From Speech to Summary: A Comprehensive Survey of Speech Summarization

Fabian Retkowski[1] **Maike Züfle**[1] **Andreas Sudmann**[2] **Dinah Pfau**[3]
Jan Niehues[1] **Alexander Waibel**[1,4]
[1]KIT [2]University Bonn [3]Deutsches Museum [4]CMU
{retkowski,zuefle,niehues,waibel}@kit.edu
asudmann@uni-bonn.de d.pfau@deutsches-museum.de

Abstract

Speech summarization has become an essential tool for efficiently managing and accessing the growing volume of spoken and audiovisual content. However, despite its increasing importance, speech summarization is still not clearly defined and intersects with several research areas, including speech recognition, text summarization, and specific applications like meeting summarization. This survey not only examines existing datasets and evaluation methodologies, which are crucial for assessing the effectiveness of summarization approaches but also synthesizes recent developments in the field, highlighting the shift from traditional systems to advanced models like fine-tuned cascaded architectures and end-to-end solutions.

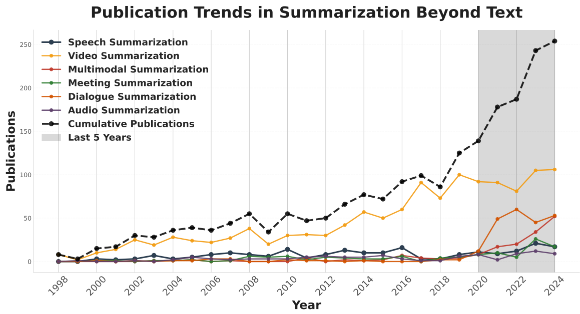

Figure 1: Publication trends in summarization beyond text, based on search results from dblp.org, showing significant growth and evolving research focus.

1 Introduction

The digital age is increasingly shaped by the high volume of spoken and audiovisual content, diverging from text-centric origins. Podcasts now number in the millions, with over 500 million global listeners and more than a million new episodes released in just two months of 2020 (Litterer et al., 2024). Platforms like YouTube and TikTok receive hundreds of thousands of hours of video every minute, a flood of content growing exponentially since the early 2000s and far outpacing human attention and capacity (Ceci, 2024). Meanwhile, everyday communication is shifting from text to voice, with users sending over 7 billion voice messages daily via apps like WhatsApp (WhatsApp, 2022).

But as audiovisual content becomes central to both media consumption and daily communication, the resulting overload of speech data creates challenges for access, navigation, and comprehension (Ghosal et al., 2022). In response, *speech summarization* (SSum) has emerged as a crucial way of making spoken content more manageable, enabling quicker information access, aiding research, and supporting everyday use across personal and professional contexts (Murray et al., 2010; Li et al., 2021). Yet despite its growing relevance, SSum remains surprisingly underdefined as a field (Rezazadegan et al., 2020; Ghosal et al., 2022). SSum occupies a unique interdisciplinary position that has not been fully defined or explored. Figure 1 reveals an interesting tension in the field: while publication counts are modest compared to video summarization, SSum exists at the intersection of multiple thriving research areas, including *automatic speech recognition* (ASR), *text summarization* (TSum), and domain-specific applications like *meeting summarization*. This is also evident in the distribution of publications across different venues (Figure 4 in Appendix C). This ambiguity in definition is both a challenge and an opportunity. SSum is not merely the application of TSum to ASR output, nor is it simply the audio component of video summarization. It requires addressing distinctive challenges, including disfluencies, prosody, speaker dynamics, and contextual elements (Zhu et al., 2020; Song et al., 2022a; Sharma et al., 2024b). The field's fragmentation across different research communities has led to parallel developments that would benefit from unification. From meeting summarization (Rennard et al., 2023a) to podcast summarization

(Jones et al., 2020) to multimodal summarization (Jangra et al., 2023), all tackle speech content but often operate in isolation, using different methodologies and benchmarks. This creates a critical need for survey work that brings these interconnected domains together.

1.1 Historical Context

In the 20th century, advances in telecommunications, military research, and information technology laid the foundations for speech processing. While early summarization efforts focused on textual data (Luhn, 1958), the challenge of summarizing speech gained prominence later. ASR began to mature in the 1980s and 1990s, particularly through statistical methods based on Markov models (Baum et al., 1970; Jelinek, 1976; Rabiner, 1989) and connectionist models (Waibel et al., 1989; Franzini et al., 1990; Renals et al., 1994), laying the groundwork for processing speech. In the 1990s, data-driven methods increasingly linked ASR and natural language processing (NLP), with early projects highlighting the potential of summarization for large-scale spoken content and identifying challenges specific to spontaneous speech, such as disfluencies, hesitations, and ASR errors through corpora like Switchboard (Godfrey et al., 1992) and programs like TIPSTER (Suhm, 1994; Zeppenfeld et al., 1997; Gee, 1998). Around 2000, research on SSum gained traction, initially adapting TSum via extractive methods for challenges like telephone dialogues (Zechner and Waibel, 2000a; McKeown et al., 2005) and broadcast news (Hori et al., 2002, 2003), selecting salient segments. Concurrently, early multimodal approaches were explored for complex meeting interactions (Bett et al., 2000; Gross et al., 2000) culminating in the development of rich, annotated corpora such as AMI (Carletta et al., 2006) and ICSI (Janin et al., 2003), foundational for meeting summarization. By the mid-2000s, extractive systems increasingly relied on features specific to speech, including prosody, speaker activity, and dialog acts (Koumpis and Renals, 2005; Maskey and Hirschberg, 2005; Murray et al., 2005). Early work raised questions about how to evaluate summaries of spoken language in the presence of ASR errors and disfluencies (Whittaker et al., 1999; Zechner and Waibel, 2000b). In subsequent years, evaluation became standardized through ROUGE (Lin, 2004). Finally, early steps toward abstractive SSum also emerged through sentence compression techniques (Hori et al., 2003).

1.2 Scope of the Survey

This survey provides a synthesis of the evolving landscape of SSum, bridging fragmented developments across ASR, TSum, dialogue summarization, and multimodal applications. The last survey of the field by Rezazadegan et al. (2020) captured pre-2020 approaches, largely based on traditional pipelines and early neural models. Since then, the field has shifted: cascaded systems now leverage fine-tuned encoder-decoder (ED) models, prompting or adapting LLMs has become common, and end-to-end (E2E) models are increasingly explored. Unlike prior surveys on meeting (Rennard et al., 2023a), dialogue (Tuggener et al., 2021; Kirstein et al., 2025a), text (Gambhir and Gupta, 2017; El-Kassas et al., 2021; Retkowski, 2023), and multimodal summarization (Jangra et al., 2023), this work focuses specifically on *spoken language* as input. We bring together diverse application domains while clearly delineating the scope of SSum from neighboring fields like video summarization.

2 Challenges of Speech Processing

Orality and Linguistic Variability. Unlike written text, spoken language lacks structural markers such as punctuation, headings, or paragraph breaks (Rehbein et al., 2020), making it harder to detect topical shifts and organize content (Zechner and Waibel, 2000a; Khalifa et al., 2021). Furthermore, speech often includes disfluencies and false starts (Khalifa et al., 2021; Kirstein et al., 2024b) and features accents, dialects, and code-switching (Keswani and Celis, 2021), all of which add complexity. Prosodic features like intonation, rhythm, and emphasis also carry meaning (Aldeneh et al., 2021) but are often lost in ASR-based pipelines. Finally, speech is often lengthy, unstructured, and semantically sparse, with important information scattered across speaker turns and interleaved with filler or redundant speech, making long-context modeling critical (Liu et al., 2019).

Acoustic Environment. External acoustic factors such as overlapping speakers or background noise (e.g., applause or sound effects) are common in spoken content. These factors can either contribute valuable context or introduce noise (Jiménez et al., 2020), posing challenges for systems that risk discarding useful cues or being disrupted by extraneous sounds (Cornell et al., 2023).

Modality Constraints. SSum presents notable technical challenges. First, real-world speech (e.g., meetings, lectures) often spans long durations, which can strain memory and processing resources (Kumar and Kabiri, 2022). Second, many pipelines rely on ASR, and transcription errors introduce noise into downstream processing (Rennard et al., 2023b; Chowdhury et al., 2024).

3 Problem Formulation

3.1 Speech Summarization

Speech summarization is the process of condensing spoken content into a shorter version while preserving essential information. It is most commonly understood as a *cross-modal* task, where an audio signal (speech) is transformed into a textual summary (*speech-to-text summarization*). However, it is often implemented as a *cascaded* approach, where an ASR system first transcribes the speech into text, followed by *unimodal text summarization* systems. Alternatively, the input may be a manually created transcript, in which case the summarization remains a form of speech summarization but is entirely text-based. The output can be either abstractive, where the summary is generated in a rephrased form, or extractive, where key sentences or phrases are directly taken from the original speech. Summarization can be performed at different granularities, such as sentence-level, segment-level, or document-level.

3.2 Input Data Modalities

The input can take the form of raw audio or transcripts, either generated via ASR or created by humans. Notably, the choice of input modality significantly impacts summary quality: direct use of speech can yield more selective and factually consistent summaries (Sharma et al., 2024b). However, incorporating speech-specific features such as prosody or speaker information in addition to the transcript has also been shown to improve the quality of summaries (Inoue et al., 2004). For cascaded systems, the quality of ASR transcripts remains a limiting factor, with clear performance gaps compared to manual transcripts (Kano et al., 2021).

3.3 Applications and Related Tasks

3.3.1 Core Applications

A core application of speech summarization is *meeting summarization*, condensing free-form discussions into concise overviews, which can range from high-level summaries (Janin et al., 2003; Carletta et al., 2006) to more structured outputs like meeting minutes (Nedoluzhko et al., 2022; Hu et al., 2023a) or action item lists (Purver et al., 2007; Mullenbach et al., 2021; Asthana et al., 2024), blurring the lines between summarization and structured information logging (Tuggener et al., 2021). More broadly, this falls under the umbrella of *dialogue summarization*, which includes not only spoken interactions such as meetings, customer service calls, and interviews but also text-based dialogues like chat transcripts. Other prominent application domains include *podcast summarization* (Clifton et al., 2020; Song et al., 2022a) and *presentation summarization*, which focuses on structured, monologic content such as lectures (Miller, 2019; Lv et al., 2021; Xie et al., 2025), TED Talks (Kano et al., 2021; Shon et al., 2023), and conference presentations (Züfle et al., 2025). A further core area is *YouTube video summarization*, which has emerged as a major testbed for SSum systems (Sanabria et al., 2018; Retkowski and Waibel, 2024a; Qiu et al., 2024). It encompasses a wide variety of content types, ranging from educational videos to interviews, vlogs, and news broadcasts, and poses unique challenges due to its diversity.

