provide a framework linking linguistic dialogue types (Walton & Krabbe, 1995) (e.g., persuasion, negotiation, inquiry) with established datasets, offering insights into dataset suitability. Gu et al. (2022) analyze dialogue summarization techniques up to early 2022, categorizing techniques by conversational context and component modeling (e.g., speaker, addressee, and utterance modeling to solve the 'Who says what to whom' paradigm), covering a subset of the conversation challenges indirectly. Jia et al. (2022b) extend this analysis with techniques from 2023, exploring approaches like hierarchical models for long inputs (Zhu et al., 2020), feature injection for enhanced understanding, auxiliary tasks for broader training objectives, and data augmentation from related tasks. Regarding evaluations, works on dialogue summarization leverage automatic metrics from traditional document summarization, with the study by Gao and Wan (2022) discussing their individual strengths and limitations.

While existing surveys provide an understanding of dialogue summarization regarding subdomains, datasets, techniques, and metrics, the field misses a link between these categories through a common framework, such as the challenges inherent in processing dialogues. Our review addresses this gap by introducing a comprehensive taxonomy of dialogue challenges and categorizing recent techniques until 2024 based on these challenges (Section 3). This approach highlights well-explored areas and points out gaps in current research. We analyze datasets (Section 4) and evaluation metrics (Section 5), linking them to the identified challenges. Additionally, we discuss the shift from smaller encoder-decoder models such as BART (Lewis et al., 2020) to LLMs and the implications for the relevance of the challenges with this new backbone model type (Section 6). Our review focuses exclusively on dialogue summarization and does not encompass broader areas such as general text summarization (R et al., 2023) or dialogue generation (Deng et al., 2023).

2. Methodology

We use the PRISMA checklist (Page et al., 2021) for organizing our literature review, as it is an established approach to set up a comprehensive systematic literature review while reducing the potential for incomplete data and biases in content selection and presentation (Fagan, 2017). This checklist ensures that our approach is structured, transparent, and reproducible.

Our methodology comprises two stages, as detailed in Figure 2.

Retrieval Stage. We retrieve literature automatically and keyword-based from two established academic databases, i.e., Semantic Scholar and DBLP, using specific retrieval criteria and defining keyword queries for systematic searching. The retrieved works are then automatically de-duplicated and saved to a list.

Manual Stage. The second stage involves manually screening full documents using strict inclusion and exclusion criteria. In this stage, we follow additional best practices for manual filtering beyond those in the PRISMA checklist. These include tracking specific items such as problem category, challenges discussed, and proposed technique category (Foltýnek et al., 2019; Spinde et al., 2024) and guidelines for human annotators (Park & Storey, 2023; Ibrahim & Shafiq, 2023).

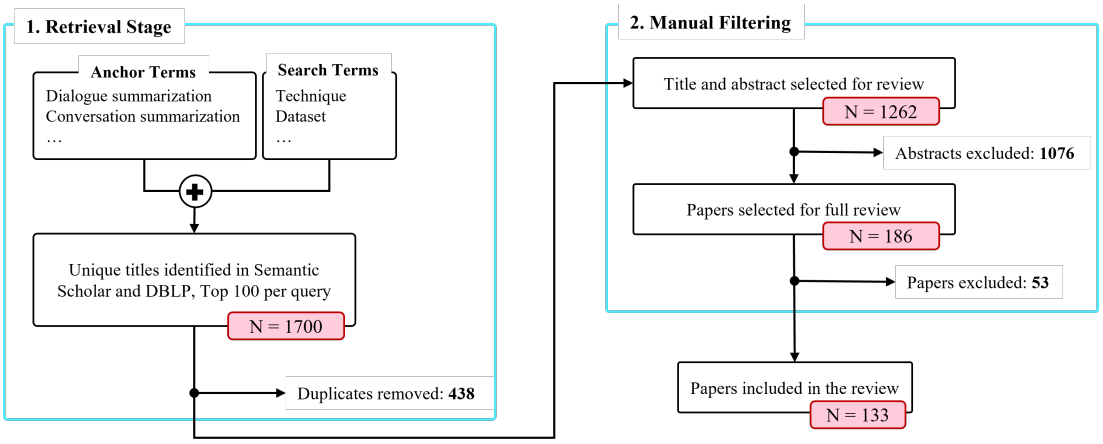

Figure 2: An overview of the different stages of our methodology and their inherent sub-stages. The red boxes indicate the number of papers considered in the respective step.

We evaluate 1262 unique papers from 2019 until 2024, offering an up-to-date perspective on the state of the art in dialogue summarization using Transformer-based architectures. We decided against including references from selected papers that did not appear in our crawl as additional sources to maintain an unbiased selection process and avoid complicating the stopping criterion for literature inclusion. Considering our exclusion criteria, such as multilingualism, non-abstractive, and non-Transformer-based approaches, we select 133 papers for our literature review.

2.1 Stage 1: Retrieving Candidate Documents

Information Sources. Our primary data sources are Semantic Scholar[2] and DBLP[3] (DataBase systems and Logic Programming), both recognized for extensive coverage in computer science (Kitchenham, 2004; Brereton et al., 2007). While Semantic Scholar allows for advanced search functionalities and complex queries (Xiong et al., 2018; Hannousse, 2021; Wang & Yu, 2021), DBLP is the most comprehensive database for computer science publications, frequently used in other surveys as sole source (Nguyen et al., 2021; Dong et al., 2022; Zhou et al., 2022) and encompassing major peer-reviewed journals and conference proceedings. This dual-source strategy ensures thorough retrieval and coverage specific to dialogue summarization, strengthening the certainty and reliability of our evidence base. To ensure our literature review remains up-to-date, we regularly re-crawl databases while conducting the review to incorporate the latest publications into our analysis.[4] While we acknowledge the potential for these databases to miss the most recent uploads, our strategy minimizes this risk. We exclude databases such as Google Scholar, which overly returned gray literature, and Web of Science, which emphasized extractive summarization techniques during our pre-testing, to maintain a focus on peer-reviewed literature in abstractive dialogue summarization. We design an automatic pipeline to retrieve query-

2. https://www.semanticscholar.org/
3. https://dblp.org/
4. Last crawled on March 25, 2024.

related works and remove duplicates. We limit the number of results per query to the top 100 based on the relevance ranking provided by the platforms, which was rarely required. We recognize that the ranking systems may be biased towards prestigious journals or highly cited authors. Hence, we retrieve all papers published after June 2023, not relying on the ranking systems to account for their limited time to gather citations.

Data Collection Process. Our search strategy uses key terms related to dialogue summarization identified during our initial related work analysis (Section 1.1). Anchor terms such as 'dialogue summarization' and 'multi-party conversation' are combined with specific search terms such as 'technique.' We use four anchor terms and 28 search terms, resulting in 112 search queries for each database. We present the complete list and search keywords in our repository[5]. Employing a Python-based pipeline (Section 9), we systematically extract documents from Semantic Scholar and DBLP APIs, merging and unifying the results into tabular data. Between January 2019 and March 2024, we retrieved 732 publications from DBLP and 968 from Semantic Scholar. After removing 438 duplicates identified across both sources, we conclude the stage with 1262 unique publications. These results are tagged with their respective queries and compiled into a CSV file for subsequent selection and analysis.

2.2 Stage 2: Manual Filtering

Eligibility Criteria. For our manual filtering process, we define eligibility criteria to ensure the focus of our literature review. We consider open-access papers published since 2019 that align with the rise of Transformer-based text summarization models such as BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020a), which have been the core model for most approaches in the field. This timeframe further allows us to concentrate on current challenges rather than including some that may have lost relevance (e.g., negation, Khalifa et al., 2021) because of advancements in the field. We exclude non-English primary datasets, multi-modal studies, and extractive or non-Transformer-based methods and focus on English-language research on abstractive dialogue summarization.

Manual Paper Processing. A team of three reviewers[6], consisting of postgraduate students and doctoral candidates with a computer science background, undertake the manual screening process, starting with titles and abstracts to narrow down the automatic selection from 1262 to 186 documents using our eligibility criteria. They then review the full texts, checking if the works still meet the eligibility criteria, and decide whether to include, exclude, or discuss each paper. Each paper undergoes a secondary review to minimize bias, reaffirming initial selections and categorizations, thereby enhancing the study's integrity. Eventual disagreements are resolved through consensus. The annotators also track features such as used datasets, employed backbone models, proposed techniques, and evaluation approaches. This step filters out 53 papers, leaving 133 to compose the literature review and detail challenges, techniques, datasets, and evaluation metrics.

5. https://github.com/FKIRSTE/LitRev-DialogueSum
6. Each team member discloses potential conflicts of interest to ensure an unbiased review process.

3. Challenges and Progress

Abstractive dialogue summarization aims to condense multi-party conversations into their key points, ranging from casual chats to expert discussions, and whether these points were explicitly stated or inferred (Cohan et al., 2018). The task was first thoroughly defined in 2002 by Zechner, outlining challenges unique to spoken dialogues such as handling speech disfluency, the lack of clear sentence boundaries, distribution of salient information across various speakers and turns, resolving references, identifying discourse structures, and dealing with inaccuracies caused by speech recognition errors. While some definitions take a mathematical approach (Jia et al., 2022b), viewing dialogues as sequences of turns that are compressed and associated with a specific speaker, other definitions adopt a broader description of the task as 'extracting useful information from a dialogue' (Liu et al., 2019). Although all definitions contain relevant, overlapping aspects, they have different viewpoints, each considering a different set of challenges to define dialogue summarization. Our goal is to organize these definitions and provide a concise, structured overview of the progress in abstractive dialogue summarization, extending the earlier challenge-based concept (Zechner, 2002) and proposing the Challenges of Abstractive Dialogue Summarization (CADS) Taxonomy (Figure 1) tailored to the current Transformer-based model architecture. This taxonomy comprises six challenges: language, structure, comprehension, speaker, salience, and factuality. The first five challenges are related to the input, while the last concerns reliability. To develop this comprehensive framework, we employed human annotators (Section 2.2) to systematically analyze and categorize recurring challenges in the literature and consolidate them into the six challenges. Each pillar encompasses related sub-challenges that represent key focus areas consistently addressed in the literature along the broader challenges. In the following Sections 3.1 to 3.6, we detail the CADS Taxonomy as shown in Figure 1. An overview of the approach strategies proposed to mitigate the individual challenges is shown in Table 1. Table 2 summarizes the expected errors when challenges are not successfully mitigated, using the analysis by Kirstein et al. (2024d) as a base for the individual error types.

For this section, we consider 91 papers, from which 73 papers define the challenge characteristics (sub-challenges in *italic*) and 74 describe explored techniques. We exclude papers mainly discussing evaluation strategies and datasets from this subset and cover them in Sections 4 and 5.

3.1 Language

Characteristics. The language challenge describes the idiosyncratic nature of spoken language and individual speech patterns. It includes *linguistic subtleties* such as informal expressions (e.g., 'yeah,'), ungrammatical structures, colloquialisms, personal vocabulary, and linguistic noise such as filler words (Koay et al., 2020; Zhang et al., 2021b; Feng et al., 2022; Kumar & Kabiri, 2022; Antony et al., 2023). It further covers *content repetition*, i.e., speakers restate or rephrase information for emphasis or clarity (Chen & Yang, 2020; Khalifa et al., 2021; Lei et al., 2021b) and *domain-specific terminology*, e.g., medical terms (Bertsch et al., 2022; Liu et al., 2022b, 2022a; Li, 2022). These elements demand that models accurately interpret and adapt to the linguistic features of dialogues (Jia et al., 2022a). Figure 3 presents a dialogue snippet showing examples of these three sub-challenges.

Challenge	Approach Category	Related Papers
Language	Pre-training	Raffel et al. (2020), Zou et al. (2021), Zhou et al. (2023a), Lyu et al. (2024b)
	Training tasks	Zhu et al. (2020), Khalifa et al. (2021), Lee et al. (2021d), Jia et al. (2022a), Bertsch et al. (2022)
	Pre-processing	Ganesh and Dingliwal (2019)
Structure	Pre-training	Lee et al. (2021c), Peysakhovich and Lerer (2023), Xu et al. (2024)
	Training tasks	Feng et al. (2021), Lee et al. (2021b), Liu et al. (2021), Yang et al. (2022)
	Architecture modification	Li (2022), Lei et al. (2021b), Gao et al. (2023), Hua et al. (2023, 2022)
	Importance measures	Reimers and Gurevych (2019), Liang et al. (2023)
Comprehension	Architecture modification	Wang et al. (2023)
Speaker	Training tasks	Gan et al. (2021), Qi et al. (2021), Asi et al. (2022), Naraki et al. (2022)
	Architecture modification	Lei et al. (2021a, 2021b), Liu et al. (2021), Hua et al. (2022)
	Pre-processing	Joshi et al. (2020), Lee et al. (2021a)
	Post-processing	Fang et al. (2022), Liu and Chen (2022)
Salience	Pre-training	Pagnoni et al. (2023), Zhang et al. (2023)
	Training tasks	Chauhan et al. (2022), Liu et al. (2022a), Ghadimi and Beigy (2022)
	Architecture modification	Li et al. (2021), Hua et al. (2023)
	Human feedback	Chen et al. (2023)
	Loss function	Huang et al. (2023)
	Pre-processing	Liu and Chen (2021), Jung et al. (2023)
Factuality	Training tasks	Gan et al. (2021), Tang et al. (2022)
	Architecture modification	Wu et al. (2021), Zhao et al. (2021a, 2021b), Nair et al. (2023)
	Human feedback	Chen et al. (2023)
	Loss function	Liu et al. (2022a), Huang et al. (2023)
	Post-processing	Fu et al. (2021), Li et al. (2023)

Table 1: Overview of challenges, major approach trends, and corresponding literature. More details are stated in the article's accompanying repository: https://github.com/FKIRSTE/LitRev-DialogueSum.

Transcript of planning a hiking trip

Alex: Hey, you know, um, the hiking trip we were, like, talking about? Are we still, uh, on for next weekend?

Jordan: Oh, yeah, definitely! I was just looking into, uh, places to go. Y'know, I found this really cool, uh, trail. It's called, uh, Echo Ridge? Yeah, Echo Ridge. Super scenic, but it's kinda, you know, tough. Lots of, um, elevation changes and, uh, rocky paths.

Alex: Sounds, like, amazing but, uh, challenging. Do we need, like, special gear or something? I mean, with all those elevation changes and stuff?

Figure 3: Dialogue snippet showing examples of the idiosyncratic nature of spoken language and individual speech patterns. Red displays disfluencies, blue highlights personal speech patterns, orange stands for colloquialism, green represent informal expressions.

Failing to handle this challenge may result in a loss of coherence and clarity, repetition of content, and factuality issues.

Approaches. Transformer-based models often struggle with the language characteristics due to a gap between their pre-training data, typically well-edited texts (e.g., news articles, research papers, Wikipedia entries, Raffel et al., 2020), and the characteristics of spoken dialogue not reflected in these edited texts (Zou et al., 2021). The transfer between these text styles is hindered by the scarcity of diverse, dialogue-oriented pre-training data, resulting in a lack of exposure to dialogue data during training, consequently reducing models' performance. Repetition of already provided information is less of an issue for current language models (Khalifa et al., 2021).