3.3.2 Related Tasks

Smart Chaptering. Many speech summarization applications benefit from *smart chaptering* (or topic segmentation), where spoken content is divided into coherent sections. This approach enables more granular summarization at the chapter level, while the chapter titles function as extreme summaries (Zechner and Waibel, 2000a; Banerjee et al., 2015; Ghazimatin et al., 2024; Retkowski and Waibel, 2024a; Xie et al., 2025).

Subtitle Compression. At an even finer granularity, *sentence-wise speech summarization* (Matsuura et al., 2024) focuses on condensing individual spoken sentences into more concise forms. This task is particularly relevant to *subtitle compression*, where subtitles may initially be transcriptions or translations of speech that are too long to fit on screen or to be read comfortably by viewers. The task of subtitle compression addresses this by automatically shortening subtitle text while preserving its meaning (Liu et al., 2020; Papi et al., 2023a; Retkowski and Waibel, 2024b; Jørgensen and Mengshoel, 2025).

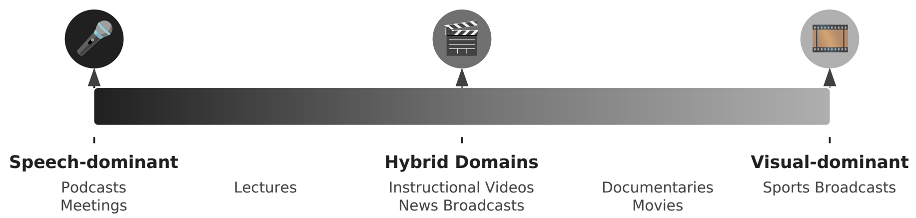

Figure 2: The Speech-Video Modality Importance Spectrum

Audio Captioning. Automated *audio captioning* (Mei et al., 2022), though not traditionally framed as summarization, shares conceptual similarities with SSum as it generates textual descriptions of audio scenes. Its focus is on environmental sounds, treating speech as just another acoustic event.

3.3.3 Additional Input Modalities

The Value of Visual Cues. Speech summarization inherently extends into *multimodal summarization* as speech is frequently embedded within environments rich with complementary visual and contextual information. As such, many datasets used in SSum, such as How2 (Sanabria et al., 2018) or AMI (Carletta et al., 2006), provide not only audio but also video. Multimodal information has been shown to provide significant value to many SSum systems. For example, incorporating modalities beyond text or audio has been demonstrated to enhance summarization of instructional videos (Palaskar et al., 2019; Khullar and Arora, 2020) while non-verbal cues like eye gaze, speaker focus, and head orientation improve meeting summarization (Nihei et al., 2018; Li et al., 2019).

The Continuum Between Speech and Video Summarization. This connection highlights a spectrum between SSum and *video summarization* (visualized in Figure 2). While speech-focused approaches treat visuals as complementary, true video summarization considers visual elements essential rather than supplementary. Different domains fall along this continuum: podcasts and meetings represent speech-dominant contexts where non-verbal cues primarily contextualize speech, while sports broadcasts and action-rich movies sit at the visual-dominant end where visual composition and action sequences carry critical narrative information.

4 Data Resources

Table 1 presents datasets relevant to speech summarization and related tasks. Given the scarcity of dedicated SSum datasets with true summaries, we also include datasets that rely on surrogate summaries (discussed below) as well as text-to-text summarization datasets if they are based on spoken content or closely resemble speech in structure and style. Subtitle compression serves as a fine-grained form of summarization, while segmentation can involve either segment-level summaries or extreme summarization, such as generating short titles.

Limitations of Surrogate Summaries. Many SSum datasets rely on what we term *surrogate summaries* such as creator descriptions (e.g., from YouTube videos and podcast episodes; Sanabria et al. 2018; Clifton et al. 2020), or paper abstracts (Liu et al., 2025; Züfle et al., 2025). However, while these summaries provide a convenient source of training data, they were not originally designed as true summaries, leading to several limitations. First, surrogate summaries are often of poor quality because they typically serve a different purpose: descriptions function as teasers, abstracts follow distinct stylistic conventions. Manakul and Gales (2022) highlights this issue by evaluating the quality of creator-provided descriptions in the Spotify Podcast Dataset, finding that 26.3% were rated as “Bad”, while only 15.6% were considered “Excellent”. Tellingly, automatic systems outperformed the original descriptions in quality (Manakul and Gales, 2020). Second, surrogate summaries may contain information not present in the original speech. Züfle et al. (2025) found that while 70.0% of paper abstracts were considered good summaries, 63.3% included content absent from the talk. Likewise, in the SummScreen dataset, TV recaps incorporate visual context (actions, settings) missing from the transcript, leading to potential content mismatches and model hallucinations (Chen et al., 2022).

Scarcity of Datasets. Our overview illustrates that the field is characterized by inconsistent benchmarks, a lack of high-quality, large-scale datasets, and a landscape of fragmented, interrelated tasks and problems rarely contextualized in the broader

Dataset	Reference	Domain	Lang.	Size	Summary Type	Transcript	Audio	Video	License
⚠[c]How2	Sanabria et al. (2018)	Instructional videos (YouTube)	EN	80k videos (2k hours)	Abstractive (video descriptions)	Manual	↓[a]	↓[a]	CC-BY-SA-4.0
YTSeg	Retkowski and Waibel (2024a)	YouTube videos (various types/topics)	EN	19.3k videos (6.5k hours)	Abstractive (segment-based, chapter titles)	Manual	✓	↓[a]	CC-BY-NC-SA-4.0
MMSum	Qiu et al. (2024)	YouTube videos (various types/topics)	EN	5.1k videos (1.2k hours)	Abstractive (segment-based, chapter titles, thumbnails)	Manual	↓[a]	↓[a]	CC-BY-NC-SA
VT-SSum	Lv et al. (2021)	Lecture videos (VideoLectures.net)	EN	9.6k videos	Abstractive (segment-based, slide text)	ASR	↓[a]	↓[a]	CC-BY-NC-ND-4.0
NUTSHELL	Züfle et al. (2025)	Conference talks (*ACL talks)	EN	6.3k talks (1.2k hours)	Abstractive (paper abstracts)	✗	✓	↓[a]	CC-BY-4.0
VISTA	Liu et al. (2025)	Conference talks (AI venues)	EN	18.6k talks (2.1k hours)	Abstractive (paper abstracts)	✗	↓[a]	↓[a]	?[h]
SLUE-TED	Shon et al. (2023)	TED talks	EN	4.2k talks (829 hours)	Abstractive (talk descriptions)	Manual	✓	↓[a]	CC-BY-NC-ND-4.0
⚠[d]TEDSummary	Kano et al. (2021)	TED talks	EN	1.5k talks	Abstractive (talk descriptions)	Manual	↓[a]	↓[a]	?[h]
⚠[e]TED Talk Teasers	Vico and Niehues (2022)	TED talks	EN	2.8k talks (739 hours)	Abstractive (talk descriptions)	Manual	↓[a]	↓[a]	CC-BY-NC-ND-4.0
StreamHover	Cho et al. (2021)	Livestreams (Behance.net)	EN	370 videos (500 hours)	Abstractive & Extractive (crowdsourced, clip-level & video-level)	ASR	↓[a]	↓[a]	?[h]
MediaSum	Zhu et al. (2021)	Media interviews (NPR, CNN)	EN	463.6k interview segments	Abstractive (topic descriptions)	Manual	✗	✗	?[h]
SummScreen	Chen et al. (2022)	TV show transcripts	EN	26k episodes	Abstractive (episode recaps)	Manual	✗	✗	?[h]
⚠[f]Spotify Podcast Dataset	Clifton et al. (2020); Garmash et al. (2023)	Podcast episodes	EN, PT	200k episodes (100k hours)	Abstractive (podcast descriptions)	ASR	✓	✗	?[h]
AMI Meeting Corpus	Carletta et al. (2006)	Business meetings (scenario-driven)	EN	137 meetings (65 hours)	Abstractive & Extractive (minutes), Topic segments	Manual	✓	✓	CC-BY-4.0
ICSI Meeting Corpus	Janin et al. (2003)	Research group meetings (naturalistic)	EN	75 meetings (72 hours)	Abstractive & Extractive (minutes), Topic segments	Manual	✓	✗	CC-BY-4.0
QMSum	Zhong et al. (2021)	AMI, ICSI & Committee meetings	EN	232 meetings	Abstractive (query-based, multiple), Topic segments	Manual	✗	✗	MIT
ELITR Minuting Corpus	Nedoluzhko et al. (2022)	Technical project & parliament meetings (naturalistic)	EN, CS	166 meetings (160 hours)	Abstractive (minutes, multiple)	Manual	✗	✗	CC-BY-NC-SA-4.0
DialogSum	Chen et al. (2021)	Diverse, spoken dialogues (EN-practicing scenarios)	EN	13.4k dialogues	Abstractive (crowdsourced)	Manual	✗	✗	CC-BY-NC-SA-4.0
MeetingBank	Hu et al. (2023a)	City council meetings (naturalistic)	EN	1.3k meetings (3.5k hours)	Abstractive (segment-level minutes)	ASR	✓	✗	CC-BY-NC-ND-4.0
◐[h]EuroParlMin	Ghosal et al. (2023)	Parliament meetings (naturalistic)	EN	2.2k sessions (1.8k hours)	Abstractive (minutes)	Manual	✗	✗	?[h]
◐[h]EuroParl Interviews	Papi et al. (2023b)	Parliament meetings (naturalistic)	EN	12 videos (1 hour)	Abstractive (sentence-level, cross-lingual)	Manual	✓	✓	CC-BY-NC-4.0
ECTSum	Mukherjee et al. (2022)	Earnings calls (The Motley Fool)	EN	2.4k transcripts	Abstractive (bullet points, from Reuters)	Manual	✗	✗	GPL-3.0
MegaSSum	Matsuura et al. (2024)	News articles (Gigaword, DUC2003)	EN	3.8M articles	Abstractive (headlines)	N/A (Articles)	≈[b]	✗	CC-BY-4.0

[a] ↓ Only a download script or source links are provided, but no direct data.
[b] ≈ Data is synthesized rather than from real recordings.
[c] ⚠ Unavailable since 12/2024 due to widespread video removals; no redistribution.
[d] ⚠ Lacks documentation on included talks, hindering reproduction (Shon et al., 2023).
[e] ⚠ Reproduction hindered; lacking documentation and TED is no longer using Amara.
[f] ⚠ Unavailable since 12/2023 due to resource constraints.
[g] ◐ Not all data partitions are available (only test set or no test set).
[h] ? No explicit license has been provided.

Table 1: English and multilingual datasets related to the SSum task. Datasets that are exclusively non-English, chat-based datasets, and derivatives or extensions of existing resources are listed in the appendix (Tables 3, 4, and 5).

field. This issue is further exacerbated by the fact that two of the most popular and largest datasets, namely How2 and the Spotify Podcast Dataset, are no longer publicly available to researchers.

Synthethic Data. A promising approach to address or mitigate data volume limitations is synthesizing data, as illustrated in recent research. For example, in the context of speech summarization, several works (Matsuura et al., 2023b, 2024; Eom et al., 2025) use a TTS system to generate synthetic speech input data from text, while LLMs can be leveraged to generate reference summaries (Jung et al., 2024; Le-Duc et al., 2024; Eom et al., 2025). Taking this approach further, LLMs have been leveraged to produce entire multi-party social conversations that achieve quality close to human-generated data (Chen et al., 2023; Suresh et al., 2025). Additionally, LLMs have been employed to synthesize ASR errors, improving the robustness of summarization models (Binici et al., 2025), while traditional audio data augmentation, such as adding background noise or reverberation, remains valuable for end-to-end speech summarization.

Out-of-Domain Data. Another strategy to overcome limited in-domain data is cross-domain pretraining, where models are first trained on large-scale text-based summarization datasets such as CNN/DailyMail, XSum, or SAMSum. These corpora help models acquire general summarization abilities before being fine-tuned on speech-specific datasets. This approach has been shown to improve performance on diverse speech summarization benchmarks, including long meeting summarization (Zhu et al., 2020; Zhang et al., 2021).

Recommended Resources. Given the limitations of current benchmarks, including the unavailability of widely used datasets and the small scale of others such as AMI, there is a clear need for viable alternatives. Among the available datasets, several stand out for their combination of *accessible audio* and *considerable scale*. SLUE-TED, NUTSHELL and VISTA offer high-quality speech aligned with abstractive summaries, based on TED talks and AI conference presentations. YTSeg, while using chapter titles as summaries, provides large-scale, manually transcribed YouTube content and is particularly well suited for long-context and structure-aware SSum. MeetingBank complements these with long-form meetings and segment-level summaries. Several other datasets in Table 1 are also promising, especially when paired with synthetic speech via TTS to compensate for the lack of audio.

5 Evaluation of Speech Summaries

Accurately evaluating SSum systems is crucial for measuring progress and ensuring reliable outputs, yet it remains challenging. First, there is no definitive ground truth for summaries, as humans emphasize different aspects and phrase information variably (Rath et al., 1961; Harman and Over, 2004; Clark et al., 2021; Cohan et al., 2022; Sharma, 2024; Zhang et al., 2024b). Second, evaluators struggle with multi-sentence summaries as their length and varied wording makes evaluation difficult (Goyal et al., 2023; Mastropaolo et al., 2023). Lastly, evaluating quality requires assessing lexical, semantic, and factual correctness (Liu et al., 2023a; Kroll and Kraus, 2024; Sharma, 2024), which makes evaluation complex. Even with reference comparisons, human evaluations are often inconsistent (Hardy et al., 2019).

While TSum evaluation already presents challenges, evaluating SSum introduces additional complexities due to the characteristics of spoken language. Colloquialisms, background noise, and multiple speakers introduce unique errors (Kirstein et al., 2024b), such as speaker misidentification affecting pronoun usage (Rennard et al., 2023b). Cascaded models further propagate transcription errors into summarization (Zechner and Waibel, 2000b; Rennard et al., 2023b; Chowdhury et al., 2024). However, current evaluation methods for SSum remain grounded in TSum approaches, which may overlook the distinct challenges of spoken content.

Evaluation methods range from human assessment to automated metrics like lexical overlap and model-based evaluation. Figure 3 in Appendix B illustrates the use of these metrics over time, showing the growing popularity of LLM-based and trained evaluator metrics over lexical overlap metrics. In the following, we discuss these metrics.

5.1 Human Evaluation.

Human evaluation is often the gold standard for summarization quality (Clark et al., 2021) but comes with challenges such as requiring extensive annotations (Card et al., 2020), being slow and costly, and lacking a standardized procedure despite multiple proposed frameworks (Nenkova and Passonneau, 2004; Hardy et al., 2019; Liu et al., 2023b; Kroll and Kraus, 2024), which complicates large-scale assessments (Iskender et al., 2020b).

Moreover, costly expert judgments are often required for high-quality evaluations (Gillick and Liu, 2010). Crowdsourcing offers a cheaper alternative, and with the right guidelines, crowd workers can match expert performance (Iskender et al., 2020b,a). However, crowdsourced evaluations tend to be more uniform and struggle with error identification (Fabbri et al., 2021). Evaluations can then be conducted either referenceless (Song et al., 2022b; Goyal et al., 2023; Schneider et al., 2025), or with references (Fabbri et al., 2021; Züfle et al., 2025). However, these setups often lack correlation (Liu et al., 2023b), making their results incomparable. Different human evaluation protocols for SSum can be found in Table 7 in Appendix B.

5.2 Lexical Overlap Metrics.

Lexical overlap metrics assess similarity based on shared surface-level units. ROUGE (Lin, 2004), designed to maximize recall, is the most widely used metric (Fabbri et al., 2021; Sharma, 2024), though implementation errors have led to incorrect evaluations in the past (Deutsch and Roth, 2020). BLEU (Papineni et al., 2002; Post, 2018) and METEOR (Banerjee and Lavie, 2005), remain common despite being developed for machine translation. Methods like BEs (Hovy et al., 2006) and the Pyramid Method (Nenkova and Passonneau, 2004) improve overlap metrics by also considering syntactic dependencies and content units.

Despite their efficiency, these metrics struggle with consistency assessment (Bhandari et al., 2020; Maynez et al., 2020; Wang et al., 2020), fail to distinguish similar or high-scoring candidates (Peyrard, 2019; Bhandari et al., 2020) and are often outperformed by model-based evaluators (Gao and Wan, 2022).

5.3 Model-Based Evaluators

Embedding-Based Metrics. Embedding-based metrics capture semantic similarity through sentence or token embeddings. Yet, they still struggle to assess factual accuracy, fully capture shared information (Deutsch and Roth, 2021), and distinguish similar candidates (Bhandari et al., 2020).

BERTScore (Zhang et al., 2020b), one of the most prominent embedding-based metrics, compares contextualized token embeddings between the summary and reference. MoverScore (Zhao et al., 2019a) measures the earth-mover distance between embeddings, assessing both shared content and divergence. SPEEDScore (Akula and Garibay, 2022) evaluates summary efficiency by balancing compression with information retention through sentence-level embeddings.

Trained Evaluators. Recent approaches have focused on training models for more holistic summary evaluation (Yuan et al., 2021; Zhong et al., 2022b), as well as for specific dimensions like factual accuracy (Kryscinski et al., 2020; Wang et al., 2020; Durmus et al., 2020; Scialom et al., 2021). Other models refine evaluations using counterfactual estimation (Xie et al., 2021) and causal graphs (Ling et al., 2025). However, even evaluation-specific models, particularly reference-free ones, may be prone to spurious correlations such as summary length (Durmus et al., 2022).

LLM-as-a-Judge. Using LLMs as evaluators is an emerging approach where models are prompted to assess summaries directly (Shen et al., 2023; Liu et al., 2023a; Zheng et al., 2024; Gong et al., 2024; Kirstein et al., 2025b). These models are applied by calculating win rates against reference models (Dubois et al., 2023, 2024), evaluating specific criteria (Liu et al., 2023a; Tang et al., 2024; Züfle et al., 2025), and performing reference-free quality estimation (Liu et al., 2023a; Gong et al., 2024; Kirstein et al., 2025b), see Table 6 in the appendix for an overview of these approaches.

LLM-as-a-Judge has shown strong performance, often surpassing traditional metrics like ROUGE (Shen et al., 2023) and, in some cases, aligning closely with human judgments (Dubois et al., 2023). However, they come with limitations: The judge model must be stronger than the systems it assesses (Dubois et al., 2023), often involving commercial models with limited reproducibility (Barnes et al., 2025). LLM judges also exhibit biases, such as favoring outputs from the same model (Dubois et al., 2023; Gong et al., 2024), struggling with factual error detection (Gong et al., 2024; Tang et al., 2024), preferring list-style over fluent text (Dubois et al., 2023), and being sensitive to prompt complexity (Thakur et al., 2025) and summary length (Dubois et al., 2024; Thakur et al., 2025). They also have difficulty distinguishing similar candidates (Shen et al., 2023) and suffer from position bias, where earlier outputs receive higher scores (Wang et al., 2024; Dubois et al., 2023).

6 Approaches

6.1 Cascaded Approaches

Cascaded approaches remain the most widely adopted paradigm in SSum. In this framework, speech is first transcribed using an ASR system and then passed to a TSum model. Two primary strategies have emerged in this paradigm: first, fine-tuning of ED models specifically for summarization, and second, prompting and adapting LLMs.

6.1.1 Fine-Tuning Encoder-Decoder Models

To enable cascaded approaches for SSum, many works focused on fine-tuning pretrained ED models such as BART, Longformer/LED, PEGASUS, DialogLM, and HMNet (e.g., Zhong et al., 2021; Hu et al., 2023b; Huang et al., 2023; Fu et al., 2024; Le-Duc et al., 2024; Zhu et al., 2025), ranging from general-purpose models such as BART and Longformer/LED to more specialized models. PEGASUS (Zhang et al., 2020a), for example, incorporates a summarization-specific pretraining using *gap sentences generation* while DialogLM/DialogLED (Zhong et al., 2022a) is trained on denoising with dialogue-inspired noise.

Handling Long Context. Long input is a particular concern for SSum, as spoken content often yields lengthy, unstructured transcripts with dispersed information. As such, many works rely on Longformer (Beltagy et al., 2020) or explore alternative sparse or windowed attention mechanisms (Zhang et al., 2021; Zhong et al., 2022a). Alternatively, hierarchical encoders (Zhu et al., 2020; Zhang et al., 2021), retrieve-then-summarize or locate-then-summarize strategies (Zhang et al., 2021; Zhong et al., 2021), and segment-level processing (Laskar et al., 2023; Retkowski and Waibel, 2024a) have been applied.

Robustness and Faithfulness. Faithfulness is a central challenge in summarization and is particularly problematic in cascaded SSum due to ASR error propagation. To improve robustness, some approaches fuse multiple ASR hypotheses (Xie and Liu, 2010; Kano et al., 2021) or ground summary segments to the transcript (Song et al., 2022a). To enhance faithfulness, other works apply symbolic knowledge distillation (Zhu et al., 2025) or incorporate fine-grained entailment signals during training (Huang et al., 2023).

Contextual and Multimodal Enrichment. Some approaches enrich SSum models with additional contextual or multimodal signals, such as speaker-role information (Zhu et al., 2020), video features combined with transcripts (Palaskar et al., 2019), or joint representations of text, video, and speech concepts (Palaskar et al., 2021).