Current research aims to bridge this gap between the formal language of the pre-training data and dialogues by retraining models with dialogue-focused tasks, considering both single- and multi-task setups. Explored tasks span masking key dialogue elements (e.g., pronouns entities, high-content tokens, and words, Khalifa et al., 2021) and part-of-speech tagging (Lee et al., 2021d). More recent tasks aim to transform informal, first-person dialogue into a structured, third-person narrative through changing speaker names, adjusting grammar, and adding emotional context (Bertsch et al., 2022), or simulating a dialogue by modifying pre-training documents into a conversation structure (Zhu et al., 2020). Teaching models to understand the 'who-did-what' structure through a specific task is further gaining traction (Jia et al., 2022a). Other explored strategies include pre-processing the input for anaphora resolution (Ganesh & Dingliwal, 2019) and adapted pre-training, e.g., to adjust to specific terminology (Zou et al., 2021; Zhou et al., 2023a; Lyu et al., 2024b).

3.2 Structure

Characteristics. Aiming to extract a conversation's built-up, the structure challenge is about segmenting a dialogue transcript and creating an ordered summary based on the extracted structures (Ma et al., 2023). This challenge involves identifying the *topic flow* of a conversation (Zhang & Zhao, 2021; Zhang et al., 2021b; Gu et al., 2022; Shinde et al., 2022), tracking the *dialogue phases* (e.g., problem identification, decision making, Tuggener et al., 2021; Li et al., 2023), and analyzing the *utterance dependency* (i.e., relationship and

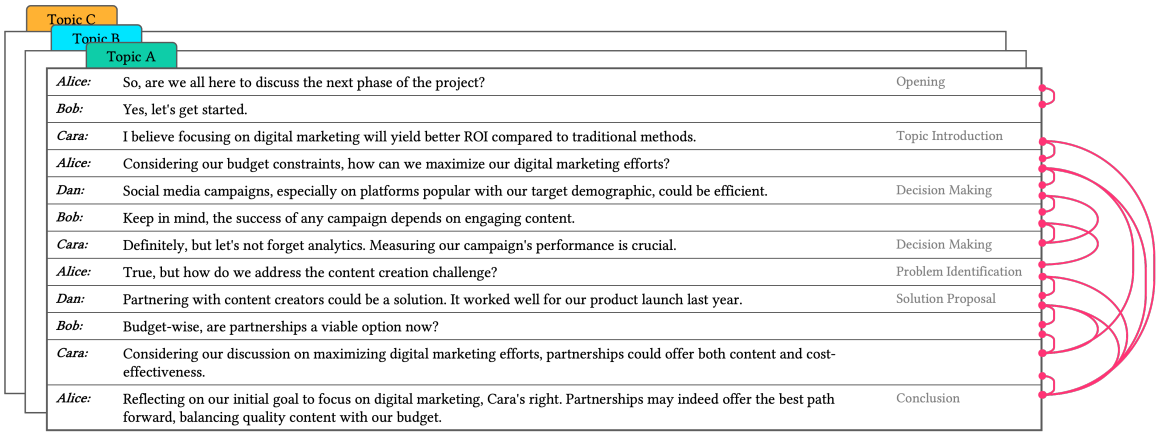

Figure 4: Excerpt from a conversation to demonstrate the structural challenge, with a detailed view on one topic out of multiple topics touched in the whole conversation. The red arcs mark the connection between individual turns, demonstrating how different dialogue phases (grey text), such as the conclusion, relate to many previous conversation phases.

dependencies between utterances, Feng et al., 2021a; Lei et al. 2021b; Hua et al., 2022). The dialogue in Figure 4 shows topic flow, dialogue phases, and utterance dependency. This challenge can lead to summaries lacking coherence, completeness, and depth (Lee et al., 2021a; Liu et al., 2021; Qi et al., 2021; Yang et al., 2022; Liang et al., 2023).

Approaches. Transformer-based models face two primary hurdles when handling the structural challenge. First, they must adapt to the variability in dialogue structures, such as differing topic sequences and conversation phases depending on the conversation type. Second, they need to track and connect long-range dependencies to link utterances. The first difficulty is related to the lack of varied datasets across different dialogue types on which models can train to improve generalization as well as that most pre-trained models used are expecting more precise structural signals due to the more structured data used during pre-training (Section 3.1). The second problem relates to a weakness of transformed-based architectures: their limitation in handling long sequences, which stems from the practical limitation of a fix-sized context window (Lee et al., 2021c), the blurred attention on a single token when many tokens are processed (Xu et al., 2024), and the bias towards recent information (Peysakhovich & Lerer, 2023).

Recent methods employ graph structures to understand and use dialogue structures more effectively. They combine static and dynamic graphs for a detailed examination of conversation dynamics (Gao et al., 2023) where the static graphs represent unchanging aspects like speaker relationships, and dynamic graphs track how dialogues evolve. Abstract Meaning Representation (AMR) graphs are employed to capture overarching themes (Hua et al., 2023), detailed sentence-level connections, and entity interactions, enhancing content comprehension (Hua et al., 2022). Meanwhile, a hybrid approach combining traditional language models with graph neural networks captures the utterance dependency and addresses the long-range dependency sub-challenge (Li, 2022). Similarly, ConceptNet (Liu & Singh, 2004) is used for thematic linking (Lei et al., 2021b).

Transcript of a team meeting

Alice:	Let's start with the updates. Bob, how's the backend going?
Bob:	I've integrated the new database schema, but it's still slow. I'm considering Redis for caching.
Charlie:	On the design front, I'm waiting for the final user flow to adjust the UI.
Alice:	Great, we need to speed up for the upcoming demo. Any blockers?
Bob:	Redis might help, but I need to sync again with Charlie on the data requirements.

Figure 5: Example conversation to demonstrate the comprehension challenge including direct and implied content. Directly stated (orange) are the explicit updates by Bob and Charlie about their respective tasks. Green marks implied content such as a deadline for a demo and a current focus on performance and finalizing features (unmentioned background knowledge), Redis implies an understanding of its role in performance enhancement (organizational knowledge and reference to prior discussions on data requirements).

To manage long dialogue transcripts beyond context-size limits and the inherent long-distance relations, researchers are moving away from simply truncating inputs to context length to avoid losing vital information. An established approach segments the summarization process into stages (e.g., topics and dialogue phases) for coherent processing (Laskar et al., 2023; Asthana et al., 2024; Mullick et al., 2024). The segmentation and multi-pass strategy iteratively summarizes parts of the dialogue and combines partial summaries with a subset of the remaining content in each iteration to build a complete summary (Sharma et al., 2023). The $Summ^N$ framework (Zhang et al., 2021a) summarizes dialogue segments independently before refining them together, offering a scalable solution.

Other techniques involve additional training tasks (Feng et al., 2021; Lee et al., 2021b; Liu et al., 2021; Yang et al., 2022), sentence importance measures to group sentences into sub-topics (Reimers & Gurevych, 2019; Liang et al., 2023), while hierarchical setups to differentiate dialogue parts based on who is speaking (word level) and what is being discussed (turn level) are also popular (Zhu et al., 2020).

3.3 Comprehension

Characteristics. The comprehension challenge involves accurately understanding and contextualizing utterances such that models can grasp and leverage these insights for summarization (Khalifa et al., 2021). This includes *direct* and *implied content*, e.g., unmentioned background knowledge, organizational knowledge, and prior discussions not explicitly referred to (examples in Figure 5). To infer and understand this content, the whole context of a dialogue is required, spanning both the *local context*, focusing on neighboring words and sentences, and the *global context*, covering the entire dialogue's informational flow (Zhang et al., 2019). Insufficient comprehension can lead to misunderstanding the content, omitting parts, or falsely presenting the information.

Approaches. Transformer-based models struggle with comprehension and contextualization due to the underlying dependency parsing challenges, particularly with unclear sentence boundaries as they appear in spoken language, which can lead to catastrophic forgetting

and result in unreliability in handling long-range dependencies (Lee et al., 2021c). Additionally, these models interpret the text literally, affecting their ability to grasp nuanced meanings (Rai & Chakraverty, 2020; Wan et al., 2021). As these models often rely on syntactic relationships between words, they miss the nuanced understanding humans use to interpret implied meanings of sarcasm, irony, euphemisms, and similar pragmatic elements (Fazly et al., 2009). Consequently, this can lead to summaries that miss deeper meanings or implications, resulting in content that may be shallow or misleading (Wang et al., 2023).

Despite its importance, our review reveals limited engagement from the dialogue summarization community to improve the contextualizing capability of models considering both the short and long context. Current efforts primarily rely on Transformer models' self-attention capabilities to comprehend local and global contexts (Beltagy et al., 2020). Despite this reliance, the context-aware extract-generate framework (Wang et al., 2023) has been proposed to reduce the risk of missing long-range contexts in complex dialogues, which uses context prompts to capture the immediate details and the overall narrative of text spans. It first pinpoints relevant text parts that guide the extraction of key information, which is then used as a prompt base for generating detailed summaries.

3.4 Speaker

Characteristics. Participants, their interactions, and how they influence the dynamics of a conversation are part of the speaker challenge. This includes the individual *participant roles* (e.g., 'project manager,' 'customer service agent', Lei et al., 2021b; Hua et al., 2022), topic-dependent role changes (Khalifa et al., 2021; Feng et al., 2022a; Gu et al., 2022; Kumar and Kabiri, 2022; Rennard et al., 2022; Shinde et al., 2022; Antony et al., 2023), their *interactions* (e.g., contributions depending on viewpoints, Beckage et al., 2021; Li, 2022; Xie et al., 2022; and on backgrounds, Lee et al., 2021; Zou et al., 2021; Geng et al., 2022), and *references to entities and other participants* (Lee et al., 2021d; Liu & Chen, 2021, 2022; Naraki et al., 2022). Examples of the speaker characteristics are shown in Figure 6. Failing to capture these speaker characteristics may lead to missing context, misaligned entity associations, and, ultimately, incorrect factual summaries. The significance of examining speaker dynamics and roles is well-recognized in other fields such as linguistics (Meskill, 1993) and is, together with the language challenge of the preceded analysis (Section 3.1), a typical hook to demonstrate and motivate techniques for dialogue summarization.

Approaches. The speaker challenge stems from identifying participants, tracking their actions and entity references, and revealing participant relationships. These aspects correspond to the difficulties of named entity recognition (NER), coreference resolution, and dependency parsing. Despite advancements with Transformer-based models regarding NER, current techniques struggle with entity identification and categorization, particularly in unstructured texts and across diverse domains (Vajjala & Balasubramaniam, 2022; Wang et al., 2022c; Pakhale, 2023). This limitation is especially pronounced in realistic dialogue scenarios. Coreference resolution poses difficulties for techniques in accurately identifying mentions and contextual links (Quan et al., 2019) and understanding the linguistic structures (Ioannides et al., 2023), extending the understanding challenge (Section 3.3).

Works typically use dense vectors (Asi et al., 2022; Naraki et al., 2022) to accurately represent speaker roles in dialogue and act as labels for utterances (Gan et al., 2021; Qi

Transcript of a customer service discourse

Jordan:	Thank you for calling TechSupport, how can I assist you today?
Alex:	My device keeps freezing ever since the last update. I've tried rebooting, but it didn't help.
Jordan:	I understand the frustration. Let me pull up your account. Meanwhile, have you tried a factory reset?
Alex:	No, I'm worried about losing my data.
Jordan:	That's a valid concern. Let's avoid a reset for now. I think we need a specialist. Taylor, I have a customer with a post-update freezing issue. We've ruled out simple fixes.
Taylor:	Got it. Let's check the device's error logs together. Could you navigate to the settings and tell me what error codes you see?
Alex:	Sure, there's a series of codes here: E-101, E-102.
Taylor:	Those codes indicate a firmware glitch. We'll need to apply a manual patch. Jordan, could you guide Alex through the consent process?
Jordan:	Absolutely, Taylor. Alex, I'll need your consent to proceed with the patch. This will not affect your data.

Figure 6: Example conversation of the speaker challenge with defined roles (Alex as customer, Jordan as customer support agent, Taylor as technical specialist). The dialogue shows a change of roles when Jordan calls Taylor (marked with orange), shifting from Jordan as a primary troubleshooter to a facilitator for Taylor's technical interventions. Further, multiple direct references to entities highlighted with blue (e.g., 'device', 'error codes') and indirect references to participants in red (e.g., 'you') are incorporated.

et al., 2021), offering insights into each speaker's function within the dialogue. Speaker vectors are either randomly initialized and trainable (Zhu et al., 2020; Qi et al., 2021) or produced by small neural networks (Gan et al., 2021; Naraki et al., 2022) and represent different roles, e.g., 'industrial designer' (Zhu et al., 2020), 'judge' (Duan et al., 2019), which are appended to the embedding of the speaker's turn. For capturing speaker dynamics and resolving coreferences, graph-based models are an established approach, e.g., representing each speaker's main ideas and their discourse, alongside their interactions, reflecting inner- and inter-speaker structures (Lei et al., 2021b; Hua et al., 2022).

Some other popular techniques involve enhancing Transformers' self-attention to analyze intra- and inter-speaker dynamics (Lei et al., 2021a; Liu et al., 2021), additional training tasks (Lee et al., 2021b; Geng et al., 2022; Zhou, 2023), and pre-and post-processing techniques to replace pronouns and correct mistakes (Joshi et al., 2020; Lee et al., 2021a; Fang et al., 2022; Liu & Chen, 2022) for coreference resolution.

3.5 Salience

Characteristics. The salience challenge involves identifying and summarizing the most relevant information within a dialogue (Chen & Yang, 2020). This includes *identifying* content critical for insightful summaries (Liu et al., 2021), where the understanding of critical shifted lately from content for general audience summaries to summaries aligning with the viewpoints, values, and ethics of specific participant (Liu & Chen, 2021), and

Transcript of a strategy team meeting

Simon:	Given our growth last quarter, it's time to think about adding a new product line.
Nadia:	Market analysis shows high demand for eco-friendly products. We should consider this for our new line.
Luca:	To support that, we'll need to vet new suppliers for sustainability. It could take some time.
Simon:	That's a worthwhile investment. Nadia, prepare a marketing strategy focusing on eco-friendliness.
Nadia:	Understood. I'll also look into partnership opportunities with eco-conscious influencers.
Luca:	I'll start researching potential suppliers and calculate the logistics costs.
Simon	Perfect. Let's aim to have initial plans ready by next month. This expansion could set us apart in the market.

Summary - Simon:

Today's discussion confirmed the strategic direction to expand our product line, focusing on eco-friendly products to capitalize on market demand. Nadia will spearhead the development of a marketing strategy to highlight our commitment to sustainability, while Luca takes on the task of vetting potential suppliers to ensure they meet our sustainability criteria. The aim is to differentiate our brand in the market through this initiative.

Summary - Nadia:

Following today's meeting, you are tasked with crafting a marketing strategy that leverages the growing demand for eco-friendly products. This includes identifying partnership opportunities with eco-conscious influencers to amplify our message. Your strategy will play a crucial role in positioning our new product line as a leader in sustainability.

Summary – General:

In the recent meeting, the company decided to pursue the addition of an eco-friendly product line, responding to significant market demand. Simon highlighted the importance of this strategic direction, emphasizing sustainability as a core value. Nadia is set to develop a comprehensive marketing strategy focused on eco-friendliness, including partnerships with influencers. Meanwhile, Luca is tasked with vetting suppliers for sustainability and calculating logistics costs. The team is aligned on having initial plans by next month, marking a significant step towards differentiating our brand in the market.