6.1.2 Prompting and Adapting LLMs

More recently, LLMs have enabled zero-shot SSum through prompting without the need for task-specific training. This capability has been explored on various models such as GPT-3.5, PaLM-2, and LLaMA 3 (Hu et al., 2023b; Fu et al., 2024; Nelson et al., 2024; Züfle et al., 2025). Building on this, several studies propose more sophisticated prompting strategies, including few-shot prompting and iterative self-refinement (Laskar et al., 2023; Kirstein et al., 2025c). To improve performance and efficiency, methods such as LoRA fine-tuning for SSum-specific adaptation (Nelson et al., 2024) and knowledge distillation into smaller models (Fu et al., 2024; Zhu et al., 2025) have been applied.

6.2 End-to-End Approaches

E2E SSum has recently gained significant traction as a research area, with models that directly map raw audio to textual summaries without relying on an intermediate transcription. They fall broadly into two categories: task-specific architectures designed and trained directly for SSum, and modular systems that integrate LLMs with audio encoders via projection mechanisms.

6.2.1 Task-Specific Models

These models often follow a two-stage training paradigm: first, a pretraining on ASR tasks to learn the mapping from speech to text and to acquire rich acoustic-linguistic representations, followed by summarization fine-tuning (e.g., Chen et al., 2024; Eom et al., 2025). However, in contrast to other speech-processing tasks like ASR, SSum effectively demands the full context of the document. This poses a challenge for the original Transformer architecture, whose self-attention mechanism scales quadratically with input length, making it inefficient for long sequences. To overcome this, researchers typically rely on input speech truncation (Matsuura et al., 2023b; Sharma et al., 2023; Chen et al., 2024) or input compression such as temporal downsampling (Chu et al., 2024; Kang and Roy, 2024) or higher-level/segment-level projections (Shang et al., 2024). Others have explored more fundamental architectural modifications, including adjusting the attention mechanism (Sharma et al., 2022, 2023, 2024a) or replacing it entirely with more efficient structures such as FNet (Kano et al., 2023b; Chen et al., 2024), convolutions (Chen et al., 2024), or state-space models like Mamba (Miyazaki et al., 2024; Eom et al., 2025).

6.2.2 LLM-Based Systems

In parallel, efforts to leverage pretrained language models have gained momentum: earlier work explored transfer learning from ED models like BART (Matsuura et al., 2023a), while more recent approaches focus on directly integrating pretrained LLMs by attaching an *audio encoder*. As shown in Table 2, these methods typically pair an audio encoder—such as Conformer (Fathullah et al., 2024; Shang et al., 2024; Microsoft et al., 2025), HuBERT (Kang and Roy, 2024; Züfle et al., 2025), or Whisper (Chu et al., 2024; Eom et al., 2025; He et al., 2025)—with a *projection module* such as a Q-Former (Shang et al., 2024; Züfle et al., 2025), MLP (He et al., 2025; Microsoft et al., 2025), or linear layer (Fathullah et al., 2024; Kang and Roy, 2024) that maps audio features into the LLM's input space. These configurations differ in how much or which part of the system is trained. While all approaches train a projection module, they vary in whether they also fine-tune the audio encoder or the LLM. Some methods keep both components frozen,

Reference	Audio Encoder	Projector	LLM
Fathullah et al. (2024)	🔥 Conformer (Gulati et al., 2020)	🔥 Linear	❄ LLaMA-2-7B-chat (Touvron et al., 2023)
Shang et al. (2024)	🔥 Conformer (Gulati et al., 2020)	🔥 Q-Former (Li et al., 2023)	≈ LLaMA-2-7B-chat (Touvron et al., 2023)
Microsoft et al. (2025)	🔥 Conformer (Gulati et al., 2020)	🔥 MLP	❄ Phi-4-mini-instruct (Microsoft et al., 2025)
Kang and Roy (2024)	🔥 HuBERT-Large (Hsu et al., 2021)	🔥 Linear	❄ MiniChat-3B (Zhang et al., 2024a)
Züfle et al. (2025)	❄ HuBERT-Large (Hsu et al., 2021)	🔥 Q-Former (Li et al., 2023)	❄ LLaMA3.1-8B-Instruct (Dubey et al., 2024)
He et al. (2025)	❄ MERaLiON-Whisper (He et al., 2025)	🔥 MLP	≈ SEA-LION V3 (He et al., 2025)
Chu et al. (2024)	🔥 Whisper-large-v3 (Radford et al., 2023)	N/A	🔥 Qwen-7B (Bai et al., 2023)
Eom et al. (2025)	❄ Whisper-large-v2 (Radford et al., 2023)	🔥 Q-Mamba (Eom et al., 2025)	🔥 Mamba-2.8B-Zephyr (hf.co/xiuyul/mamba-2.8b-zephyr)

Table 2: Overview of Audio Encoder → Projector → LLM Architectures (🔥 trainable, ❄ frozen, ≈ LoRA)

training only the projector (Züfle et al., 2025). Others (Fathullah et al., 2024; Kang and Roy, 2024; Microsoft et al., 2025) train the projector alongside the audio encoder. Several approaches fine-tune the LLM using parameter-efficient techniques such as LoRA (Shang et al., 2024; He et al., 2025). Chu et al. (2024) omit a projection module, training all parameters of the audio encoder and LLM end to end. Eom et al. (2025) propose an alternative to transformer-based systems using Q-Mamba and a pretrained Mamba LLM.

Zero-Shot E2E SSum. LLM-based open-source models now, for the first time, make E2E SSum accessible with minimal setup. Models like Qwen2-Audio (Chu et al., 2024) have been used for zero-shot SSum without task-specific training (He et al., 2025; Züfle et al., 2025). Similarly, Phi-4 (Microsoft et al., 2025) supports audio inputs and shows potential for general-purpose SSum.

7 Challenges and Future Directions

Limited Reliability of Evaluation. A key bottleneck remains the lack of trustworthy evaluation practices for SSum. Most existing datasets rely on surrogate summaries, often lack audio data, and are limited by availability[1]. The majority also focus solely on English, restricting broader applicability. Simultaneously, ROUGE remains the dominant metric, despite its limited suitability for SSum. While LLM-based judges are gaining traction, common evaluation protocols are lacking. Human evaluations are often incomparable due to differences in setups, and few approaches account for speech-specific phenomena such as disfluencies, speaker variation, and background noise.

Personalization and Controllability. Summary needs vary by domain, audience, and intent. As Tuggener et al. (2021) outline, meeting summaries alone span formats from action items to narrative recaps, highlighting the mismatch between surrogate summaries and real user needs. Future work should enable controllable summarization along dimensions like length, focus, or style, and support personalization to user roles or preferences.

Underexplored Frontiers. Several promising directions in SSum remain underexplored. Online and real-time summarization has seen limited work, with only a few streaming-capable approaches (LeDuc et al., 2024; Schneider et al., 2025). Multi-document or multi-source SSum, where models process multiple speech inputs or supplemental materials, is also rare despite its relevance in collaborative settings (Kirstein et al., 2024a). Cross-lingual SSum is an emerging area, explored through cascaded setups with an intermediate MT module (Nelson et al., 2024) or integrated models that jointly translate and summarize (Kano et al., 2023a), yet remains largely untapped in E2E settings.

8 Conclusion

Despite the progress made in speech summarization, challenges remain, particularly in the development of multilingual datasets and evaluation benchmarks that can truly reflect real-world use cases. Future work will need to address these gaps while continuing to refine models for better faithfulness and efficiency. This survey takes a step toward addressing these challenges by providing a com-

[1] Most E2E approaches presented in Section 6.2 exclusively benchmarked on How2, a dataset that is now unavailable and based on surrogate summaries.

prehensive overview of existing datasets, summarization approaches, and evaluation methods and by promoting a more holistic view of SSum as a distinct and multifaceted research domain. As the field advances, SSum is poised to play a crucial role in enabling scalable, accessible insights from large, diverse collections of audiovisual content.

Limitations

Limitations of the Survey. While we have made efforts to provide a thorough review of the literature on speech summarization, some relevant works may have been overlooked due to variations in search criteria or keywords. Additionally, given the scope of this survey, we focus on the high-level aspects of the approaches and do not delve into an exhaustive, detailed experimental comparison. It is also worth noting that the field is evolving rapidly, particularly with the recent emergence of all-purpose language models. While we present these advancements, the widespread adoption of such models may significantly alter the landscape of speech summarization in the near future.

Ethical Considerations. Although several critical issues related to AI systems—such as bias, explainability, and fairness—have received increasing attention in recent work (Mei et al., 2023; Brandl et al., 2024; Gallegos et al., 2024), speech summarization remains a comparatively under-explored area (Liu et al., 2023c). Some researchers have begun to highlight the gap in assessing its ethical, legal, and societal implications (Shandilya et al., 2021; Keswani and Celis, 2021; Merine and Purkayastha, 2022; Steen and Markert, 2024).

SSum systems are active media agents that selectively extract and re-present information from audio or video sources, condensing spoken content into a more concise or structured written summary. In doing so, SSum serves as a powerful tool for controlling the selection and presentation of knowledge. These dynamics raise important questions about the broader consequences of algorithmic and engineering decisions, especially in how meaning is conveyed, distorted, or lost. The societal impact of automated summaries goes beyond sensitive domains like medicine, where inaccuracies could lead to misdiagnosis or harmful health outcomes (Otmakhova et al., 2022). Also in fields like scientific communication or news reporting, fluent but incorrect summaries can mislead and misinform (Zhao et al., 2020). These risks are further amplified in speech summarization, where disfluencies, ambiguity, and the lack of structural cues in spoken language make faithful abstraction especially challenging (Kirstein et al., 2025a). As language models become increasingly fluent and persuasive, the threat of confidently wrong summaries becomes all the more pressing.

Acknowledgments

This research is supported by the project "How is AI Changing Science? Research in the Era of Learning Algorithms" (HiAICS), funded by the Volkswagen Foundation, and partially by the European Union's Horizon research and innovation programme under grant agreement No. 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People).

References

Ramya Akula and Ivan Garibay. 2022. Sentence Pair Embeddings Based Evaluation Metric for Abstractive and Extractive Summarization. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6009–6017, Marseille, France. European Language Resources Association.

Zakaria Aldeneh, Matthew Perez, and Emily Mower Provost. 2021. Learning paralinguistic features from audiobooks through style voice conversion. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4736–4745, Online. Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Sumit Asthana, Sagih Hilleli, Pengcheng He, and Aaron Halfaker. 2024. Summaries, Highlights, and Action items: Design, implementation and evaluation of an LLM-powered meeting recap system. *arXiv preprint*. ArXiv:2307.15793 [cs].

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, ..., and Tianhang Zhu. 2023. Qwen Technical Report. *arXiv preprint*. ArXiv:2309.16609 [cs].

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. 2015. Generating Abstractive Summaries from Meeting Transcripts. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, DocEng '15, pages 51–60, New York, NY, USA. Association for Computing Machinery. Event-place: Lausanne, Switzerland.

Jeremy Barnes, Naiara Perez, Alba Bonet-Jover, and Begoña Altuna. 2025. Summarization metrics for spanish and basque: Do automatic scores and llm-judges correlate with humans? *Preprint*, arXiv:2503.17039.

Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1):164–171. Publisher: JSTOR.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. _eprint: 2004.05150.

Michael Bett, Ralph Gross, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel. 2000. Multimodal meeting tracker. In *RIAO*, pages 32–45. Paris, France.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, and Pengfei Liu. 2020. Metrics also disagree in the low scoring range: Revisiting summarization evaluation metrics. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5702–5711, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kuluhan Binici, Abhinav Ramesh Kashyap, Viktor Schlegel, Andy T. Liu, Vijay Prakash Dwivedi, Thanh-Tung Nguyen, Xiaoxue Gao, Nancy F. Chen, and Stefan Winkler. 2025. MEDSAGE: Enhancing Robustness of Medical Dialogue Summarization to ASR Errors with LLM-generated Synthetic Dialogues. *arXiv preprint*. ArXiv:2408.14418 [cs].

Stephanie Brandl, Emanuele Bugliarello, and Ilias Chalkidis. 2024. On the interplay between fairness and explainability. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 94–108, Mexico City, Mexico. Association for Computational Linguistics.

Brian Brazil, Ellen Troutman-Zaig, and Rebecca A Demarest. 2019. Prometheus. *Poems for the Millennium, Volume Three*.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A Pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer.

Laura Ceci. 2024. Hours of video uploaded to YouTube every minute 2007-2022. Publisher: Statista.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting Language Models for Social Conversation Synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 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)*, pages 8602–8615, Dublin, Ireland. Association for Computational Linguistics.

William Chen, Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe. 2024. Train Long and Test Long:Leveraging Full Document Contexts in Speech Processing. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13066–13070.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A Real-Life Scenario Dialogue Summarization Dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Sangwoo Cho, Franck Dernoncourt, Tim Ganter, Trung Bui, Nedim Lipka, Walter Chang, Hailin Jin, Jonathan Brandt, Hassan Foroosh, and Fei Liu. 2021. StreamHover: Livestream Transcript Summarization and Annotation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6457–6474, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Priyanjana Chowdhury, Nabanika Sarkar, Sanghamitra Nath, and Utpal Sharma. 2024. Analyzing the effects of transcription errors on summary generation of bengali spoken documents. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(9).

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-Audio Technical Report. *arXiv preprint*. ArXiv:2407.10759 [eess].

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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)*, pages 7282–7296, Online. Association for Computational Linguistics.

Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Arman Cohan, Guy Feigenblat, Tirthankar Ghosal, and Michal Shmueli-Scheuer. 2022. Overview of the first shared task on multi perspective scientific document summarization (MuP). In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 263–267, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Samuele Cornell, Matthew Wiesner, Shinji Watanabe, Desh Raj, Xuankai Chang, Paola García, Yoshiki Masuyama, Zhong-Qiu Wang, Stefano Squartini, and Sanjeev Khudanpur. 2023. The chime-7 DASR challenge: Distant meeting transcription with multiple devices in diverse scenarios. *CoRR*, abs/2306.13734.

Daniel Deutsch and Dan Roth. 2020. SacreROUGE: An open-source library for using and developing summarization evaluation metrics. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 120–125, Online. Association for Computational Linguistics.

Daniel Deutsch and Dan Roth. 2021. Understanding the extent to which content quality metrics measure the information quality of summaries. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 300–309, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, ..., and Zhiyu Ma. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. *CoRR*, abs/2404.04475. ArXiv: 2404.04475.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. 2022. Spurious Correlations in Reference-Free Evaluation of Text Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1443–1454, Dublin, Ireland. Association for Computational Linguistics.

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

SooHwan Eom, Jay Shim, Eunseop Yoon, Hee Suk Yoon, Hyeonmok Ko, Mark A. Hasegawa-Johnson, and Chang D. Yoo. 2025. SQuBa: Speech Mamba Language Model with Querying-Attention for Efficient Summarization.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *arXiv preprint*. ArXiv:2007.12626 [cs].

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2024. AudioChatLlama: Towards General-Purpose Speech Abilities for LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5522–5532, Mexico City, Mexico. Association for Computational Linguistics.

Guy Feigenblat, Chulaka Gunasekara, Benjamin Sznajder, Sachindra Joshi, David Konopnicki, and Ranit Aharonov. 2021. TWEETSUMM - A Dialog Summarization Dataset for Customer Service. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 245–260, Punta Cana, Dominican Republic. Association for Computational Linguistics.

M. Franzini, K.-F. Lee, and A. Waibel. 1990. Connectionist Viterbi training: a new hybrid method for continuous speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 425–428 vol.1.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and Shashi Tn. 2024. Tiny Titans: Can Smaller Large Language Models Punch Above Their Weight in the Real World for Meeting Summarization? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 387–394, Mexico City, Mexico. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artif. Intell. Rev.*, 47(1):1–66.

Mingqi Gao and Xiaojun Wan. 2022. DialSummEval: Revisiting summarization evaluation for dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5693–5709, Seattle, United States. Association for Computational Linguistics.

Ekaterina Garmash, Edgar Tanaka, Ann Clifton, Joana Correia, Sharmistha Jat, Winstead Zhu, Rosie Jones, and Jussi Karlgren. 2023. Cem Mil Podcasts: A Spoken Portuguese Document Corpus for Multi-modal, Multi-lingual and Multi-dialect Information Access Research. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 48–59, Cham. Springer Nature Switzerland.

F Ruth Gee. 1998. The TIPSTER text program overview. In *TIPSTER Text Program Phase III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 3–5, Baltimore, Maryland.

Azin Ghazimatin, Ekaterina Garmash, Gustavo Penha, Kristen Sheets, Martin Achenbach, Oguz Semerci, Remi Galvez, Marcus Tannenberg, Sahitya Mantravadi, Divya Narayanan, Ofeliya Kalaydzhyan, Douglas Cole, Ben Carterette, Ann Clifton, Paul N. Bennett, Claudia Hauff, and Mounia Lalmas. 2024. PODTILE: Facilitating Podcast Episode Browsing with Auto-generated Chapters. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 4487–4495. ACM.

Tirthankar Ghosal, Ondřej Bojar, Marie Hledíková, Tom Kocmi, and Anna Nedoluzhko. 2023. Overview of the Second Shared Task on Automatic Minuting (AutoMin) at INLG 2023. In *Proceedings of the 16th International Natural Language Generation Conference: Generation Challenges*, pages 138–167, Prague, Czechia. Association for Computational Linguistics.

Tirthankar Ghosal, Muskaan Singh, Anja Nedoluzhko, and Ondřej Bojar. 2022. Report on the SIGDial 2021 special session on summarization of dialogues and multi-party meetings (SummDial). *SIGIR Forum*, 55(2). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

Ziwei Gong, Lin Ai, Harshsaiprasad Deshpande, Alexander Johnson, Emmy Phung, Zehui Wu, Ahmad Emami, and Julia Hirschberg. 2024. CREAM: Comparison-Based Reference-Free ELO-Ranked Automatic Evaluation for Meeting Summarization. *arXiv preprint*. ArXiv:2409.10883 [cs].

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3. *Preprint*, arXiv:2209.12356.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, ..., and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ralph Gross, Michael Bett, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel. 2000. Towards a multimodal meeting record. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, volume 3, pages 1593–1596. IEEE.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*, pages 5036–5040. ISSN: 2958-1796.

Hardy Hardy, Shashi Narayan, and Andreas Vlachos. 2019. HighRES: Highlight-based reference-less evaluation of summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3381–3392, Florence, Italy. Association for Computational Linguistics.

Donna Harman and Paul Over. 2004. The effects of human variation in DUC summarization evaluation. In *Text Summarization Branches Out*, pages 10–17, Barcelona, Spain. Association for Computational Linguistics.

Yingxu He, Zhuohan Liu, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy F. Chen, and Ai Ti Aw. 2025. MERaLiON-AudioLLM: Bridging Audio and Language with Large Language Models. *arXiv preprint*. ArXiv:2412.09818 [cs].

Chiori Hori, Sadaoki Furui, Rob Malkin, Hua Yu, and Alex Waibel. 2002. Automatic speech summarization applied to english broadcast news speech. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–9. IEEE.

Takaaki Hori, Chiori Hori, and Yasuhiro Minami. 2003. Speech summarization using weighted finite-state transducers. In *INTERSPEECH*, pages 2817–2820. Citeseer.

Eduard Hovy, Chin-Yew Lin, Liang Zhou, and Junichi Fukumoto. 2006. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460. Publisher: IEEE Press.

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

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

Kung-Hsiang Huang, Siffi Singh, Xiaofei Ma, Wei Xiao, Feng Nan, Nicholas Dingwall, William Yang Wang, and Kathleen McKeown. 2023. SWING: Balancing Coverage and Faithfulness for Dialogue Summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 512–525, Dubrovnik, Croatia. Association for Computational Linguistics.

Akira Inoue, Takayoshi Mikami, and Yoichi Yamashita. 2004. Improvement of speech summarization using prosodic information. In *Speech Prosody 2004*, pages 599–602. ISCA.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020a. Towards a reliable and robust methodology for crowd-based subjective quality assessment of query-based extractive text summarization. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 245–253, Marseille, France. European Language Resources Association.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2020b. Best Practices for Crowd-based Evaluation of German Summarization: Comparing Crowd, Expert and Automatic Evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 164–175, Online. Association for Computational Linguistics.

Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A Survey on Multi-modal Summarization. *ACM Comput. Surv.*, 55(13s). Place: New York, NY, USA Publisher: Association for Computing Machinery.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI Meeting Corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I. ISSN: 1520-6149.

Frederick Jelinek. 1976. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556.

Rafael Zequeira Jiménez, Babak Naderi, and Sebastian Möller. 2020. Effect of environmental noise in speech quality assessment studies using crowdsourcing. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6.

Rosie Jones, Ben Carterette, Ann Clifton, Jussi Karlgren, Aasish Pappu, Sravana Reddy, Yongze Yu, Maria Eskevich, and Gareth J. F. Jones. 2020. TREC 2020 Podcasts Track Overview. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Jee-Weon Jung, Roshan S. Sharma, William Chen, Bhiksha Raj, and Shinji Watanabe. 2024. AugSumm: Towards Generalizable Speech Summarization Using

Synthetic Labels from Large Language Models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 12071–12075. IEEE.

Tollef Emil Jørgensen and Ole Jakob Mengshoel. 2025. Cross-Lingual Sentence Compression for Length-Constrained Subtitles in Low-Resource Settings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6447–6458, Abu Dhabi, UAE. Association for Computational Linguistics.