Figure 7: Example of the salience challenge with a conversation being summarized for a general audience and tailored summaries for participants with a specific focus on content relevant to them.

locating text spans containing relevant content (Lei et al., 2021b). Figure 7 demonstrates the differences between personalized and generally applicable summaries, as the generated texts for Simon and Nadia are tailored to their role, stands, and tasks, whereas the general summary provides a neutral overview. In contrast to traditional summarization, where salience may be related to the position of sentences (positional salience), dialogues are more dynamic, and salience emerges from how ideas evolve over the conversation. Failure to effectively locate the scattered information or understand what is considered salient can impact the depth and relevance of the generated summaries, focusing on unimportant content while omitting relevant information bits (Chen & Yang, 2020; Zhang et al., 2021b; Zou et al., 2022; Rennard et al., 2023).

Transcript of scheduling a team meeting

John:	Are we still on for the team meeting tomorrow? I think we need to finalize the budget report.
Viktor:	Yes, the meeting is set for 10 AM. I agree, the budget report needs our immediate attention, especially with the upcoming audit. Also, remind me to discuss the new marketing strategy; I've gathered some interesting insights.
John:	Will do. And I've noticed some discrepancies in the budget projections we might want to address.
Viktor:	Good catch. Let's delve into that tomorrow as well.

Summary:

The team meeting to discuss the budget report and finalize the marketing strategy was successfully held yesterday, with key decisions made on addressing discrepancies in budget projections. Additionally, concerns about an upcoming audit led to a detailed review of financial strategies, ensuring all angles were covered for a robust approach.

Figure 8: Example of the factuality challenge where the generated summary exhibits extrinsic errors through hallucinating events (blue background) such as the finalization of a marketing strategy and a detailed review of financial strategies. The intrinsic error reporting the wrong date of the meeting is marked with red.

Approaches. The first hurdle for current models to the salience challenge is the request for subjective salience and ranking content's importance to align with the role and knowledge of the summary addressee. This is difficult as there is no training task for inferring personal attributes and ranking information according to them. Further, locating this content is difficult as salient information is typically scattered across multiple turns within the idiosyncrasies of spoken language (Lei et al., 2021b; Feng et al., 2022; Tan et al., 2023), which forms an extension of the language challenge (Section 3.1).

To nest the general understanding of what is considered salient into models, established strategies include additional training stages to better distinguish between salient and non-salient content. Huang et al. (2023) introduce the uncovered loss to point out if salient information is missing and the contrastive loss to distinguish between relevant and irrelevant sentences, respectively. Negative samples, categorized as redundant (i.e., text with unnecessary utterances) and null (i.e., text with relevant utterances removed), help models prioritize important content (Liu et al., 2022a, 2022b). To better determine salient content for specific participants, works propose a fine-tuning task in which the topic or user perspective to be focused on is passed as an input along the dialogue text (Chauhan et al., 2022), apply question-driven pre-training (Pagnoni et al., 2023) and use dynamic prompts to direct model attention to key dialogue aspects (Zhang et al., 2023). Additional strategies involve direct text manipulation (Jung et al., 2023), personal named entity planning to concentrate on specific entities (Liu & Chen, 2021), adapting attention mechanisms (Li et al., 2021), and integrating human feedback (Chen et al., 2023) or graph structures (Hua et al., 2023) to further tailor summaries to the dialogue's core information.

Challenge	Related Errors
Language	incoherence, repetition, hallucination
Structure	incoherence, omission, lack of depth
Comprehension	lack of context understanding, omission, hallucination
Speaker	lack of context understanding, false coreference resolution, hallucination
Salience	irrelevance, lack of context understanding
Factuality	hallucination

Table 2: Definition of eight to-be-expected error types in dialogue summarization, based on existing meeting summarization related error types (Kirstein et al., 2024d).

3.6 Factuality

In contrast to the previous input-related challenges, the factuality challenge describes the problem of correcting false content in a generated summary caused by a non-robust architecture. Factual incorrect content contains *extrinsic errors*, such as hallucinating events not present in the original transcript, and *intrinsic errors*, such as incorrect coreference resolution misrepresenting actual event details, and wrong conclusions stemming from negation (Wang et al., 2022). Extrinsic and intrinsic errors are shown in a sample dialogue in Figure 8. Given the importance of ensuring summaries accurately reflect the conversation, research focuses on creating safety measures to maintain credibility.

Approaches. Research interest predominantly targets post-processing approaches that rewrite the generated summary to correct factual errors. Notable approaches are the usage of LLMs for error detection and correcting (Li et al., 2023) and reformulating the summary to closely match the original dialogue's predictive capabilities for subsequent content (Fu et al., 2021). Other approaches explore the use of auxiliary tasks (Tang et al., 2022) to estimate the factual aspects and predict the missing aspects in summary (Gan et al., 2021), losses such as encouraging a model to generate sentences about content not yet covered and differentiating factual from non-factual sentences (Liu et al., 2022a; Huang et al., 2023), and human feedback (Chen et al., 2023). Further, architecture-modifying approaches are explored, such as a new encoder to grasp dialogue states (Zhao et al., 2021b) or to handle graph structures, where dialogue events are captured and organized in a graph. A slot-driven beam search algorithm is used in a filling-in-the-blanks setup to give priority to generating salient elements in the summary (Zhao et al., 2021a), and a hierarchical approach builds on sub-summaries (Nair et al., 2023). Another noteworthy technique is the generation of a sketch to structure the final summary along this plan (Wu et al., 2021). Negation remains a long-standing challenge for language models often neglected in computational studies (Hossain et al., 2020; Zhang & Zhao, 2021).

4. Datasets

In this review, we collected data generation methods and downstream datasets from the selected publications. We only consider datasets used at least five times in our analysis to avoid including resources with little community impact. This setup yields 18 datasets that we group into six categories based on the dialogue subdomains they stem from: daily chat,

online chat, meeting, screenplay, customer service, medical, and debates. We also create an 'others' category for datasets that do not fit neatly into these subdomain groups.

These datasets typically condense a short dialogue of about 1k tokens between natural persons into a third-person summary of about 100 tokens. More complex scenarios, such as meetings, parliamentary debates, and TV series dialogues, involve around 10k to 20k tokens in the transcript and at least four participants. Due to privacy concerns, a significant portion of currently publicly available datasets are reenactments based on actual events, e.g., by Amazon Mechanical Turk[7] workers, or established datasets such as SAMSum (Gliwa et al., 2019) are modified to generate new data, e.g., by changing the dialogue transcript to a third-person report (Bertsch et al., 2022). This underscores that the variety of existing datasets is comparably limited, and these datasets may not capture real-life scenarios. The 18 available established datasets are detailed in Sections 4.1 to 4.8. Given the limited number of available datasets, we present methods for creating artificial datasets and strategies for optimizing the use of existing datasets in Section 4.9.

For this section, we consider 93 papers to identify the established datasets, excluding works that do not mention a dataset, datasets that are no longer publicly available, and datasets in a language other than English. We did not assess dataset quality or perform a detailed analysis of the inherent challenges, leaving that for future research. In Table 3, we connect the datasets with their respective primary challenge, as noted by their creators, to provide an overview of the distribution of challenges in datasets.

4.1 Daily Chat

DialogSum (Chen et al., 2021) integrates dialogues from DailyDialog (Li et al., 2017), Dream (Sun et al., 2019), Mutual (Cui et al., 2020), and English-speaking practice websites, featuring two-speaker interactions across daily-life scenarios like work, leisure, and travel. Annotators were given guidelines on writing summaries, such as length constraints (no longer than 20% of conversation length). DialogSum includes 13k dialogues, with inputs of around 1k tokens and summaries of 130 tokens.

4.2 Online Chat

SAMSum (Gliwa et al., 2019) comprises 16k written online dialogues from messaging apps, which linguists craft asked to create conversations similar to those they write daily, reflecting the proportion of topics of their real-life messenger conversations. The messages, each written by one person, contain chit-chats, gossip, arranging meetings, discussing politics, and consulting university assignments. They typically involve two speakers, with an average of 94 tokens per conversation, varying between three and 30 turns[8]. A subset of SAMSum has been adapted into formal third-person language (Bertsch et al., 2022) to help models transition from informal dialogue to edited text, addressing the linguistic gap between pre-training and downstream task (detailed in the language challenge, Section 3.1).

Forum (Bhatia et al., 2014) consists of random samples of 100 threads from the online discussion forums on Ubuntu and TripAdvisor, with a total of 556 and 916 posts,

7. https://www.mturk.com/
8. A turn is a contribution made by a speaker in the form of a single utterance or a statement.

respectively. Two human evaluators created summaries of the discussion threads, resulting in two human-written summaries per sample.

CRD3 (Rameshkumar & Bailey, 2020) consists of dialogues from the 'Critical Role Dungeon and Dragon' show with summaries collected from the Fandom wiki, featuring 159 episodes with an average dialogue length of 2550 turns.

4.3 Meeting

AMI (Carletta et al., 2006) contains business meetings on the product design process, detailing 137 staged meetings on designing a new remote control. Participants are project managers, marketing experts, user interface, and industrial designers. The dataset includes transcripts and human summaries, with dialogues averaging 6k tokens over 535 turns and four speakers. A modified version, AMI-ITS (Manuvinakurike et al., 2021), offers additional annotations for incremental temporal summaries, which provide summaries for 100-second time durations on a subset of AMI.

ICSI (Janin et al., 2003) has 59 academic group meetings with computer scientists, linguists, and engineers at the International Computer Science Institute (ICSI) in Berkeley, along with their summaries written by hired annotators. The meetings have an average of 819 turns and 13k tokens with six speakers and are research discussions among students.

QMSUM (Zhong et al., 2021) introduces query-based summarization across diverse meeting domains, compiling 1.8k query-summary pairs from 232 meetings, encompassing product design (AMI), academic discussions (ICSI), and committee deliberations. The dialogues feature up to 13k tokens and six speakers. The original task is to summarize the meeting given a stated question. MACSUM-DIAL (Zhang et al., 2023) is a modification of QMSUM, designed for controllable summarization, highlighting mixed attributes such as length, attractiveness, and topic specificity.

ELITR (Nedoluzhko et al., 2022) features 120 English technical project meetings in computer science, with each transcript averaging 7k words, 730 turns, and six speakers. The duration of the meetings varies from ten minutes to more than two hours, with an average of one hour long. Meetings shorter than half an hour are exceptions, whereas meetings longer than two hours are topic-oriented mini-workshops, also rather occasional. The original task of the dataset differs from abstractive summarization as a model is required to produce not an abstractive summary but a set of meeting minutes in bullet points.

MEETINGBANK (Hu et al., 2023) contains meetings of city councils from six major U.S. cities occurring over the past decade. It contains 1366 meetings spanning 3.6k hours, with a council meeting lasting an average of 2.6 hours and the transcript containing 28k tokens.

4.4 Screenplay

MEDIASUM (Zhu et al., 2021) contains 463.6k media interview transcripts with summaries from National Public Radio (NPR, Majumder et al., 2020) and CNN, spanning a range of topics, including politics, news, crime, and economy. The summaries are based on NPR's interview overviews and CNN's topic descriptions, with the latter segmented at commercial breaks to match topics to corresponding interview segments. On average, each transcript in this dataset contains 1.5k words, with summaries of around 11 words, typically involving seven speakers.

SummScreen (Chen et al., 2022) consists of 27k instances of TV series transcripts paired with human-written recaps sourced from TV MegaSite and ForeverDreaming. The recaps of ForeverDreaming are based on community contributions stemming from Wikipedia and TVmaze. Transcripts, typically including 28 speakers, average 6.6k tokens, and summaries around 380 tokens.

4.5 Customer Service

TweetSumm (Feigenblat et al., 2021) contains 1.1k dialogues derived from Twitter customer support exchanges, each with three extractive and three abstractive human-written summaries. Originating from the Kaggle Customer Support On Twitter dataset, these real-world interactions span various industries, such as airlines and retail, averaging ten turns per dialogue and 36 tokens per abstractive summary.

TodSum (Zhao et al., 2021b) is a customer service dataset based on MultiWOZ (Budzianowski et al., 2018), from which they select the five domains: restaurant, hotel, attraction, taxi, and train. The dataset spans 10k samples with an average dialogue length of 187 utterances and 45 words in the summary.

4.6 Medical

MTS-Dialog (Ben Abacha et al., 2023) is a collection of 1.7k simulated doctor-patient conversations with corresponding clinical notes serving as summaries, sourced from the public Mtsamples collection (Moramarco et al., 2021). This dataset encompasses various medical fields, including general medicine, neurology, orthopedics, dermatology, and immunology, adhering to the SOAP (Subjective, Objective, Assessment, Plan) note format. Dialogues average 18 sentences and 242 words, with summaries typically around 81 words.

4.7 Debates

ADSC (Misra et al., 2015) features sequences of two-party dialogue chains derived from the Internet Argument Corpus (Walker et al., 2012), focusing on significant social and political topics like gun control, gay marriage, the death penalty, and abortion. It includes 45 dialogues, each accompanied by five unique human-generated summaries, with each summary approximately 150 words in length.

4.8 Others

ConvoSumm (Fabbri et al., 2021a) consolidates dialogues from four sources: New York Times comments, StackExchange, W3C emails, and Reddit, totaling 2k dialogues with 500 from each domain. Crowdsourced workers from Amazon Mechanical Turk wrote the abstractive summaries with at most 90 tokens. Inputs average 1.1k tokens, with summaries about 70 tokens in length.

Ferranti (Gao et al., 2023) is a dataset designed for factual error correction in dialogue summarization, featuring 4k manually annotated items. The original task is to evaluate the factuality of summaries and how to correct these summaries with minimal effort. Drawing on SAMSum and DialogSum, Ferranti includes summaries produced by models like BART

(Lewis et al., 2020) and UniLM (Dong et al., 2019). Annotators assess these summaries for accuracy and identify errors, providing a focused tool for improving summary factualness.

DialSum (Fang et al., 2021) is based on the VisDial dataset (Das et al., 2017), where two participants discuss images from the MSCOCO dataset (Lin et al., 2015), which features ∼120k images. Each image has five captions from five different annotators.

4.9 Data Augmentation and Utility

Besides highlighting datasets with a notable community interest, we summarize in this subsection the research on techniques to cope with the data scarcity in dialogue summarization, covering artificial generation techniques of datasets through text generation and curriculum learning strategies to use the available data more effectively.

Artificial Dataset Generation. As discussed earlier, dialogue summarization faces challenges due to the need for adaptability across various domains, structures, and speaker dynamics, which typically would be addressed by training models on diverse datasets (Feng et al., 2022). However, datasets' scarcity and small size, with utterances containing just two to ten turns in areas like customer service and medical, restrict model training for more general applications.

Creating real-world datasets is costly and may conflict with data security and information communication policies, so artificially generated datasets are explored (Ben Abacha et al., 2023). Methods range from simple heuristic-based weak labeling, such as selecting the leading or longest utterance as proxy summary (Sznajder et al., 2022), to paraphrasing with updating the summary to maintain coherence (Liu et al., 2022; Wahle et al., 2022, 2023), to random alterations (e.g., swapping, deletion, dialogue-act-guided insertions) and changing conversation structures (Chen & Yang, 2021; Park et al., 2022). The most common method is to use language models to create ground truth summaries either directly (Asi et al., 2022; Nair et al., 2023; Zhou et al., 2023a; Zhu et al., 2023a) or after training them on human summarization patterns through few-shot learning (Chintagunta et al., 2021).

Techniques to Maximize Dataset Utility. As the training of Transformer-based models requires extensive data to generalize across various topics and dialogue formats (Fu et al., 2021), researchers have explored ways to use the limited data available more effectively. A key strategy involves a prompt-based curriculum learning strategy that gradually increases the degree of prompt perturbation (e.g., word swapping, content cutting) to improve the generalization ability of models (Li et al., 2022). Another variation are dynamic prompts, which select best-fitting few-shot samples considering dialogue content, size, and speaker number (Prodan & Pelican, 2022). Also explored are prompt transfer techniques from related dialogue domains (Xie et al., 2023) to bolster the usage of dialogue state information (i.e., data used to represent the underlying intentions and goals within a dialogue). Prompts are further used to split inputs into domain-invariant and domain-specific content to enhance model generalization (Zhao et al., 2022; Li et al., 2023). Further techniques include freezing most model parameters and training only specific parameters for domain adaption (Chen et al., 2023; Suri et al., 2023; Zhu et al., 2023b).

Type	Dataset	Challenge	Descriptive Tags	Usage
Daily Chat	DialogSum	Speaker	daily life scenarios	26
Online Chat	SAMSum	Speaker, Salience	messaging apps	68
	FORUM	Salience	threads	5
	CRD3	Language	live-streamed show	6
Meeting	AMI	Salience, Comprehension	staged business meetings	63
	ICSI	Salience, Language	academic group meetings	21
	QMSum	Salience, Language	query-based summarization	18
	Elitr	Salience	technical project meetings in computer science	7
	MeetingBank	Structure	parliament meetings	7
Screenplay	MediaSum	Structure	media interview	15
	SummScreen	Speaker, Salience	TV series transcripts	8
Customer Service	TweetSumm	Language	Twitter customer support	7
	TODSum	Language, Factuality	open-domain task-oriented dialogues	6
Medical	MTS-Dialog	Comprehension	simulated doctor-patient conversations with clinical notes	5
Debates	ADSC	Comprehension	two-party dialogues on social and political topics	5
Others	ConvoSumm	Salience	dialogues from comments, emails, and threads	5
	Ferranti	Factuality	human assessed automatic summaries	5
	DialSum	Comprehension	two-party discussion on images	10

Table 3: Matching between established datasets and their primary reported challenge. 'Usage' indicates the number of papers reporting the dataset in their publication.

5. Evaluation

Researchers have adopted metrics from related fields and introduced new ones designed to assess the effectiveness of dialogue summarization methods. These metrics proxy for quality characteristics, such as coherence, fluency, factuality, and accuracy. Our literature review identifies 15 metrics used more than twice in 93 papers. We categorize these metrics into four groups: count-based (e.g., ROUGE Lin, 2004), model-based (e.g., BARTScore Yuan et al., 2021), QA-based (e.g., QAEval Deutsch et al., 2021), and human evaluation. The automatic metrics strive for alignment with human judgment. However, the correlation between these two is weak for dialogue summarization, and automatic metrics sometimes reward low-quality texts (Gatt & Krahmer, 2018). Due to these recognized limitations in fully capturing the nuances of summarization quality, human evaluation is considered the gold standard across most studies. Recent analyses (Gao & Wan, 2022; Kirstein et al., 2024d) conclude that established automatic metrics may work for a superficial understand-

ing of a model's performance but do not align well with human judgment in discerning error nuances. These metrics show individual weaknesses where they individually do not adequately reflect occurrence across all error types (e.g., of hallucinated content) or cannot show the impact on the quality in their scores (e.g., when information is omitted). Hence, a composite metric of count-based, model-based, and QA-based metrics may be required for a more reliable automatic interpretation, with the individual metrics focusing on specific error types.

Section 5.1 details the characteristics typically evaluated by metrics and Sections 5.2 to 5.5 further summarizes the 15 established metrics in dialogue summarization. Table 4 provides an overview of the identified metrics and their usage throughout the papers considered for this literature review.

5.1 Assessed Characteristics

Metrics assess the quality of a generated text, often in relation to a reference text. Low-quality text may contain more redundancy, incoherence, grammatical errors, poor structure, or inappropriate language, while high-quality text would closely align with the reference. We identify ten key characteristics discussed in the literature (in *italic*) and organize them into four overarching groups: accuracy, content, readability, and style.

Accuracy is the core of a high-quality summary (Neto et al., 2002), making it crucial to maintain the truth of the original text and ensure *factuality*. A summary must accurately reflect the facts, events, and details from the source without any distortion.

Content quality is mainly driven by *relevance*. A summary must contain the most relevant information, directly addressing the addressee's informational needs (Williams et al., 2014). This effort to overcome the salience challenge ensures that summaries prioritize the most important points for the reader. *Coverage* complements relevance by ensuring that all key topics and arguments from the dialogue are included, offering a comprehensive understanding without significant omissions (Mullick et al., 2024). The goal is to track both the salience and structure challenges with this. *Informativeness* goes a step further by selecting crucial information that conveys the depth of the input transcript, enabling readers to grasp the main points without needing to refer back to the original text (See et al., 2017).

Readability is essential for making summaries accessible and easy to grasp the key points quickly. *Coherence* ensures that information is presented logically with smooth transitions, making the summary easy to follow (Mullick et al., 2024). The *structure and organization* of a summary further enhances its readability and facilitates information processing, ensuring that it is well-organized and logical (Carbonell & Goldstein, 1998). *Conciseness* (Biswas & Iakubovich, 2022) and *non-redundancy* (Yang et al., 2020) are also essential as a summary should contain only the essence of the input transcript without unnecessary details.

Style pertains to the text's perceived quality and formal presentation, contributing to its professional polish. *Consistency* in perspective, tense, and stylistic choices throughout the summary contributes to the overall perception of the text's professionalism (King et al., 2022). *Fluency* focuses on grammar, vocabulary, and sentence structure and aims for a natural, easy-to-read summary free from phrasing errors, enhancing the information's overall clarity and accessibility (Kryscinski et al., 2019).

5.2 Count-Based

Count-based metrics, including N-gram-based measures like BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004), and statistical measures such as Perplexity (Jelinek et al., 1977), are often static and rule-based algorithms that have a long history for evaluating text summarization. However, due to their simplicity, they have been criticized for their limitations in capturing overall meaning, fluency, coherence, or factuality (Sai et al., 2022).

BLEU (Papineni et al., 2002) measures the precision of generated summaries against reference texts by examining the overlap of word sequences (typically 1- to 3-word N-grams, Yang et al., 2019), emphasizing similarity to the original phrasing but risking a linear relationship to noise (Vaibhav et al., 2019).

ROUGE (Lin, 2004) focuses on lexical similarity (Ng & Abrecht, 2015) and aims to capture the extent to which key content from the source is included in the summary, prioritizing content coverage through recall of N-gram overlap, thus attempting to address BLEU's limitations by emphasizing content inclusion over mere precision. A frequently used variation is ROUGE-L, which measures the longest common subsequence between a candidate and a reference.

METEOR (Banerjee & Lavie, 2005) advances BLEU and ROUGE by incorporating both precision and recall, along with synonym matching for semantic analysis, thus enabling the capturing of the semantic similarity between a candidate and a reference. Despite its more balanced approach and sentence-level focus, METEOR, like its predecessors, is prone to noise interference (Vaibhav et al., 2019).

chrF++ (Popović, 2017) further expands the previous metrics by considering precision, recall, and F-score-based N-gram overlap at both character and word levels, offering a more granular analysis (Popović, 2015).

CIDEr (Vedantam et al., 2015) is a consensus-based metric that compares the similarity of a generated sentence against a set of human-written reference sentences. The score is an aggregation of cosine similarity scores between the TF-IDF weighted N-grams of the generated and reference sentences, inherently capturing precision, recall, grammaticality, and salience (Li & Liang, 2021; Fabbri et al., 2021b; Lu et al., 2022).

Perplexity (Jelinek et al., 1977) diverges from the previous methods by the statistical approach of gauging a model's uncertainty in predicting word sequences, with lower scores indicating better alignment with expected language patterns. This measure is, in contrast to the previous, an intrinsic evaluation metric that directly evaluates the language modeling objectives through text predictability rather than assessing the performance on the downstream task. Therefore, additional metrics are required for a holistic evaluation of text generation quality.

5.3 Model-Based

With advancements in language models, there is an increasing focus on model-based evaluation metrics due to their higher correlation with human judgment, though count-based metrics remain popular. Model-based metrics encompass a wide range of approaches, including those that calculate semantic similarity (e.g., BERTScore, MoverScore), text generation likelihood (e.g., BARTScore), and entailment probability (e.g., FactCC). These metrics typically represent text in a latent space using pre-trained embeddings or contex-

tual representations, aiming to provide a more nuanced assessment by focusing on semantic similarity, likelihood of text generation, and factual consistency. While these metrics have the potential to adapt to evolving language use (e.g., distribution drifts) (Sellam et al., 2020), they can still be error-prone (Ji et al., 2023), slower than count-based metrics, and may not directly measure specific characteristics (as outlined in Section 5.1), making it challenging to discern which particular aspect influences their scores.

BERTScore (Zhang et al., 2020b) leverages BERT embeddings (Devlin et al., 2019) to assess textual similarity. The metric first contextually embeds the reference and candidate texts, then constructs a similarity matrix through pairwise cosine similarities on the token level. The final score is computed by greedily selecting the highest similarity scores and calculating the harmonic mean of precision and recall, enabling the metric to capture semantic nuances beyond surface-level matching (Zhao et al., 2019).

MoverScore (Zhao et al., 2019) was introduced concurrently to BERTScore following a similar approach, as it also leverages BERT embeddings but instead considers the distance between reference and candidate text, hence employing the Word Mover's Distanc (Kusner et al., 2015), a special case of the Earth Mover's Distance (Rubner et al., 2000), to measure semantic distance. This change in measure allows MoverScore to map semantically related words from one sequence to one word in another sequence (many-to-one).

BARTScore (Yuan et al., 2021) evaluates the plausibility of generating a reference text from a given generated text, and vice versa, by calculating the log-likelihood of a sequence that a BART (Lewis et al., 2020) model would typically generate based on the given context. This evaluation focuses on assessing both the fluency and the semantic accuracy.

BLEURT (Sellam et al., 2020) extends beyond mere embedding comparisons by incorporating a BERT model pre-trained on lexical- and semantic-level supervision signals and fine-tuned on human judgments, enabling it to make detailed assessments of text quality, including coherence and relevance.

BLANC (Vasilyev et al., 2020) leverages BERT to perform a fill-in-the-blank task, both with and without the generated summary, to assess how informative or helpful the generated text is. The difference in prediction accuracy indicates the utility of the summary in helping understand the text.

FactCC (Kryscinski et al., 2020) evaluates a summary's factual consistency with its source document using an entailment classifier, scoring based on the proportion of sentences classified as factual consistent by a BERT model. The metric struggles with complex inferences and subtle nuances beyond direct comparison.

5.4 QA-Based

The QA-based metrics we identify in the literature focus on the factual correctness of a summary by using external question-answering systems. These metrics leverage pre-trained transformer-based models to generate questions and assess whether the summary contains the correct answers. Their effectiveness largely depends on the quality of the underlying QA systems, which may not always align perfectly with human evaluation standards.

FEQA (Durmus et al., 2020) generates questions based on the summary content and verifies whether their answers can be found in the source document.

SummaQA (Scialom et al., 2019) derives questions from the source text and tries to answer these using the summary. It expands the QA-based evaluation scope by incorporating various question types and emphasizing the summary's informativeness.

QuestEval (Deutsch et al., 2021) combines the approaches of FEQA and SummaQA, thereby adopting a bidirectional strategy, generating questions from the summary to compare with the source text and vice versa. This bidirectional evaluation offers a balanced and holistic assessment by considering included and omitted information in the summary.

5.5 Human Evaluation

Human evaluation is often considered the gold standard for assessing the quality of a summary. It is typically performed through crowdsourcing annotators (e.g., Amazon Mechanical Turk) who label samples. We identify established approaches for human evaluations on summary performance and annotator agreement.

Performance. Summary performance is typically evaluated using *Likert* scales (Likert, 1932; Qader et al., 2018; Feng et al., 2021; Lu et al., 2022), which provide a simple rating system for quality assessment but lack detailed feedback on specific text issues (Dou et al., 2022). Alternatives like *pairwise comparison* (Elder et al., 2018; Liu et al., 2021) and *best-worst scaling* (Finn & Louviere, 1992; Rothe et al., 2020) offer more nuanced evaluations, with best-worst scaling noted for its higher reliability (Kiritchenko & Mohammad, 2017). Despite these options, the Likert scale remains the predominant method.

Agreement. The reliability of human-generated evaluations hinges on annotator agreement, also referred to as meta-evaluation (Yuan et al., 2021). Despite its significance, the incorporation of agreement metrics is frequently overlooked. We identify three established measures to determine inner-annotator agreements: Krippendorff's Alpha, Cohen's Kappa, and Fleiss' Kappa. Their scores range from 0 (poor reliability) to 1 (perfect reliability), with reported scores typically between 0.65 and 0.85. Krippendorff's Alpha (Krippendorff, 1970) offers a flexible approach suitable for multiple raters, applicable to any level of measurement, i.e., nominal, ordinal, interval, or ratio, and accommodates both qualitative and quantitative assessments. This measure is particularly valuable with a broad range of variables or when comparing agreements across different measurement scales but assumes that all annotators assess all samples. Cohen's Kappa (Cohen, 1960) is used to measure the agreement between two raters who each classify a set of items into mutually exclusive categories (e.g., yes/no). Fleiss' Kappa (Fleiss, 1971) is also tailored for scenarios involving fixed-category classifications but extends Cohen's Kappa to multiple annotators, excelling in the evaluation of how consistently annotators categorize text segments or dialogue turns into predefined groups.

6. Discussion

Throughout this work, we provide an overview of the current state of challenges, datasets, and evaluation. In Section 6.1, we analyze the emerging trends such as LLMs on mitigating the individual challenges, finding that our challenge taxonomy remains up to date. Section 6.2 discusses the increasing interest in datasets covering personalized summarization and realistic settings, while Section 6.3 covers recent research on improving evaluation metrics.

Type	Category	Metric	Descriptive Tags	Usage
Count-based	N-gram	Bleu	word overlap, precision, multiple references	21
		Rouge	word overlap, recall, one reference	127
		Meteor	word overlap, precision and recall, one reference	3
		ChrF++	character-level, F1 score, one reference	2
		Cider	consensus-based, multiple references	3
	Statistical	Perplexity	likelihood of word sequences	3
Model-based	Hybrid	Bertscore	token similarity, cosine similarity, one-to-one	38
		Moverscore	token similarity, mover distance, one-to-many	3
		Factcc	entailment classifier, scoring based on consistency	9
	Trained	Bleurt	mimics human judgment	7
		Bartscore	mimics human judgment, promptable	13
		Blanc	fill-in-the-blank task with and without the summary	2
QA-based		Feqa	questions based on summary, answered through input	38
		Summaqa	questions based on input, answered through summary	3
		Questeval	combination of Feqa and Summaqa	3
Human Evaluation	Performance	Likert Scale	ordinal scale, e.g., 1 (worst) to 5 (best)	6
		Pairwise Comparison	pick best sample out of two	9
		Best-Worst Scaling	rank a list of samples	5
	Agreement	Krippendorff's Alpha	measure disagreement, different data formats	2
		Cohen's Kappa	measure agreements, categorical data, 2 raters	3
		Fleiss' Kappa	measure agreement, nominal data, 2 raters	2

Table 4: Relevant metrics and evaluation measures employed in dialogue summarization with more than twice reported use. 'Usage' states the number of papers reporting the evaluation measure in their publication.