Wonjune Kang and Deb Roy. 2024. Prompting Large Language Models with Audio for General-Purpose Speech Summarization. In *Interspeech 2024*, pages 1955–1959. ISSN: 2958-1796.

Takatomo Kano, Atsunori Ogawa, Marc Delcroix, Kohei Matsuura, Takanori Ashihara, William Chen, and Shinji Watanabe. 2023a. Summarize while translating: Universal model with parallel decoding for summarization and translation. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Takatomo Kano, Atsunori Ogawa, Marc Delcroix, Roshan Sharma, Kohei Matsuura, and Shinji Watanabe. 2023b. Speech Summarization of Long Spoken Document: Improving Memory Efficiency of Speech/Text Encoders. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Takatomo Kano, Atsunori Ogawa, Marc Delcroix, and Shinji Watanabe. 2021. Attention-based Multi-hypothesis Fusion for Speech Summarization. *arXiv preprint*. ArXiv:2111.08201 [eess].

Vijay Keswani and L. Elisa Celis. 2021. Dialect diversity in text summarization on twitter. In *Proceedings of the Web Conference 2021*, WWW '21, page 3802–3814, New York, NY, USA. Association for Computing Machinery.

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

Aman Khullar and Udit Arora. 2020. MAST: Multimodal Abstractive Summarization with Trimodal Hierarchical Attention. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 60–69, Online. Association for Computational Linguistics.

Frederic Kirstein, Terry Ruas, Robert Kratel, and Bela Gipp. 2024a. Tell me what I need to know: Exploring LLM-based (Personalized) Abstractive Multi-Source Meeting Summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 920–939, Miami, Florida, US. Association for Computational Linguistics.

Frederic Kirstein, Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2025a. CADS: A Systematic Literature Review on the Challenges of Abstractive Dialogue Summarization. *J. Artif. Int. Res.*, 82. Place: El Segundo, CA, USA Publisher: AI Access Foundation.

Frederic Kirstein, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2024b. What's under the hood: Investigating automatic metrics on meeting summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6709–6723, Miami, Florida, USA. Association for Computational Linguistics.

Frederic Kirstein, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2024c. What's under the hood: Investigating Automatic Metrics on Meeting Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6709–6723, Miami, Florida, USA. Association for Computational Linguistics.

Frederic Thomas Kirstein, Terry Lima Ruas, and Bela Gipp. 2025b. Is my Meeting Summary Good? Estimating Quality with a Multi-LLM Evaluator. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 561–574, Abu Dhabi, UAE. Association for Computational Linguistics.

Frederic Thomas Kirstein, Terry Lima Ruas, and Bela Gipp. 2025c. What's Wrong? Refining Meeting Summaries with LLM Feedback. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2100–2120, Abu Dhabi, UAE. Association for Computational Linguistics.

Konstantinos Koumpis and Steve Renals. 2005. Automatic summarization of voicemail messages using lexical and prosodic features. *ACM Transactions on Speech and Language Processing*, 2(1):1.

Margaret Kroll and Kelsey Kraus. 2024. Optimizing the role of human evaluation in llm-based spoken document summarization systems. In *Interspeech 2024*, interspeech2024, page 1935–1939. ISCA.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Lakshmi Prasanna Kumar and Arman Kabiri. 2022. Meeting summarization: A survey of the state of the art. *Preprint*, arXiv:2212.08206.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 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*, pages 343–352, Singapore. Association for Computational Linguistics.

Khai Le-Duc, Khai-Nguyen Nguyen, Long Vo-Dang, and Truong-Son Hy. 2024. Real-time Speech Summarization for Medical Conversations. In *Interspeech 2024*, pages 1960–1964. ISCA.

Daniel Li, Thomas Chen, Albert Tung, and Lydia B Chilton. 2021. Hierarchical Summarization for Long-form Spoken Dialog. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21, pages 582–597, New York, NY, USA. Association for Computing Machinery. Event-place: Virtual Event, USA.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org. Place: Honolulu, Hawaii, USA.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep Meeting Summaries on Topic: Abstractive Multi-Modal Meeting Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

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

Haitao Lin, Liqun Ma, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 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*, pages 4436–4451, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhenqing Ling, Yuexiang Xie, Chenhe Dong, and Ying Shen. 2025. Enhancing factual consistency in text summarization via counterfactual debiasing. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7912–7924, Abu Dhabi, UAE. Association for Computational Linguistics.

Benjamin Litterer, David Jurgens, and Dallas Card. 2024. Mapping the Podcast Ecosystem with the Structured Podcast Research Corpus. *arXiv preprint*. ArXiv:2411.07892 [cs].

Danni Liu, Jan Niehues, and Gerasimos Spanakis. 2020. Adapting End-to-End Speech Recognition for Readable Subtitles. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 247–256, Online. Association for Computational Linguistics.

Dongqi Liu, Chenxi Whitehouse, Xi Yu, Louis Mahon, Rohit Saxena, Zheng Zhao, Yifu Qiu, Mirella Lapata, and Vera Demberg. 2025. What Is That Talk About? A Video-to-Text Summarization Dataset for Scientific Presentations. *arXiv preprint*. ArXiv:2502.08279 [cs].

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11807–11823, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yixin Liu, Alex Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023b. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4140–4170, Toronto, Canada. Association for Computational Linguistics.

Yu Lu Liu, Meng Cao, Su Lin Blodgett, Jackie Chi Kit Cheung, Alexandra Olteanu, and Adam Trischler. 2023c. Responsible AI considerations in text summarization research: A review of current practices. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6246–6261, Singapore. Association for Computational Linguistics.

Zhengyuan Liu, Angela Ng, Sheldon Lee, Ai Ti Aw, and Nancy F. Chen. 2019. Topic-Aware Pointer-Generator Networks for Summarizing Spoken Conversations. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 814–821.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Tengchao Lv, Lei Cui, Momcilo Vasilijevic, and Furu Wei. 2021. VT-SSum: A Benchmark Dataset for Video Transcript Segmentation and Summarization. *arXiv preprint*. ArXiv:2106.05606 [cs].

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

Kikuo Maekawa. 2003. Corpus of spontaneous Japanese: its design and evaluation. page paper MMO2.

Potsawee Manakul and Mark J. F. Gales. 2020. CUED_speech at TREC 2020 Podcast Summarisation Track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Potsawee Manakul and Mark J. F. Gales. 2022. Podcast Summary Assessment: A Resource for Evaluating Summary Assessment Methods. *arXiv preprint*. ArXiv:2208.13265 [cs].

Sameer Maskey and Julia Hirschberg. 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Interspeech*, pages 621–624.

Antonio Mastropaolo, Matteo Ciniselli, Massimiliano Di Penta, and Gabriele Bavota. 2023. Evaluating code summarization techniques: A new metric and an empirical characterization. *Preprint*, arXiv:2312.15475.

Kohei Matsuura, Takanori Ashihara, Takafumi Moriya, Masato Mimura, Takatomo Kano, Atsunori Ogawa, and Marc Delcroix. 2024. Sentence-wise Speech Summarization: Task, Datasets, and End-to-End Modeling with LM Knowledge Distillation. In *Interspeech 2024*, pages 1945–1949.

Kohei Matsuura, Takanori Ashihara, Takafumi Moriya, Tomohiro Tanaka, Takatomo Kano, Atsunori Ogawa, and Marc Delcroix. 2023a. Transfer Learning from Pre-trained Language Models Improves End-to-End Speech Summarization. In *Interspeech 2023*, pages 2943–2947. ISSN: 2958-1796.

Kohei Matsuura, Takanori Ashihara, Takafumi Moriya, Tomohiro Tanaka, Atsunori Ogawa, Marc Delcroix, and Ryo Masumura. 2023b. Leveraging Large Text Corpora For End-To-End Speech Summarization. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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

Kathleen McKeown, Julia Hirschberg, Michel Galley, and Sameer Maskey. 2005. From text to speech summarization. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v–997. IEEE.

Alex Mei, Sharon Levy, and William Yang Wang. 2023. Foveate, attribute, and rationalize: Towards physically safe and trustworthy AI. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11021–11036, Toronto, Canada. Association for Computational Linguistics.

Xinhao Mei, Xubo Liu, Mark D. Plumbley, and Wenwu Wang. 2022. Automated audio captioning: an overview of recent progress and new challenges. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1):26.

Regina Merine and Saptarshi Purkayastha. 2022. Risks and benefits of ai-generated text summarization for expert level content in graduate health informatics. In *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*, pages 567–574.

Microsoft, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yiling Chen, Qi Dai, Xiyang Dai, ..., and Xiren Zhou. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. *arXiv preprint*. ArXiv:2503.01743 [cs].

Derek Miller. 2019. Leveraging BERT for Extractive Text Summarization on Lectures. *arXiv preprint*. ArXiv:1906.04165 [cs].

Koichi Miyazaki, Yoshiki Masuyama, and Masato Murata. 2024. Exploring the Capability of Mamba in Speech Applications. In *Interspeech 2024*, pages 237–241. ISSN: 2958-1796.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. ECTSum: A New Benchmark Dataset For Bullet Point Summarization of Long Earnings Call Transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

James Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer Seale, Jordan Swartz, Greg McKelvey, Hui Dai, Yi Yang, and David Sontag. 2021. CLIP: A Dataset for Extracting Action Items for Physicians from Hospital Discharge Notes. 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)*, pages 1365–1378, Online. Association for Computational Linguistics.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. Generating and Validating Abstracts of Meeting Conversations: a User Study. In *Proceedings of the 6th International Natural Language Generation Conference*. Association for Computational Linguistics.

Gabriel Murray, Steve Renals, and Jean Carletta. 2005. Extractive summarization of meeting recordings. In *Interspeech 2005*, pages 593–596. ISSN: 2958-1796.

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

Max Nelson, Shannon Wotherspoon, Francis Keith, William Hartmann, and Matthew Snover. 2024. Cross-Lingual Conversational Speech Summarization with Large Language Models. *arXiv preprint*. ArXiv:2408.06484 [cs].

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Fumio Nihei, Yukiko I. Nakano, and Yutaka Takase. 2018. Fusing Verbal and Nonverbal Information for Extractive Meeting Summarization. In *Proceedings of the Group Interaction Frontiers in Technology*, GIFT'18, New York, NY, USA. Association for Computing Machinery. Event-place: Boulder, CO, USA.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, ..., and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yulia Otmakhova, Karin Verspoor, Timothy Baldwin, and Jey Han Lau. 2022. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5111, Dublin, Ireland. Association for Computational Linguistics.

Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal Abstractive Summarization for How2 Videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, Florence, Italy. Association for Computational Linguistics.

Shruti Palaskar, Ruslan Salakhutdinov, Alan W. Black, and Florian Metze. 2021. Multimodal Speech Summarization Through Semantic Concept Learning. In *Interspeech 2021*, pages 791–795. ISSN: 2958-1796.