6.1 Remarks on Challenges

In this study, we have introduced the CADS taxonomy to organize the inherent challenges of abstractive dialogue summarization (i.e., language, speaker, salience, comprehension, structure, and factuality) and unify their underlying definitions. While presented separately for clarity, the challenges are interdependent and influence each other. Consequently, approaches should be researched to tackle the challenges simultaneously whenever possible. Despite advancements in addressing the challenges, we observe that most of them are still a hurdle for models due to limitations in the Transformer components used in these models, the lack of capabilities in contextualization and few-shot adaption of the encoder-decoder backbone models, and missing mitigation techniques.

Since 2023, NLP has seen a significant shift with the exploration of LLMs and optimized Transformer architecture components (e.g., ring attention, Liu et al., 2023, and multi-token generation, Gloeckle et al., 2024), with their application to dialogue summarization being explored later (Zhou et al., 2023b; Lyu et al., 2024b; Mullick et al., 2024). Initial studies indicate that LLMs match or exceed the performance of task-specific models like DialogLED (Zhong et al., 2022), even with their limitations (e.g., hallucinations, salience) (Laskar et al., 2023). Kirstein et al. (2024d) show that encoder-decoder models may perform better against the speaker challenge, particularly regarding coreference resolution, and struggle with the structure challenge. LLMs handle the comprehension challenge but seem to be sub-optimal regarding robustness to language and speakers. These early observations state LLMs as a noteworthy alternative to more traditional techniques (e.g., dialogue-style pre-training, Zhong et al., 2022, graph structures to represent speaker structures, Hua et al., 2022, and special losses tailored to determine salience, Huang et al., 2023).

Following, we discuss how the hurdles of the currently employed encoder-decoder models (identified in Section 3) align with challenges in NLP and discuss techniques introduced to mitigate these challenges, thereby deriving how they can aid dialogue summarization.

Bridging the language gap. Adapting established summarization models from structured pre-training data to the less formal nature of dialogue remains challenging. LLMs, especially those trained on non-edited text such as models from the GPT and GEMINI series, are promising in this regard (Lyu et al., 2024a). Their few-shot learning capabilities allow for rapid adaptation to new linguistic patterns more effectively. At the same time, the pre-training on a large and diverse corpus improves robustness in handling informal language, style variations, and repetition (Wang et al., 2024).

Long distance handling and dependency parsing. Recent progress in LLMs such as GPT-4, Claude-3, and Phi-3 shows a significant extension of the processable context length to over 100k tokens, exceeding the context size of earlier models like LED (Beltagy et al., 2020), which handle up to 16k tokens. This increase in context allows for handling long dependencies without chunking, which is crucial for understanding complex dialogue structures, speaker dynamics, and content understanding. While traditional models could theoretically handle such lengthy contexts, practical limitations arise from the input-length-dependent quadratic computational costs associated with Transformer attention mechanisms (Vaswani et al., 2017). To address these issues, techniques like sparse attention (Beltagy et al., 2020), flash attention during the training stage (Dao et al., 2024) and ring attention for inference (Liu et al., 2023), sharing weights across attention heads

(Shazeer, 2019), and conditional computation to reduce the overall computational load (Ainslie et al., 2023) are proposed to manage large inputs.

Generalizability. Modifying pre-trained models to handle the salience and structure challenges independent of the conversation's domain is essential for practical applications (e.g., customer service). Established generalization techniques involving prompting (Ma et al., 2022; Wang et al., 2022a), few-shot learning (Dang et al., 2022; Qin & Joty, 2022), and meta-learning (Vilalta & Drissi, 2002; Hospedales et al., 2022) are incipient in dialogue summarization. Meanwhile, LLMs build on these techniques to leverage patterns observed during the training stage and boost their generalization capabilities (Wilson et al., 2023). However, LLMs may still struggle with bias mitigation and ensuring broad applicability across various contexts when encountered words are not present in the training corpus or in examples (Bakker et al., 2024; Talat et al., 2022; Wolf et al., 2024).

Coreference resolution. Large-scale pre-trained and fine-tuned language models currently set the benchmark within NLP for coreference resolution (Liu et al., 2023). Challenges remain with the ambiguities in reference and context (Khurana et al., 2023), and the usage of LLMs for this task.

Hallucination reduction. Avoiding hallucinated content in generated text is a major challenge in NLP tasks using LLMs, especially in culturally sensitive contexts (Ji et al., 2023; McIntosh et al., 2023). Mitigation techniques include confidence penalty regularization (Lu et al., 2021; Liu et al., 2023) to reduce overconfidence and enhance accuracy and refinement methods to review and post-process a summary (Kirstein, Ruas, & Gipp, 2024b).

These advancements throughout NLP suggest upcoming shifts in the key challenges of dialogue summarization.

- The *language* challenge can lose impact if few-shot learning is used to prompt LLMs, making it easier to apply document summarization techniques to dialogues directly without additional finetuning.

- We expect progress for the *structure* challenge using LLMs, but robustness issues may persist due to issues with generalization and processing long texts.

- The *comprehension* challenge, especially in grasping implied meanings, remains largely unsolved and holds significant research potential. Research from related NLP fields such as sentiment analysis (Srivastava et al., 2020) suggests that current models still struggle to understand implied content. At the same time, research on retrieval-augmented generation (Balaguer et al., 2024) could enhance contextualization, helping models better grasp the context of conversations.

- For the hurdles implied in the *speaker* challenge, i.e., coreference resolution and dependency parsing, LLMs are explored but have not yet succeeded due to robustness weaknesses (e.g., errors due to uncommon names, overseeing details in long inputs), signaling that further research is required.

- The *salience* challenge builds on personalizing summaries and understanding long discussions, which both are gaining attention lately. Due to the novelty, the personalized summaries require more research on techniques and benchmarking datasets.

- The ongoing issue of hallucination in language models continues to pose a challenge in ensuring the *factuality* in generated summaries.

6.2 Trends in Datasets

Personalized Summarization. In contrast to the typically general summaries, personalized summaries tailored to the users' needs are becoming increasingly popular (Tepper et al., 2018; Khurana et al., 2024; Kirstein et al., 2024c). Recent advancements have explored cross-attention and decoder self-attention to enhance role-specific information capturing (Lin et al., 2022). By integrating detailed personal attributes, models can better understand each participant's background and motivations, leading to more targeted summaries. Dataset-wise, only a few, such as the Chinese CSDS (Lin et al., 2021), incorporate personalized summaries, whereas established English datasets typically provide only a single, general summary.
Realistic Datasets. As models improve in few-shot learning and do not necessarily require finetuning, the reliance on large-scale datasets is shrinking. Considering this trend, we expect a rise in smaller datasets, which allow for high-quality, realistic examples. These examples may be sufficient to adapt models to new domains, make approaches more robust, and test them against realistic scenarios.

6.3 Trends in Evaluation

Evaluation metrics, vital for indicating performance and comparing new techniques, are currently mostly borrowed from related NLP tasks such as translation or text generation (Section 5). However, the effectiveness of the established metrics (e.g., ROUGE for dialogue summarization as identified in Table 4) is limited (Gao & Wan, 2022; Kirstein et al., 2024d), and while it can serve as a proxy (Wang et al., 2022b) it provides insufficient insights into the true efficacy of new techniques. The also popular model-based metrics (e.g., BARTScore) seem unable to align well with human judgments (Gao & Wan, 2022) for dialogue summarization. Recent developments in NLP use LLM-based metrics such as GEMBA (Kocmi & Federmann, 2023) and ICE (Jain et al., 2023) for evaluation (Nair et al., 2023), building on LLMs' advanced text comprehension. This set of metrics thereby mimics human judgment and assesses common aspects (e.g., fluency, coverage, coherence) with continuous (Jain et al., 2023), Likert scale (Likert, 1932; Chiang & Lee, 2023), or probability scores (Fu et al., 2024). LLM-based metrics offer a promising direction for dialogue summarization evaluation due to their customizability (Kirstein, Ruas, & Gipp, 2024a), though this area remains largely unexplored.

7. Final Considerations

In this article, we reviewed works on Transformer-based abstractive dialogue summarization. We unified the existing definitions and concepts of dialogue summarization into a taxonomy (CADS) of the field's six main challenge blocks: language, structure, comprehension, speaker, salience, and reliability-related factuality. We demonstrated how these challenges appear in dialogue summarization and discussed why they occur for established encoder-decoder-based summarization systems. We then grouped and highlighted tech-

niques introduced since 2019 under the challenges they tackle. Despite advancements, we observed that the field is still in its infancy and offers ample opportunities for further research. We also listed datasets used in existing studies to illustrate the extent of data scarcity in dialogue summarization. Additionally, we evaluated common metrics and noted a strong reliance on the ROUGE metric, coupled with a lack of human evaluation reports, which raises doubts about the actual effectiveness of current techniques. In our discussion, we evaluated how current NLP techniques address these challenges and derived implications of using LLMs for the field. We thereby identify a lack of exploring such new models and borrowing effective techniques from related fields to overcome the limitations currently holding back progress. We conclude that while challenges like language may become less relevant, others, such as comprehension and factuality, still require more exploration. We recommend that future work should further discuss and adapt this taxonomy as new challenges, techniques, datasets, and evaluation measures emerge and as current challenges may be solved naturally through technological innovation. The considered literature can be found in the accompanying repository[9] which will be regularly updated.

In the following, we discuss the discuss limitations of evidence (Section 7.1) and in our review process (Section 7.2).

7.1 Limitations of Evidence

Given our selection of papers on dialogue summarization, we found limitations in current works that may bias our findings. First, the crawled research studies in dialogue summarization predominantly focus on English-language dialogues, which renders their performance in multilingual or cross-cultural contexts unknown. While we frequently came across Chinese datasets, other languages are underrepresented. As a consequence, we focus our work on English dialogue summarization. Second, most datasets stem from specific sectors, such as customer service or business meetings, which may not reflect the various dialogue types in different environments. Lastly, the lack of detailed human evaluation reporting across research works, including inter-annotator agreement scores, further complicates the assessment of new techniques' usefulness.

7.2 Limitations of Review Process

In our systematic review of dialogue summarization, we encountered methodological challenges that we had to mitigate. First, we focused our search on studies published in English and mainly using English datasets. While this is a limiting factor, most research on dialogue summarization considers only English dialogues, making this a feasible approximation. However, we encourage others to assess the issues we point out for other languages. Second, despite considering multiple databases, the timing and methods of our search may have missed studies, thus not fully capturing the dialogue summarization field. To mitigate this limitation, we updated our crawl several times throughout writing this article and retrieved works from two databases. We considered both peer-reviewed and non-peer-reviewed papers, which risks including less credible studies or misinformation. To mitigate the potential weaknesses of non-peer-reviewed works, we use the ranking from semantic

9. https://github.com/FKIRSTE/LitRev-DialogueSum

scholar and DBLP, which should account for the paper's quality and evaluate the studies' quality based on adherence to established reporting standards of top-tier conferences.

Acknowledgments

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

References

Ainslie, J., Lei, T., de Jong, M., Ontanon, S., Brahma, S., Zemlyanskiy, Y., Uthus, D., Guo, M., Lee-Thorp, J., Tay, Y., Sung, Y.-H., & Sanghai, S. (2023). CoLT5: Faster Long-Range Transformers with Conditional Computation. In Bouamor, H., Pino, J., & Bali, K. (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5100 Singapore. Association for Computational Linguistics.

Altmami, N. I., & Menai, M. E. B. (2022). Automatic Summarization of Scientific Articles: A Survey. *Journal of King Saud University - Computer and Information Sciences*, 34, 4, 1011–1028.

Antony, D., Abhishek, S., Singh, S., Kodagali, S., Darapaneni, N., Rao, M., Paduri, A. R., & BG, S. (2023). A Survey of Advanced Methods for Efficient Text Summarization. In *2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0962–0968.

Asi, A., Wang, S., Eisenstadt, R., Geckt, D., Kuper, Y., Mao, Y., & Ronen, R. (2022). An End-to-End Dialogue Summarization System for Sales Calls. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pp. 45–53. Association for Computational Linguistics.

Asthana, S., Hilleli, S., He, P., & Halfaker, A. (2024). Summaries, Highlights, and Action Items: Design, Implementation and Evaluation of an LLM-powered Meeting Recap System. arXiv. arXiv:2307.15793.

Bakker, M. A., Chadwick, M. J., Sheahan, H. R., Tessler, M. H., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M. M., & Summerfield, C. (2024). Fine-Tuning Language Models to Find Agreement among Humans with Diverse Preferences. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 38176–38189 Red Hook, NY, USA. Curran Associates Inc.

Balaguer, A., Benara, V., Cunha, R. L. d. F., Filho, R. d. M. E., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V., & Chandra, R. (2024). RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture.

Banerjee, S., & Lavie, A. (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*, pp. 65–72. Association for Computational Linguistics.

Beckage, N., H Kumar, S., Sahay, S., & Manuvinakurike, R. (2021). Context or No Context? A Preliminary Exploration of Human-in-the-Loop Approach for Incremental Temporal Summarization in Meetings. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pp. 96–106. Association for Computational Linguistics.

Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. arXiv. arXiv:2004.05150.

Ben Abacha, A., Yim, W.-w., Fan, Y., & Lin, T. (2023). An Empirical Study of Clinical Note Generation from Doctor-Patient Encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2291–2302. Association for Computational Linguistics.

Bertsch, A., Neubig, G., & Gormley, M. R. (2022). He Said, She Said: Style Transfer for Shifting the Perspective of Dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4823–4840. Association for Computational Linguistics.

Bhatia, S., Biyani, P., & Mitra, P. (2014). Summarizing Online Forum Discussions – Can Dialog Acts of Individual Messages Help?. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2127–2131. Association for Computational Linguistics.

Biswas, P. K., & Iakubovich, A. (2022). Extractive Summarization of Call Transcripts. *IEEE access : practical innovations, open solutions*, 10, 119826–119840.

Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from Applying the Systematic Literature Review Process within the Software Engineering Domain. *Journal of Systems and Software*, 80, 4, 571–583.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., & Gašić, M. (2018). MultiWOZ - a Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5016–5026. Association for Computational Linguistics.

Carbonell, J., & Goldstein, J. (1998). The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pp. 335–336. Association for Computing Machinery.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., & Wellner, P. (2006). The AMI Meeting Corpus: A Pre-announcement. In Renals, S., & Bengio, S. (Eds.), *Machine Learning for Multimodal Interaction*, pp. 28–39. Springer.

Chauhan, V., Roy, P., Dey, L., & Goel, T. (2022). TCS_WITM_2022 @ DialogSum : Topic Oriented Summarization Using Transformer Based Encoder Decoder Model. In *Proceedings of the 15th International Conference on Natural Language Generation: Generation Challenges*, pp. 104–109. Association for Computational Linguistics.

Chen, J., Dodda, M., & Yang, D. (2023). Human-in-the-Loop Abstractive Dialogue Summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 9176–9190. Association for Computational Linguistics.

Chen, J., & Yang, D. (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)*, pp. 4106–4118. Association for Computational Linguistics.

Chen, J., & Yang, D. (2021). Simple Conversational Data Augmentation for Semi-Supervised Abstractive Dialogue Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6605–6616. Association for Computational Linguistics.

Chen, M., Chu, Z., Wiseman, S., & Gimpel, K. (2022). SummScreen: A Dataset for Abstractive Screenplay Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8602–8615. Association for Computational Linguistics.

Chen, Y., Liu, Y., Chen, L., & Zhang, Y. (2021). DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 5062–5074. Association for Computational Linguistics.

Chen, Y., Liu, Y., Xu, R., Yang, Z., Zhu, C., Zeng, M., & Zhang, Y. (2023). UniSumm and SummZoo: Unified Model and Diverse Benchmark for Few-Shot Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12833–12855. Association for Computational Linguistics.

Chiang, C.-H., & Lee, H.-y. (2023). Can Large Language Models Be an Alternative to Human Evaluations?. In Rogers, A., Boyd-Graber, J., & Okazaki, N. (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631. Association for Computational Linguistics.

Chintagunta, B., Katariya, N., Amatriain, X., & Kannan, A. (2021). Medically Aware GPT-3 as a Data Generator for Medical Dialogue Summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pp. 66–76. Association for Computational Linguistics.

Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., & Goharian, N. (2018). A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 615–621. Association for Computational Linguistics.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales - Jacob Cohen, 1960.

Cui, L., Wu, Y., Liu, S., Zhang, Y., & Zhou, M. (2020). MuTual: A Dataset for Multi-Turn Dialogue Reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1406–1416. Association for Computational Linguistics.

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models. arXiv. arXiv:2209.01390.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., & Ré, C. (2024). FLASHATTENTION: Fast and Memory-Efficient Exact Attention with IO-awareness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 16344–16359 Red Hook, NY, USA. Curran Associates Inc.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., & Batra, D. (2017). Visual Dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1080–1089. IEEE Computer Society.

Deng, J., Cheng, J., Sun, H., Zhang, Z., & Huang, M. (2023). Towards Safer Generative Language Models: A Survey on Safety Risks, Evaluations, and Improvements. arXiv. arXiv:2302.09270.

Deutsch, D., Bedrax-Weiss, T., & Roth, D. (2021). Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary. *Transactions of the Association for Computational Linguistics*, 9, 774–789. MIT Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics.

Dong, C., Li, Y., Gong, H., Chen, M., Li, J., Shen, Y., & Yang, M. (2022). A Survey of Natural Language Generation. *ACM Computing Surveys*, 55, 8, 173:1–173:38.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., & Hon, H.-W. (2019). Unified Language Model Pre-Training for Natural Language Understanding and Generation. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13042–13054.

Dou, Y., Forbes, M., Koncel-Kedziorski, R., Smith, N. A., & Choi, Y. (2022). Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7250–7274. Association for Computational Linguistics.

Duan, X., Zhang, Y., Yuan, L., Zhou, X., Liu, X., Wang, T., Wang, R., Zhang, Q., Sun, C., & Wu, F. (2019). Legal Summarization for Multi-Role Debate Dialogue via Controversy Focus Mining and Multi-Task Learning. In Zhu, W., Tao, D., Cheng, X., Cui, P., Rundensteiner, E. A., Carmel, D., He, Q., & Yu, J. X. (Eds.), *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pp. 1361–1370. ACM.

Durmus, E., He, H., & Diab, M. (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*, pp. 5055–5070. Association for Computational Linguistics.

Elder, H., Gehrmann, S., O'Connor, A., & Liu, Q. (2018). E2E NLG Challenge Submission: Towards Controllable Generation of Diverse Natural Language. In *Proceedings of the 11th International Conference on Natural Language Generation*, pp. 457–462. Association for Computational Linguistics.

Fabbri, A., Rahman, F., Rizvi, I., Wang, B., Li, H., Mehdad, Y., & Radev, D. (2021a). ConvoSumm: Conversation Summarization Benchmark and Improved Abstractive Summarization with Argument Mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6866–6880. Association for Computational Linguistics.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021b). SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9, 391–409. MIT Press.

Fagan, J. C. (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. American Library Association.

Fang, T., Pan, H., Zhang, H., Song, Y., Xu, K., & Yu, D. (2021). Do Boat and Ocean Suggest Beach? Dialogue Summarization with External Knowledge. *AKBC 2021*.

Fang, Y., Zhang, H., Chen, H., Ding, Z., Long, B., Lan, Y., & Zhou, Y. (2022). From Spoken Dialogue to Formal Summary: An Utterance Rewriting for Dialogue Summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3859–3869. Association for Computational Linguistics.

Fazly, A., Cook, P., & Stevenson, S. (2009). Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35, 1, 61–103. MIT Press.

Feigenblat, G., Gunasekara, C., Sznajder, B., Joshi, S., Konopnicki, D., & Aharonov, R. (2021). TWEETSUMM - a Dialog Summarization Dataset for Customer Service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 245–260. Association for Computational Linguistics.

Feng, X., Feng, X., & Qin, B. (2022). A Survey on Dialogue Summarization: Recent Advances and New Frontiers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 5453–5460 Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization.

Feng, X., Feng, X., Qin, L., Qin, B., & Liu, T. (2021). Language Model as an Annotator: Exploring DialoGPT for Dialogue Summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1479–1491. Association for Computational Linguistics.

Finn, A., & Louviere, J. J. (1992). Determining the Appropriate Response to Evidence of Public Concern: The Case of Food Safety. *Journal of Public Policy & Marketing*, 11, 2, 12–25. American Marketing Association.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement among Many Raters. *Psychological Bulletin*, 76, 5, 378–382. American Psychological Association.

Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic Plagiarism Detection: A Systematic Literature Review. *ACM Computing Surveys*, 52, 6, 112:1–112:42.

Fu, J., Ng, S.-K., Jiang, Z., & Liu, P. (2024). GPTScore: Evaluate as You Desire. In Duh, K., Gomez, H., & Bethard, S. (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6556–6576 Mexico City, Mexico. Association for Computational Linguistics.

Fu, X., Zhang, Y., Wang, T., Liu, X., Sun, C., & Yang, Z. (2021). RepSum: Unsupervised Dialogue Summarization Based on Replacement Strategy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6042–6051. Association for Computational Linguistics.

Gan, L., Zhang, Y., Kuang, K., Yuan, L., Li, S., Sun, C., Liu, X., & Wu, F. (2021). Dialogue Inspectional Summarization with Factual Inconsistency Awareness. arXiv. arXiv:2111.03284.

Ganesh, P., & Dingliwal, S. (2019). Restructuring Conversations Using Discourse Relations for Zero-shot Abstractive Dialogue Summarization.. *arXiv: Computation and Language*.

Gao, M., Wan, X., Su, J., Wang, Z., & Huai, B. (2023). Reference Matters: Benchmarking Factual Error Correction for Dialogue Summarization with Fine-grained Evaluation Framework | Semantic Scholar.

Gao, M., & Wan, X. (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*, pp. 5693–5709. Association for Computational Linguistics.

Gao, S., Cheng, X., Li, M., Chen, X., Li, J., Zhao, D., & Yan, R. (2023). Dialogue Summarization with Static-Dynamic Structure Fusion Graph. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13858–13873. Association for Computational Linguistics.

Gatt, A., & Krahmer, E. (2018). Survey of the State of the Art in Natural Language Generation: Core Tasks, Applications and Evaluation. *Journal of Artificial Intelligence Research*, 61, 65–170. AI Access Foundation.

Geng, Z., Zhong, M., Yin, Z., Qiu, X., & Huang, X. (2022). Improving Abstractive Dialogue Summarization with Speaker-Aware Supervised Contrastive Learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6540–6546. International Committee on Computational Linguistics.

Ghadimi, A., & Beigy, H. (2022). Hybrid Multi-Document Summarization Using Pre-Trained Language Models. *Expert Systems With Applications*, 192, C. Pergamon Press, Inc.

Gliwa, B., Mochol, I., Biesek, M., & Wawer, A. (2019). SAMSum Corpus: A Human-Annotated Dialogue Dataset for Abstractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79. Association for Computational Linguistics.

Gloeckle, F., Idrissi, B. Y., Rozière, B., Lopez-Paz, D., & Synnaeve, G. (2024). Better & Faster Large Language Models via Multi-token Prediction. arXiv. arXiv:2404.19737.

Gu, J.-C., Tao, C., & Ling, Z.-H. (2022). Who Says What to Whom: A Survey of Multi-Party Conversations. In *Thirty-First International Joint Conference on Artificial Intelligence*, Vol. 6, pp. 5486–5493.

Hannousse, A. (2021). Searching Relevant Papers for Software Engineering Secondary Studies: Semantic Scholar Coverage and Identification Role. *IET Software*, 15, 1, 126–146.

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2022). Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 09, 5149–5169. IEEE Computer Society.

Hossain, M. M., Hamilton, K., Palmer, A., & Blanco, E. (2020). Predicting the Focus of Negation: Model and Error Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8389–8401. Association for Computational Linguistics.

Hu, Y., Ganter, T., Deilamsalehy, H., Dernoncourt, F., Foroosh, H., & Liu, F. (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)*, pp. 16409–16423. Association for Computational Linguistics.

Hua, Y., Deng, Z., & McKeown, K. (2023). Improving Long Dialogue Summarization with Semantic Graph Representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13851–13883. Association for Computational Linguistics.

Hua, Y., Deng, Z., & Xu, Z. (2022). AMRTVSumm: AMR-augmented Hierarchical Network for TV Transcript Summarization. In *Proceedings of the Workshop on Automatic Summarization for Creative Writing*, pp. 36–43. Association for Computational Linguistics.

Huang, K.-H., Singh, S., Ma, X., Xiao, W., Nan, F., Dingwall, N., Wang, W. Y., & McKeown, K. (2023). SWING: Balancing Coverage and Faithfulness for Dialogue Summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 512–525. Association for Computational Linguistics.

Ibrahim, R., & Shafiq, M. O. (2023). Explainable Convolutional Neural Networks: A Taxonomy, Review, and Future Directions. *ACM Computing Surveys*, 55, 10, 206:1–206:37.

Ioannides, G., Jadhav, A., Sharma, A., Navali, S., & Black, A. W. (2023). Compressed Models for Co-Reference Resolution: Enhancing Efficiency with Debiased Word Embeddings. *Scientific Reports*, 13, 1, 18510. Nature Publishing Group.

Jain, S., Keshava, V., Sathyendra, S. M., Fernandes, P., Liu, P., Neubig, G., & Zhou, C. (2023). Multi-Dimensional Evaluation of Text Summarization with In-Context Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8487–8495.

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., & Wooters, C. (2003). The ICSI Meeting Corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, Vol. 1, pp. I–I.

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a Measure of the Difficulty of Speech Recognition Tasks. *The Journal of the Acoustical Society of America*, 62, S1, S63-S63.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Chan, H. S., Dai, W., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55, 12, 1–38.

Jia, Q., Liu, Y., Tang, H., & Zhu, K. (2022a). Post-Training Dialogue Summarization Using Pseudo-Paraphrasing. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 1660–1669. Association for Computational Linguistics.

Jia, Q., Ren, S., Liu, Y., & Zhu, K. Q. (2022b). Taxonomy of Abstractive Dialogue Summarization: Scenarios, Approaches and Future Directions.

Jones, Q., Ravid, G., & Rafaeli, S. (2004). Information Overload and the Message Dynamics of Online Interaction Spaces: A Theoretical Model and Empirical Exploration. *Information Systems Research*, 15, 2, 194–210. Institute for Operations Research & the Management Sciences (INFORMS).

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving Pre-Training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8, 64–77. MIT Press.

Jung, J., Seo, H., Jung, S., Chung, R., Ryu, H., & Chang, D.-S. (2023). Interactive User Interface for Dialogue Summarization. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 934–957. ACM.

Khalifa, M., Ballesteros, M., & McKeown, K. (2021). A Bag of Tricks for Dialogue Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8014–8022. Association for Computational Linguistics.

Khurana, A., Bhatnagar, V., & Kumar, V. (2024). Personalized Summarization of Scientific Scholarly Texts. arXiv. arXiv:2306.09604.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural Language Processing: State of the Art, Current Trends and Challenges. *Multimedia Tools and Applications*, 82, 3, 3713–3744.

King, D., Shen, Z., Subramani, N., Weld, D. S., Beltagy, I., & Downey, D. (2022). Don't Say What You Don't Know: Improving the Consistency of Abstractive Summarization by Constraining Beam Search. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pp. 555–571. Association for Computational Linguistics.

Kiritchenko, S., & Mohammad, S. (2017). Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 465–470. Association for Computational Linguistics.

Kirstein, F., Ruas, T., & Gipp, B. (2024a). Is My Meeting Summary Good? Estimating Quality with a Multi-LLM Evaluator. arXiv. arXiv:2411.18444.

Kirstein, F., Ruas, T., & Gipp, B. (2024b). What's Wrong? Refining Meeting Summaries with LLM Feedback. arXiv. arXiv:2407.11919.

Kirstein, F., Ruas, T., Kratel, R., & Gipp, B. (2024c). Tell Me What I Need to Know: Exploring LLM-based (Personalized) Abstractive Multi-Source Meeting Summarization. In Dernoncourt, F., Preoţiuc-Pietro, D., & Shimorina, A. (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 920–939 Miami, Florida, US. Association for Computational Linguistics.

Kirstein, F., Wahle, J. P., Ruas, T., & Gipp, B. (2024d). What's under the Hood: Investigating Automatic Metrics on Meeting Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

Kitchenham, B. (2004). Procedures for Performing Systematic Reviews. *Keele University, Keele*.

Koay, J. J., Roustai, A., Dai, X., Burns, D., Kerrigan, A., & Liu, F. (2020). How Domain Terminology Affects Meeting Summarization Performance. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5689–5695. International Committee on Computational Linguistics.

Kocmi, T., & Federmann, C. (2023). Large Language Models Are State-of-the-Art Evaluators of Translation Quality. In Nurminen, M., Brenner, J., Koponen, M., Latomaa, S., Mikhailov, M., Schierl, F., Ranasinghe, T., Vanmassenhove, E., Vidal, S. A., Aranberri, N., Nunziatini, M., Escartín, C. P., Forcada, M., Popovic, M., Scarton, C., & Moniz, H. (Eds.), *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 193–203 Tampere, Finland. European Association for Machine Translation.

Krippendorff, K. (1970). Bivariate Agreement Coefficients for Reliability of Data. *Sociological Methodology*, 2, 139–150. [American Sociological Association, Wiley, Sage Publications, Inc.].

Kryscinski, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). Neural Text Summarization: A Critical Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 540–551. Association for Computational Linguistics.

Kryscinski, W., McCann, B., Xiong, C., & Socher, R. (2020). Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346. Association for Computational Linguistics.

Kumar, L. P., & Kabiri, A. (2022). Meeting Summarization: A Survey of the State of the Art. arXiv. arXiv:2212.08206.

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015). From Word Embeddings to Document Distances. In Bach, F. R., & Blei, D. M. (Eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, Vol. 37 of *JMLR Workshop and Conference Proceedings*, pp. 957–966. JMLR.org.

Laskar, M. T. R., Fu, X.-Y., Chen, C., & Bhushan Tn, S. (2023). Building Real-World Meeting Summarization Systems Using Large Language Models: A Practical Perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 343–352. Association for Computational Linguistics.