Sara Papi, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2023a. Direct speech translation for automatic subtitling. *Transactions of the Association for Computational Linguistics*, 11:1355–1376.

Sara Papi, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2023b. Direct Speech Translation for Automatic Subtitling. *Transactions of the Association for Computational Linguistics*, 11:1355–1376.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100, Florence, Italy. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and Summarizing Action Items in Multi-Party Dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 18–25, Antwerp, Belgium. Association for Computational Linguistics.

Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, Bo Li, and Lijuan Wang. 2024. MMSum: A Dataset for Multimodal Summarization and Thumbnail Generation of Videos. pages 21909–21921.

Lawrence R Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286. Publisher: Ieee.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023.

Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

G. J. Rath, A. Resnick, and T. R. Savage. 1961. The formation of abstracts by the selection of sentences. Part I. Sentence selection by men and machines. *American Documentation*, 12(2):139–141.

Ines Rehbein, Josef Ruppenhofer, and Thomas Schmidt. 2020. Improving sentence boundary detection for spoken language transcripts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7102–7111, Marseille, France. European Language Resources Association.

S. Renals, N. Morgan, H. Bourlard, M. Cohen, and H. Franco. 1994. Connectionist probability estimators in HMM speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(1):161–174.

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2023a. Abstractive Meeting Summarization: A Survey. *Transactions of the Association for Computational Linguistics*, 11:861–884.

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2023b. Abstractive meeting summarization: A survey. *Transactions of the Association for Computational Linguistics*, 11:861–884.

Fabian Retkowski. 2023. The current state of summarization. In Andreas Sudmann, Anna Echterhölter, Markus Ramsauer, Fabian Retkowski, Jens Schröter, and Alexander Waibel, editors, *Beyond Quantity. Research with Subsymbolic AI*, 1 edition, pages 291–312. transcript Verlag, Bielefeld, Germany.

Fabian Retkowski and Alexander Waibel. 2024a. From Text Segmentation to Smart Chaptering: A Novel Benchmark for Structuring Video Transcriptions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 406–419, St. Julian's, Malta. Association for Computational Linguistics.

Fabian Retkowski and Alexander Waibel. 2024b. Zero-Shot Strategies for Length-Controllable Summarization. *arXiv preprint*. ArXiv:2501.00233 [cs].

Dana Rezazadegan, Shlomo Berkovsky, Juan C. Quiroz, A. Baki Kocaballi, Ying Wang, Liliana Laranjo, and Enrico Coiera. 2020. Automatic Speech Summarisation: A Scoping Review. *arXiv preprint*. ArXiv:2008.11897 [cs].

Bareera Sadia, Farah Adeeba, Sana Shams, and Kashif Javed. 2024. Meeting the challenge: A benchmark corpus for automated Urdu meeting summarization. *Inf. Process. Manage.*, 61(4). Place: USA Publisher: Pergamon Press, Inc.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset for Multimodal Language Understanding. *arXiv preprint*. ArXiv:1811.00347 [cs].

Felix Schneider, Marco Turchi, and Alex Waibel. 2025. Policies and evaluation for online meeting summarization. *Preprint*, arXiv:2502.03111.

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

Anurag Shandilya, Abhisek Dash, Abhijnan Chakraborty, Kripabandhu Ghosh, and Saptarshi Ghosh. 2021. Fairness for whom? understanding the reader's perception of fairness in text summarization. *CoRR*, abs/2101.12406.

Hengchao Shang, Zongyao Li, Jiaxin Guo, Shaojun Li, Zhiqiang Rao, Yuanchang Luo, Daimeng Wei, and Hao Yang. 2024. An End-to-End Speech Summarization Using Large Language Model. In *Interspeech 2024*, pages 1950–1954. ISSN: 2958-1796.

Roshan Sharma. 2024. *End-to-End Modeling for Abstractive Speech Summarization*. Ph.D. thesis.

Roshan Sharma, Siddhant Arora, Kenneth Zheng, Shinji Watanabe, Rita Singh, and Bhiksha Raj. 2023. BASS: Block-wise Adaptation for Speech Summarization. In *Interspeech 2023*, pages 1454–1458. ISSN: 2958-1796.

Roshan Sharma, Shruti Palaskar, Alan W Black, and Florian Metze. 2022. End-to-End Speech Summarization Using Restricted Self-Attention. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8072–8076.

Roshan Sharma, Ruchira Sharma, Hira Dhamyal, Rita Singh, and Bhiksha Raj. 2024a. R-BASS : Relevance-aided Block-wise Adaptation for Speech Summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 848–857, Mexico City, Mexico. Association for Computational Linguistics.

Roshan Sharma, Suwon Shon, Mark Lindsey, Hira Dhamyal, and Bhiksha Raj. 2024b. Speech vs. Transcript: Does It Matter for Human Annotators in Speech Summarization? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14779–14797, Bangkok, Thailand. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-yi Lee, Karen Livescu, and Shinji Watanabe. 2023. SLUE Phase-2: A Benchmark Suite of Diverse Spoken Language Understanding Tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8906–8937, Toronto, Canada. Association for Computational Linguistics.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Fei Liu. 2022a. Towards Abstractive Grounded Summarization of Podcast Transcripts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4407–4418, Dublin, Ireland. Association for Computational Linguistics.

Kaiqiang Song, Chen Li, Xiaoyang Wang, Dong Yu, and Fei Liu. 2022b. Towards abstractive grounded summarization of podcast transcripts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4407–4418, Dublin, Ireland. Association for Computational Linguistics.

Julius Steen and Katja Markert. 2024. Bias in news summarization: Measures, pitfalls and corpora. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5962–5983, Bangkok, Thailand. Association for Computational Linguistics.

Bernhard Suhm. 1994. Towards better language models for spontaneous speech. In *Proc. ICSLP'94*, volume 2, pages 831–834.

Sathya Krishnan Suresh, Wu Mengjun, Tushar Pranav, and Eng Siong Chng. 2025. DiaSynth: Synthetic Dialogue Generation Framework for Low Resource Dialogue Applications. *arXiv preprint*. ArXiv:2409.19020 [cs].

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. TofuEval: Evaluating Hallucinations of LLMs on Topic-Focused Dialogue Summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *Preprint*, arXiv:2406.12624.

David Thulke, Yingbo Gao, Rricha Jalota, Christian Dugast, and Hermann Ney. 2024. Prompting and Fine-Tuning of Small LLMs for Length-Controllable Telephone Call Summarization. *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 305–312.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, ..., and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*. ArXiv:2307.09288 [cs].

Don Tuggener, Margot Mieskes, Jan Deriu, and Mark Cieliebak. 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*, pages 107–118, Online and in Dominican Republic. Association for Computational Linguistics.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 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*, pages 11–20, Online. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. pages 4566–4575.

Gianluca Vico and Jan Niehues. 2022. TED Talk Teaser Generation with Pre-Trained Models. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8067–8071. ISSN: 2379-190X.

A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang. 1989. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

WhatsApp. 2022. Making Voice Messages Better. Publisher: WhatsApp.

Steve Whittaker, Julia Hirschberg, John Choi, Don Hindle, Fernando Pereira, and Amit Singhal. 1999. SCAN: Designing and evaluating user interfaces to support retrieval from speech archives. pages 26–33.

Han Wu, Mingjie Zhan, Haochen Tan, Zhaohui Hou, Ding Liang, and Linqi Song. 2023. VCSUM: A Versatile Chinese Meeting Summarization Dataset. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6065–6079, Toronto, Canada. Association for Computational Linguistics.

Shasha Xie and Yang Liu. 2010. Using Confusion Networks for Speech Summarization. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 46–54, Los Angeles, California. Association for Computational Linguistics.

Tao Xie, Yuanyuan Kuang, Ying Tang, Jian Liao, and Yunong Yang. 2025. Using LLM-supported lecture summarization system to improve knowledge recall and student satisfaction. *Expert Systems with Applications*, 269:126371.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 100–110, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Klaus Zechner and Alex Waibel. 2000a. DIASUMM: Flexible Summarization of Spontaneous Dialogues in Unrestricted Domains. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Klaus Zechner and Alex Waibel. 2000b. Minimizing Word Error Rate in Textual Summaries of Spoken Language. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Torsten Zeppenfeld, Michael Finke, Klaus Ries, Martin Westphal, and Alex Waibel. 1997. Recognition of conversational telephone speech using the janus speech engine. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1815–1818. IEEE.

Chen Zhang, Dawei Song, Zheyu Ye, and Yan Gao. 2024a. Towards the Law of Capacity Gap in Distilling Language Models. *arXiv preprint*. ArXiv:2311.07052 [cs].

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 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.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024b. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Yusen Zhang, Ansong Ni, Tao Yu, Rui Zhang, Chenguang Zhu, Budhaditya Deb, Asli Celikyilmaz, Ahmed Hassan Awadallah, and Dragomir Radev. 2021. An Exploratory Study on Long Dialogue Summarization: What Works and What‘s Next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4426–4433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019a. 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)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019b. 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)*, pages 563–578, Hong Kong, China. Association for Computational Linguistics.

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Com-*

putational Linguistics: EMNLP 2020, pages 2237–2249, Online. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022a. DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11765–11773.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022b. Towards a Unified Multi-Dimensional Evaluator for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 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*, pages 5927–5934, Online. Association for Computational Linguistics.

Chenguang Zhu, Ruochen Xu, Michael Zeng, and Xuedong Huang. 2020. A Hierarchical Network for Abstractive Meeting Summarization with Cross-Domain Pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 194–203, Online. Association for Computational Linguistics.

Rongxin Zhu, Jey Han Lau, and Jianzhong Qi. 2025. Factual Dialogue Summarization via Learning from Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4474–4492, Abu Dhabi, UAE. Association for Computational Linguistics.

Maike Züfle, Sara Papi, Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, and Jan Niehues. 2025. NUTSHELL: A Dataset for Abstract Generation from Scientific Talks. *arXiv preprint*. ArXiv:2502.16942 [cs].

A Datasets

A.1 Non-English Datasets

Dataset	Reference	Domain	Lang.	Size	Summary Type	Transcript	Audio	Video	License
CSJ	Maekawa (2003)	Academic speech (various types)	JA	3.3k recordings (661 hours)	Abstractive & Extractive	Manual	✓	✗	Paid
VCSum	Wu et al. (2023)	Roundtable meetings (from Chinese video-sharing websites)	ZH	239 meetings (230 hours)	Abstractive (overall, segment-level, and chapter titles) & Extractive	ASR	✗	✗	MIT
CLE Meeting Corpus	Sadia et al. (2024)	Administrative & technical meetings (virtual, mostly scenario-driven)	UR	240 meetings	Abstractive (overall summaries, multiple)	Manual	✗	✗	?[h]
MNSC	He et al. (2025)	Conversations of various nature (IMDA NSC Corpus)	SGE	~100 hours	Abstractive	Manual	✓	✗	Singapore Open Data License
VietMed-Sum	Le-Duc et al. (2024)	Medical conversations	VI	16 hours	Abstractive (local & global)	Manual	✓	✗	?[a]

[a] ? No explicit license has been provided.