Lee, D., Lim, J., Whang, T., Lee, C., Cho, S., Park, M., & Lim, H. (2021a). Capturing Speaker Incorrectness: Speaker-focused Post-Correction for Abstractive Dialogue Summarization. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pp. 65–73. Association for Computational Linguistics.

Lee, H., Yun, J., Choi, H., Joe, S., & Gwon, Y. L. (2021b). Enhancing Semantic Understanding with Self-Supervised Methods for Abstractive Dialogue Summarization. In Hermansky, H., Cernocký, H., Burget, L., Lamel, L., Scharenborg, O., & Motlícek, P. (Eds.), *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pp. 796–800. ISCA.

Lee, K., Kayaalp, M., Henry, S., & Uzuner, Ö. (2021c). A Context-Enhanced De-identification System. *ACM Transactions on Computing for Healthcare*, 3, 1, 6:1–6:14.

Lee, S., Yang, K., Park, C., Sedoc, J., & Lim, H. (2021d). Who Speaks Like a Style of Vitamin: Towards Syntax-Aware Dialogue Summarization Using Multi-Task Learning. *IEEE access : practical innovations, open solutions*, 9, 168889–168898.

Lei, Y., Yan, Y., Zeng, Z., He, K., Zhang, X., & XuS, W. (2021a). Hierarchical Speaker-Aware Sequence-to-Sequence Model for Dialogue Summarization. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7823–7827. IEEE.

Lei, Y., Zheng, F., Yan, Y., He, K., & Xu, W. (2021b). A Finer-Grain Universal Dialogue Semantic Structures Based Model for Abstractive Dialogue Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1354–1364. Association for Computational Linguistics.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (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*, pp. 7871–7880. Association for Computational Linguistics.

Li, B., Wang, R., Guo, J., Song, K., Tan, X., Hassan, H., Menezes, A., Xiao, T., Bian, J., & Zhu, J. (2023). Deliberate Then Generate: Enhanced Prompting Framework for Text Generation. arXiv. arXiv:2305.19835.

Li, C., Wang, L., Lin, X., de Melo, G., & He, L. (2022). Curriculum Prompt Learning with Self-Training for Abstractive Dialogue Summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1096–1106. Association for Computational Linguistics.

Li, C., Huber, P., Xiao, W., Amblard, M., Braud, C., & Carenini, G. (2023). Discourse Structure Extraction from Pre-Trained and Fine-Tuned Language Models in Dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2562–2579. Association for Computational Linguistics.

Li, H., Xu, S., Yuan, P., Wang, Y., Wu, Y., He, X., & Zhou, B. (2021). Learn to Copy from the Copying History: Correlational Copy Network for Abstractive Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4091–4101. Association for Computational Linguistics.

Li, H. (2022). URAMDS: Utterances Relation Aware Model for Dialogue Summarization: A Combined Model for Dialogue Summarization. In *2022 2nd International Conference on Bioinformatics and Intelligent Computing*, pp. 397–401. ACM.

Li, J., Xia, Y., Cheng, X., Zhao, D., & Yan, R. (2023). Learning Disentangled Representation via Domain Adaptation for Dialogue Summarization. In *Proceedings of the ACM Web Conference 2023*, pp. 1693–1702. ACM.

Li, X. L., & Liang, P. (2021). Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597. Association for Computational Linguistics.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). DailyDialog: A Manually Labelled Multi-Turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995. Asian Federation of Natural Language Processing.

Liang, X., Wu, S., Cui, C., Bai, J., Bian, C., & Li, Z. (2023). Enhancing Dialogue Summarization with Topic-Aware Global- and Local- Level Centrality. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 27–38. Association for Computational Linguistics.

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 22 140, 55–55.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics.

Lin, H., Ma, L., Zhu, J., Xiang, L., Zhou, Y., Zhang, J., & Zong, C. (2021). CSDS: A Fine-Grained Chinese Dataset for Customer Service Dialogue Summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4436–4451. Association for Computational Linguistics.

Lin, H., Zhu, J., Xiang, L., Zhou, Y., Zhang, J., & Zong, C. (2022). Other Roles Matter! Enhancing Role-Oriented Dialogue Summarization via Role Interactions. In *Proceedings*

of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2545–2558. Association for Computational Linguistics.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015). Microsoft COCO: Common Objects in Context. arXiv. arXiv:1405.0312.

Liu, A., Sap, M., Lu, X., Swayamdipta, S., Bhagavatula, C., Smith, N. A., & Choi, Y. (2021). DExperts: Decoding-time Controlled Text Generation with Experts and Anti-Experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6691–6706. Association for Computational Linguistics.

Liu, C., Wang, P., Xu, J., Li, Z., & Ye, J. (2019). Automatic Dialogue Summary Generation for Customer Service. In Teredesai, A., Kumar, V., Li, Y., Rosales, R., Terzi, E., & Karypis, G. (Eds.), *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pp. 1957–1965. ACM.

Liu, H., & Singh, P. (2004). ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, 22, 4, 211–226.

Liu, H., Zaharia, M., & Abbeel, P. (2023). Ring Attention with Blockwise Transformers for Near-Infinite Context. arXiv. arXiv:2310.01889.

Liu, J., Zou, Y., Xi, Y., Li, S., Ma, M., & Ding, Z. (2022a). Summarizing Dialogues with Negative Cues. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6050–6056. International Committee on Computational Linguistics.

Liu, J., Zou, Y., Xi, Y., Li, S., Ma, M., Ding, Z., & Long, B. (2022b). Negative Guided Abstractive Dialogue Summarization. In *Interspeech 2022*, pp. 3253–3257. ISCA.

Liu, J., Zou, Y., Zhang, H., Chen, H., Ding, Z., Yuan, C., & Wang, X. (2021). Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1229–1243. Association for Computational Linguistics.

Liu, R., Mao, R., Luu, A. T., & Cambria, E. (2023). A Brief Survey on Recent Advances in Coreference Resolution. *Artificial Intelligence Review*, 56, 12, 14439–14481.

Liu, Y., Maynez, J., Simões, G., & Narayan, S. (2022). Data Augmentation for Low-Resource Dialogue Summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 703–710. Association for Computational Linguistics.

Liu, Y., Yu, S., & Lin, T. (2023). Hessian Regularization of Deep Neural Networks: A Novel Approach Based on Stochastic Estimators of Hessian Trace. *Neurocomputing*, 536, 13–20.

Liu, Z., & Chen, N. (2021). Controllable Neural Dialogue Summarization with Personal Named Entity Planning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 92–106. Association for Computational Linguistics.

Liu, Z., & Chen, N. (2022). Entity-Based de-Noising Modeling for Controllable Dialogue Summarization. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 407–418. Association for Computational Linguistics.

Liu, Z., Shi, K., & Chen, N. (2021). Coreference-Aware Dialogue Summarization. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 509–519. Association for Computational Linguistics.

Lu, X., Welleck, S., West, P., Jiang, L., Kasai, J., Khashabi, D., Le Bras, R., Qin, L., Yu, Y., Zellers, R., Smith, N. A., & Choi, Y. (2022). NeuroLogic A*esque Decoding: Constrained Text Generation with Lookahead Heuristics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 780–799. Association for Computational Linguistics.

Lu, Y., Bo, Y., & He, W. (2021). Confidence Adaptive Regularization for Deep Learning with Noisy Labels. arXiv. arXiv:2108.08212.

Lyu, C., Du, Z., Xu, J., Duan, Y., Wu, M., Lynn, T., Aji, A. F., Wong, D. F., & Wang, L. (2024a). A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., & Xue, N. (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 1339–1352 Torino, Italia. ELRA and ICCL.

Lyu, M., Peng, C., Li, X., Balian, P., Bian, J., & Wu, Y. (2024b). Automatic Summarization of Doctor-Patient Encounter Dialogues Using Large Language Model through Prompt Tuning. arXiv. arXiv:2403.13089.

Ma, C., Zhang, W. E., Guo, M., Wang, H., & Sheng, Q. Z. (2023). Multi-Document Summarization via Deep Learning Techniques: A Survey. *ACM Computing Surveys*, 55, 5, 1–37.

Ma, R., Zhou, X., Gui, T., Tan, Y., Li, L., Zhang, Q., & Huang, X. (2022). Template-Free Prompt Tuning for Few-Shot NER. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5721–5732. Association for Computational Linguistics.

Majumder, B. P., Li, S., Ni, J., & McAuley, J. (2020). Interview: A Large-Scale Open-Source Corpus of Media Dialog. arXiv. arXiv:2004.03090.

Mane, Khadtare, Hingmire, Pawar, & Vasal (2024). Exploring Advances in Meeting Minutes Generation and Face Attendance Systems: A Comprehensive Literature Survey. *IJSREM Journal*.

Manuvinakurike, R., Sahay, S., Chen, W., & Nachman, L. (2021). Incremental Temporal Summarization in Multi-Party Meetings. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 530–541. Association for Computational Linguistics.

McIntosh, T. R., Liu, T., Susnjak, T., Watters, P., Ng, A., & Halgamuge, M. N. (2023). A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination.

Meskill, C. (1993). ESL and Multimedia: A Study of the Dynamics of Paired Student Discourse. *System*, 21, 3, 323–341.

Misra, A., Anand, P., Fox Tree, J. E., & Walker, M. (2015). Using Summarization to Discover Argument Facets in Online Idealogical Dialog. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 430–440. Association for Computational Linguistics.

Moramarco, F., Juric, D., Savkov, A., Flann, J., Lehl, M., Boda, K., Grafen, T., Zhelezniak, V., Gohil, S., Korfiatis, A. P., & Hammerla, N. (2021). Towards More Patient Friendly Clinical Notes through Language Models and Ontologies. *AMIA Annual Symposium proceedings. AMIA Symposium*, 2021, 881–890.

Mullick, A., Bhowmick, A. K., R, R., Kokku, R., Dey, P., Goyal, P., & Ganguly, N. (2024). Long Dialog Summarization: An Analysis. arXiv. arXiv:2402.16986.

Nair, V., Schumacher, E., & Kannan, A. (2023). Generating Medically-Accurate Summaries of Patient-Provider Dialogue: A Multi-Stage Approach Using Large Language Models. In Naumann, T., Ben Abacha, A., Bethard, S., Roberts, K., & Rumshisky, A. (Eds.), *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pp. 200–217 Toronto, Canada. Association for Computational Linguistics.

Naraki, Y., Sakai, T., & Hayashi, Y. (2022). Evaluating the Effects of Embedding with Speaker Identity Information in Dialogue Summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 298–304. European Language Resources Association.

Nedoluzhko, A., Singh, M., Hledíková, M., Ghosal, T., & Bojar, O. (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*, pp. 3174–3182. European Language Resources Association.

Neto, J. L., Freitas, A. A., & Kaestner, C. A. A. (2002). Automatic Text Summarization Using a Machine Learning Approach. In Bittencourt, G., & Ramalho, G. L. (Eds.), *Advances in Artificial Intelligence*, pp. 205–215. Springer.

Ng, J.-P., & Abrecht, V. (2015). Better Summarization Evaluation with Word Embeddings for ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1925–1930. Association for Computational Linguistics.

Nguyen, T. T. H., Jatowt, A., Coustaty, M., & Doucet, A. (2021). Survey of Post-OCR Processing Approaches. *ACM Computing Surveys*, 54, 6, 124:1–124:37.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *Systematic Reviews*, 10, 1, 89.

Pagnoni, A., Fabbri, A., Kryscinski, W., & Wu, C.-S. (2023). Socratic Pretraining: Question-Driven Pretraining for Controllable Summarization. In Rogers, A., Boyd-Graber, J.,

& Okazaki, N. (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12737–12755 Toronto, Canada. Association for Computational Linguistics.

Pakhale, K. (2023). Comprehensive Overview of Named Entity Recognition: Models, Domain-Specific Applications and Challenges. arXiv. arXiv:2309.14084.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics.

Park, E. H., & Storey, V. C. (2023). Emotion Ontology Studies: A Framework for Expressing Feelings Digitally and Its Application to Sentiment Analysis. *ACM Computing Surveys*, 55, 9, 181:1–181:38.

Park, S., Shin, D., & Lee, J. (2022). Leveraging Non-Dialogue Summaries for Dialogue Summarization. In *Proceedings of the First Workshop on Transcript Understanding*, pp. 1–7. International Conference on Computational Linguistics.

Peysakhovich, A., & Lerer, A. (2023). Attention Sorting Combats Recency Bias In Long Context Language Models. arXiv. arXiv:2310.01427.

Popović, M. (2015). chrF: Character n-Gram F-score for Automatic MT Evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395. Association for Computational Linguistics.

Popović, M. (2017). chrF++: Words Helping Character n-Grams. In *Proceedings of the Second Conference on Machine Translation*, pp. 612–618. Association for Computational Linguistics.

Prodan, G., & Pelican, E. (2022). Prompt Scoring System for Dialogue Summarization Using GPT-3.

Qader, R., Jneid, K., Portet, F., & Labbé, C. (2018). Generation of Company Descriptions Using Concept-to-Text and Text-to-Text Deep Models: Dataset Collection and Systems Evaluation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pp. 254–263. Association for Computational Linguistics.

Qi, M., Liu, H., Fu, Y., & Liu, T. (2021). Improving Abstractive Dialogue Summarization with Hierarchical Pretraining and Topic Segment. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1121–1130. Association for Computational Linguistics.

Qin, C., & Joty, S. R. (2022). LFPT5: A Unified Framework for Lifelong Few-Shot Language Learning Based on Prompt Tuning of T5. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Quan, J., Xiong, D., Webber, B., & Hu, C. (2019). GECOR: An End-to-End Generative Ellipsis and Co-Reference Resolution Model for Task-Oriented Dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4547–4557. Association for Computational Linguistics.

R, V., Ramesh, D., M, H., R, V., Ramesh, D., & M, H. (2023). Automatic Text Summarization–A Systematic Literature Review. *World Journal of Advanced Engineering Technology and Sciences*, 8, 2, 126–129. World Journal of Advanced Engineering Technology and Sciences.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21, 140:1–140:67.

Rai, S., & Chakraverty, S. (2020). A Survey on Computational Metaphor Processing. *ACM Computing Surveys*, 53, 2, 24:1–24:37.

Rameshkumar, R., & Bailey, P. (2020). Storytelling with Dialogue: A Critical Role Dungeons and Dragons Dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5121–5134. Association for Computational Linguistics.

Ravaut, M., Joty, S., & Chen, N. (2022). SummaReranker: A Multi-Task Mixture-of-Experts Re-Ranking Framework for Abstractive Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4504–4524. Association for Computational Linguistics.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings Using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992. Association for Computational Linguistics.

Rennard, V., Shang, G., Hunter, J., & Vazirgiannis, M. (2023). Abstractive Meeting Summarization: A Survey. *Transactions of the Association for Computational Linguistics*, 11, 861–884. MIT Press, Cambridge, MA.

Rothe, S., Narayan, S., & Severyn, A. (2020). Leveraging Pre-Trained Checkpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics*, 8, 264–280. MIT Press.

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40, 2, 99–121.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. *Language*, 50, 4, 696–735. Linguistic Society of America.

Sai, A. B., Mohankumar, A. K., & Khapra, M. M. (2022). A Survey of Evaluation Metrics Used for NLG Systems. *ACM Computing Surveys*, 55, 2, 26:1–26:39.

Scialom, T., Dray, P.-A., Lamprier, S., Piwowarski, B., Staiano, J., Wang, A., & Gallinari, P. (2021). QuestEval: Summarization Asks for Fact-Based Evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6594–6604. Association for Computational Linguistics.

Scialom, T., Lamprier, S., Piwowarski, B., & Staiano, J. (2019). Answers Unite! Unsupervised Metrics for Reinforced Summarization Models. In *Proceedings of the 2019*

Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3246–3256. Association for Computational Linguistics.

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083. Association for Computational Linguistics.

Sellam, T., Das, D., & Parikh, A. (2020). BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892. Association for Computational Linguistics.

Sharma, A., Feldman, D., & Jain, A. (2023). Team Cadence at MEDIQA-Chat 2023: Generating, Augmenting and Summarizing Clinical Dialogue with Large Language Models. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pp. 228–235. Association for Computational Linguistics.

Shazeer, N. (2019). Fast Transformer Decoding: One Write-Head Is All You Need. arXiv. arXiv:1911.02150.

Shinde, K., Ghosal, T., Singh, M., & Bojar, O. (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*, pp. 691–702. De La Salle University.

Spinde, T., Hinterreiter, S., Haak, F., Ruas, T., Giese, H., Meuschke, N., & Gipp, B. (2024). The Media Bias Taxonomy: A Systematic Literature Review on the Forms and Automated Detection of Media Bias. arXiv. arXiv:2312.16148.

Srivastava, H., Varshney, V., Kumari, S., & Srivastava, S. (2020). A Novel Hierarchical BERT Architecture for Sarcasm Detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pp. 93–97. Association for Computational Linguistics.

Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., & Cardie, C. (2019). DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 7, 217–231. MIT Press.

Suri, K., Mishra, P., Saha, S., & Singh, A. (2023). SuryaKiran at MEDIQA-Sum 2023: Leveraging LoRA for Clinical Dialogue Summarization. arXiv.

Sznajder, B., Gunasekara, C., Lev, G., Joshi, S., Shnarch, E., & Slonim, N. (2022). Heuristic-Based Inter-training to Improve Few-shot Multi-perspective Dialog Summarization. arXiv. arXiv:2203.15590.

Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., Luccioni, S., Masoud, M., Mitchell, M., Radev, D., Sharma, S., Subramonian, A., Tae, J., Tan, S., Tunuguntla, D., & Van Der Wal, O. (2022). You Reap What You Sow: On the Challenges of Bias Evaluation under Multilingual Settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pp. 26–41. Association for Computational Linguistics.

Tan, H., Wu, H., Shao, W., Zhang, X., Zhan, M., Hou, Z., Liang, D., & Song, L. (2023). Reconstruct Before Summarize: An Efficient Two-Step Framework for Condensing

and Summarizing Meeting Transcripts. In Bouamor, H., Pino, J., & Bali, K. (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13128–13141 Singapore. Association for Computational Linguistics.

Tang, X., Nair, A., Wang, B., Wang, B., Desai, J., Wade, A., Li, H., Celikyilmaz, A., Mehdad, Y., & Radev, D. (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*, pp. 5657–5668. Association for Computational Linguistics.

Tepper, N., Hashavit, A., Barnea, M., Ronen, I., & Leiba, L. (2018). Collabot: Personalized Group Chat Summarization. In Chang, Y., Zhai, C., Liu, Y., & Maarek, Y. (Eds.), *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*, pp. 771–774. ACM.

Tuggener, D., Mieskes, M., Deriu, J., & Cieliebak, M. (2021). Are We Summarizing the Right Way? A Survey of Dialogue Summarization Data Sets. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pp. 107–118. Association for Computational Linguistics.

Vaibhav, V., Singh, S., Stewart, C., & Neubig, G. (2019). Improving Robustness of Machine Translation with Synthetic Noise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1916–1920. Association for Computational Linguistics.

Vajjala, S., & Balasubramaniam, R. (2022). What Do We Really Know about State of the Art NER?. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5983–5993. European Language Resources Association.

Vasilyev, O., Dharnidharka, V., & Bohannon, J. (2020). Fill in the BLANC: Human-free Quality Estimation of Document Summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pp. 11–20. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008.

Vedantam, R., Zitnick, C. L., & Parikh, D. (2015). CIDEr: Consensus-based Image Description Evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4566–4575. IEEE Computer Society.

Vilalta, R., & Drissi, Y. (2002). A Perspective View and Survey of Meta-Learning | Artificial Intelligence Review.

Wahle, J., Gipp, B., & Ruas, T. (2023). Paraphrase Types for Generation and Detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12148–12164. Association for Computational Linguistics.

Wahle, J. P., Ruas, T., Kirstein, F., & Gipp, B. (2022). How Large Language Models Are Transforming Machine-Paraphrase Plagiarism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 952–963. Association for Computational Linguistics.

Walker, M. A., Anand, P., Abbott, R., Tree, J. E. F., Martell, C., & King, J. (2012). That Is Your Evidence?: Classifying Stance in Online Political Debate. *Decision Support Systems*, 53, 4, 719–729.

Walton, D. N., & Krabbe, E. C. W. (1995). *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. State University of New York Press.

Wan, H., Lin, J., Du, J., Shen, D., & Zhang, M. (2021). Enhancing Metaphor Detection by Gloss-Based Interpretations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 1971–1981. Association for Computational Linguistics.

Wang, B., Zhang, C., Zhang, Y., Chen, Y., & Li, H. (2022). Analyzing and Evaluating Faithfulness in Dialogue Summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4897–4908. Association for Computational Linguistics.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S. T., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., Song, D., & Li, B. (2024). DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. arXiv. arXiv:2306.11698.

Wang, J., & Yu, B. (2021). News2PubMed: A Browser Extension for Linking Health News to Medical Literature. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pp. 2605–2609. Association for Computing Machinery.

Wang, S., Tang, L., Majety, A., Rousseau, J. F., Shih, G., Ding, Y., & Peng, Y. (2022a). Trustworthy Assertion Classification through Prompting. *Journal of Biomedical Informatics*, 132, 104139.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A. S., Arunkumar, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Pal, K. K., Patel, M., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P. R., Verma, P., Puri, R. S., Karia, R., Doshi, S., Sampat, S. K., Mishra, S., Reddy A, S., Patro, S., Dixit, T., & Shen, X. (2022b). Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5109. Association for Computational Linguistics.

Wang, Y., Tong, H., Zhu, Z., & Li, Y. (2022c). Nested Named Entity Recognition: A Survey. *ACM Transactions on Knowledge Discovery from Data*, 16, 6, 108:1–108:29.

Wang, Z., Cao, Z., & Li, W. (2023). Preserve Context Information for Extract-Generate Long-Input Summarization Framework | Proceedings of the AAAI Conference on Artificial Intelligence.

Williams, J., Tam, S., & Shen, W. (2014). Finding Good Enough: A Task-Based Evaluation of Query Biased Summarization for Cross-Language Information Retrieval. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 657–669. Association for Computational Linguistics.

Wilson, M., Petty, J., & Frank, R. (2023). How Abstract Is Linguistic Generalization in Large Language Models? Experiments with Argument Structure. *Transactions of the Association for Computational Linguistics*, 11, 1377–1395.

Wolf, Y., Wies, N., Avnery, O., Levine, Y., & Shashua, A. (2024). Fundamental Limitations of Alignment in Large Language Models. arXiv. arXiv:2304.11082.

Wu, C.-S., Liu, L., Liu, W., Stenetorp, P., & Xiong, C. (2021). Controllable Abstractive Dialogue Summarization with Sketch Supervision. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 5108–5122. Association for Computational Linguistics.

Xie, K., Yu, T., Wang, H., Wu, J., Zhao, H., Zhang, R., Mahadik, K., Nenkova, A., & Riedl, M. (2023). Few-Shot Dialogue Summarization via Skeleton-Assisted Prompt Transfer.. arXiv.

Xie, K., He, D., Zhuang, J., Lu, S., & Wang, Z. (2022). View Dialogue in 2D: A Two-stream Model in Time-speaker Perspective for Dialogue Summarization and Beyond. In Calzolari, N., Huang, C.-R., Kim, H., Pustejovsky, J., Wanner, L., Choi, K.-S., Ryu, P.-M., Chen, H.-H., Donatelli, L., Ji, H., Kurohashi, S., Paggio, P., Xue, N., Kim, S., Hahm, Y., He, Z., Lee, T. K., Santus, E., Bond, F., & Na, S.-H. (Eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6075–6088 Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xiong, C., Liu, Z., Callan, J., & Liu, T.-Y. (2018). Towards Better Text Understanding and Retrieval through Kernel Entity Salience Modeling. In Collins-Thompson, K., Mei, Q., Davison, B. D., Liu, Y., & Yilmaz, E. (Eds.), *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pp. 575–584. ACM.

Xu, R., Zhu, C., & Zeng, M. (2022). Narrate Dialogues for Better Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3565–3575. Association for Computational Linguistics.

Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination Is Inevitable: An Innate Limitation of Large Language Models. arXiv. arXiv:2401.11817.

Yang, M., Li, C., Sun, F., Zhao, Z., Shen, Y., & Wu, C. (2020). Be Relevant, Non-Redundant, and Timely: Deep Reinforcement Learning for Real-Time Event Summarization. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, the Thirty-Second Innovative Applications of Artificial Intelligence Conference,*

IAAI 2020, the Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pp. 9410–9417. AAAI Press.

Yang, Z., Wang, C., Tian, Z., Wu, W., & Li, Z. (2022). TANet: Thread-Aware Pretraining for Abstractive Conversational Summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2594–2607. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., & Garnett, R. (Eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5754–5764.

Yuan, W., Neubig, G., & Liu, P. (2021). BARTScore: Evaluating Generated Text as Text Generation. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., & Vaughan, J. W. (Eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual*, pp. 27263–27277.

Zechner, K. (2002). Automatic Summarization of Open-Domain Multiparty Dialogues in Diverse Genres. *Computational Linguistics*, 28, 4, 447–485. MIT Press.

Zhang, J., Zhao, Y., Saleh, M., & Liu, P. J. (2020a). PEGASUS: Pre-Training with Extracted Gap-Sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Vol. 119 of *Proceedings of Machine Learning Research*, pp. 11328–11339. PMLR.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020b). BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhang, Y., Liu, Y., Yang, Z., Fang, Y., Chen, Y., Radev, D., Zhu, C., Zeng, M., & Zhang, R. (2023). MACSum: Controllable Summarization with Mixed Attributes. *Transactions of the Association for Computational Linguistics*, 11, 787–803. MIT Press, Cambridge, MA.

Zhang, Y., Ni, A., Mao, Z., Wu, C. H., Zhu, C., Deb, B., Awadallah, A., Radev, D. R., & Zhang, R. (2021a). Summˆ N: A Multi-Stage Summarization Framework for Long Input Dialogues and Documents. In *Annual Meeting of the Association for Computational Linguistics*.

Zhang, Y., Ni, A., Yu, T., Zhang, R., Zhu, C., Deb, B., Celikyilmaz, A., Awadallah, A. H., & Radev, D. (2021b). An Exploratory Study on Long Dialogue Summarization: What Works and What's Next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4426–4433. Association for Computational Linguistics.

Zhang, Z., Huang, M., Zhao, Z., Ji, F., Chen, H., & Zhu, X. (2019). Memory-Augmented Dialogue Management for Task-Oriented Dialogue Systems. *ACM Transactions on Information Systems*, 37, 3, 34:1–34:30.

Zhang, Z., & Zhao, H. (2021). Advances in Multi-turn Dialogue Comprehension: A Survey. arXiv. arXiv:2103.03125.

Zhao, L., Zeng, W., Xu, W., & Guo, J. (2021a). Give the Truth: Incorporate Semantic Slot into Abstractive Dialogue Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2435–2446. Association for Computational Linguistics.

Zhao, L., Zheng, F., He, K., Zeng, W., Lei, Y., Jiang, H., Wu, W., Xu, W., Guo, J., & Meng, F. (2021b). TODSum: Task-Oriented Dialogue Summarization with State Tracking. *ArXiv*.

Zhao, L., Zheng, F., Zeng, W., He, K., Ruotong, G., Jiang, H., Wu, W., & Xu, W. (2022). *ADPL: Adversarial Prompt-based Domain Adaptation for Dialogue Summarization with Knowledge Disentanglement*.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., & Eger, S. (2019). MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 563–578. Association for Computational Linguistics.

Zhong, M., Liu, Y., Xu, Y., Zhu, C., & Zeng, M. (2022). DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, the Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 11765–11773. AAAI Press.

Zhong, M., Yin, D., Yu, T., Zaidi, A., Mutuma, M., Jha, R., Awadallah, A. H., Celikyilmaz, A., Liu, Y., Qiu, X., & Radev, D. (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*, pp. 5905–5921. Association for Computational Linguistics.

Zhou, F., Xu, X., Trajcevski, G., & Zhang, K. (2022). A Survey of Information Cascade Analysis: Models, Predictions, and Recent Advances. *ACM Computing Surveys*, 54, 2, 1–36.

Zhou, T. (2023). WHORU: Improving Abstractive Dialogue Summarization with Personal Pronoun Resolution. *Electronics*, 12, 14, 3091. Multidisciplinary Digital Publishing Institute.

Zhou, W., Li, G., Cheng, X., Liang, X., Zhu, J., Zhai, F., & Li, Z. (2023a). Multi-Stage Pre-training Enhanced by ChatGPT for Multi-Scenario Multi-Domain Dialogue Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6893–6908. Association for Computational Linguistics.

Zhou, Y., Ringeval, F., & Portet, F. (2023b). Can GPT Models Follow Human Summarization Guidelines? Evaluating ChatGPT and GPT-4 for Dialogue Summarization. arXiv. arXiv:2310.16810.

Zhu, C., Liu, Y., Mei, J., & Zeng, M. (2021). MediaSum: A Large-Scale Media Interview Dataset for Dialogue Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5927–5934. Association for Computational Linguistics.

Zhu, C., Xu, R., Zeng, M., & Huang, X. (2020). A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 194–203. Association for Computational Linguistics.

Zhu, R., Qi, J., & Lau, J. H. (2023a). Annotating and Detecting Fine-grained Factual Errors for Dialogue Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6825–6845. Association for Computational Linguistics.

Zhu, Y., Yang, X., Wu, Y., & Zhang, W. (2023b). Parameter-Efficient Fine-Tuning with Layer Pruning on Free-Text Sequence-to-Sequence Modeling. arXiv. arXiv:2305.08285.

Zou, Y., Song, K., Tan, X., Fu, Z., Zhang, Q., Li, D., & Gui, T. (2022). Towards Understanding Omission in Dialogue Summarization. arXiv.

Zou, Y., Zhu, B., Hu, X., Gui, T., & Zhang, Q. (2021). Low-Resource Dialogue Summarization with Domain-Agnostic Multi-Source Pretraining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 80–91. Association for Computational Linguistics.

Citation for this Paper

```
@article{Kirstein2025,
 author={Kirstein, Frederic and Wahle, Jan Philip and Ruas, Terry and Gipp, Bela},
 title={A Literature Review on the Challenges of Abstractive Dialogue Summarization CADS: A
Systematic Literature Review on the Challenges of Abstractive Dialogue Summarization},
 journal={Journal of Artificial Intelligence Research},
 topic={nlp},
 year={2025},
 month={01}
}
```