Table 3: Non-English Datasets Related to the Speech Summarization Task

A.2 Chat-Based Datasets

Dataset	Reference	Domain	Lang.	Size	Summary Type	Transcript	Audio	Video	License
TweetSumm	Feigenblat et al. (2021)	Customer service chats (Twitter)	EN	1.1k dialogues	Abstractive & Extractive (multiple)	N/A (Chat)	✗	✗	CDLA-Sharing-1.0
CSDS	Lin et al. (2021)	Customer service chats (JD.com)	ZH	2.5k dialogues	Extractive & Abstractive (role-oriented, topic-structured, multiple)	N/A (Chat)	✗	✗	?[a]
SAMSum	Gliwa et al. (2019)	Chat conversations (scenario-driven)	EN	16k dialogues	Abstractive	N/A (Chat)	✗	✗	CC-BY-NC-ND-4.0

[a] ? No explicit license has been provided.

Table 4: Chat-Based Summarization Datasets Structurally Similar to Speech

A.3 Dataset Derivates and Augmentations

Dataset	Reference	Base Dataset	Lang.	Extension Type	License
AugSumm	Jung et al. (2024)	How2	EN	Synthetic summaries generated by GPT-3.5 Turbo (direct + paraphrased) to enrich summary diversity	?[a]
QMSum-I	Fu et al. (2024)	QMSum	EN	Instruction-based summaries (long, medium, short) generated by GPT-4	?[a]
MS-AMI	Kirstein et al. (2024a)	AMI	EN	Enriches the source data with processed, supplementary materials (whiteboard drawings, slides, notes) using GPT-4o and Aspose for text extraction	Apache-2.0
YTSeg-LC	Retkowski and Waibel (2024b)	YTSeg	EN	Length-controlled summaries generated by LLaMA 3 and other LLMs	CC-BY-NC-SA 4.0

[a] ? No explicit license has been provided.

Table 5: Derivatives of and Augmentations to Existing Speech Summarization Sources

B Evaluation of Speech Summaries

B.1 Usage of Speech Summarization Metrics over Time

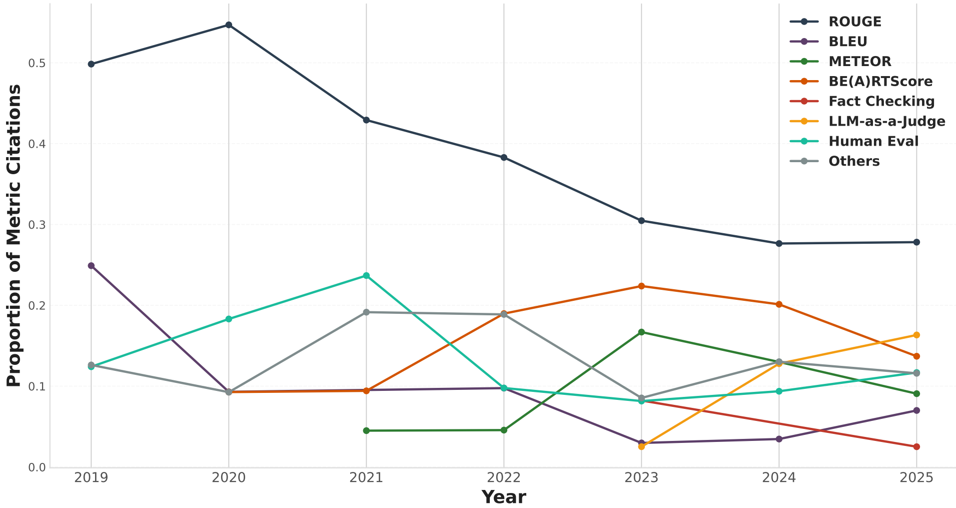

Figure 3: Proportion of citations for different evaluation metrics over time (based on the SSum papers included in this survey; after 2018), normalized by the total number of citations per year. *Others* includes all metrics with three or fewer citations.

Trends in Usage of Metrics. Figure 3 shows the normalized proportion of citations for various evaluation metrics from 2019 to 2025. We observe an increase of using other metrics than ROUGE (Lin, 2004) from 2019 to 2023, followed by stagnation. BE(A)RTScore (BERTScore, [Zhang et al., 2020b] and BARTScore [Yuan et al., 2021]) grows steadily from 2020 to 2023. Human evaluation has remained relatively stable throughout the years. By 2025, LLM-as-a-Judge becomes the second most used metric, emerging in 2024 and rapidly gaining popularity. A detailed overview of the different LLM-as-a-Judge methods can be found in Table 6.

Fact Checking. The *Fact Checking* category includes the following metrics: FactCC (Huang et al., 2023), QUALS (Huang et al., 2023), QAGS (Suresh et al., 2025), QAFactEval (Huang et al., 2023), FactVC (Liu and Wan, 2023), and SummaC-Conv (Laban et al., 2022).

Others. The *Others* category includes metrics less frequently used for speech summarization, such as F-score (Lv et al., 2021; Palaskar et al., 2019), Perplexity (Kirstein et al., 2024c; Retkowski and Waibel, 2024b), ChrF (Popović, 2015; Jørgensen and Mengshoel, 2025), Sentence Cosine Similarity (Li et al., 2021), BoC (Bag of Characters) (Chen et al., 2022), BLANC (Kirstein et al., 2024c; Vasilyev et al., 2020), LENS (Maddela et al., 2023; Kirstein et al., 2024c), MoverScore (Zhao et al., 2019b; Hu et al., 2023b), CIDEr (Vedantam et al., 2015; Qiu et al., 2024), and SPICE (Anderson et al., 2016; Qiu et al., 2024).

B.2 LLM-as-a-Judge for Speech Summarization

Method	Judge Model	Criteria (Framework)	Data	Reference	Total
Absolute Score/Scale	Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	Relevance, Coherence, Conciseness, Factual Accuracy	Output Summary, Reference	Züfle et al. (2025)	7
	Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	General Alignment with Reference	Output Summary, Reference	Züfle et al. (2025)	
	Meta-Llama-3-70B (Grattafiori et al., 2024)	Content, Accuracy, and Relevance	Output Summary, Reference	He et al. (2025)	
	GPT-4 (OpenAI et al., 2024)	Overall Quality, Instruction Adherence	Transcript, Output Summary	Microsoft et al. (2025)	
	Prometheus-8x7B (Brazil et al., 2019)	Honesty, Factual Validity, Completeness (Prometheus-Eval, Brazil et al., 2019)	Transcript, Output Summary	Thulke et al. (2024)	
	GPT-4o (OpenAI et al., 2024)	Redundancy, Incoherence, Language, Omission, Coreference, Hallucination, Structure, Irrelevance (MESA, Kirstein et al., 2025b)	Transcript, Output Summary	Kirstein et al. (2025b)	
	GPT-4-32k (OpenAI et al., 2024)	Adequacy, Relevance, Topicality, Fluency, Grammaticality	Transcript, Output Summary	Ghosal et al. (2023)	
Ranking	Llama-3.1-8B-Instruct (Grattafiori et al., 2024)	Relevance, Coherence, Conciseness, Factual Accuracy	Output Summaries, Reference	Züfle et al. (2025)	2
	GPT-based models (OpenAI et al., 2024)	Completeness, Conciseness, Factuality (CREAM, Gong et al., 2024)	Output Summaries	Gong et al. (2024)	
Pairwise Comparison	GPT4-Turbo (OpenAI et al., 2024)	Not Specified	Output Summary, Reference	Matsuura et al. (2024)	2
	GPT-4o (OpenAI et al., 2024)	General Performance (Alpaca Eval, Dubois et al., 2023)	Transcript, Output Summary, Baseline Summary	Retkowski and Waibel (2024b)	
Accuracy	GPT-4 (OpenAI et al., 2024)	Hallucination	Transcript, Output Summary	Microsoft et al. (2025)	3
	GPT-4o (OpenAI et al., 2024)	Faithfulness, Completeness, Conciseness (FineSureE, Song et al., 2024)	Transcript, Output Summary	Thulke et al. (2024)	
	GPT-4 (OpenAI et al., 2024) among other, weaker judges	Factual Correctness	Transcript, Output Summary/Sentence	Tang et al. (2024)	

Table 6: Different ways of LLM-as-a-Judge for SSum, based on the SSum papers included in this survey.

B.3 Human Evaluation for Speech Summarization

Method	Annotators	Criteria	Data	Reference	Total
Likert Scale	Crowdsourced	Readability, Relevance	Transcript, Output Summary	Zhu et al. (2020)	12
	Crowdsourced	Informativeness, Relevance, Coherence	Video, Output Summary, Reference	Palaskar et al. (2019)	
	Crowdsourced	Informativeness, Redundancy	Transcript, Output Summary	Song et al. (2022b)	
	Crowdsourced	Informativeness, Factuality, Fluency, Coherence, Redundancy	Video, Transcript, Output Summary	Hu et al. (2023b)	
	Graduate Students	Frequency of Transcript Challenges	Transcripts, Output Summary Reference	Kirstein et al. (2024b)	
	Domain Experts	Adequacy, Fluency, Relevance	Transcript, Summary	Schneider et al. (2025)	
	Not Specified	Fluency, Coherence, Factual Consistency	Not Specified	Fu et al. (2024)	
	Domain Experts	Fluency, Consistency, Relevance, Coherence	Transcript, Output Summary	Le-Duc et al. (2024)	
	Graduate Students	Error Types Detection	Transcript, Output Summary	Kirstein et al. (2025b)	
	Not Specified	Fluency, Consistency, Relevance, Coherence	Source (Dialog), Output Summary	Chen et al. (2021)	
	Experienced Annotators	Adequacy (Informativeness), Fluency, Grammatical Correctness, Relevance	Transcripts, Output Summary	Ghosal et al. (2023)	
	Well-Educated Volunteers	Informativeness, Redundancy, Fluency, Matching Rate	Transcripts, Output Summary	Lin et al. (2021)	
Best-Worst Scaling	Domain Experts	Relevance, Coherence, Conciseness, Factual Accuracy	Output Summaries, Reference	Züfle et al. (2025)	2
	Graduate Students	Fluency, Informativeness, Faithfulness	Source (Dialog), Output Summaries	Zhong et al. (2022a)	
Pairwise Comparison	Crowdsourced	Coherence, Informativeness, Overall quality	Transcript, Output Summaries	Cho et al. (2021)	5
	Crowdsourced	Factual Consistency, Informativeness	Source (Dialog), Output Summaries	Zhu et al. (2025)	
	Crowdsourced	Recall, Precision, Faithfulness	Source (Dialog), Output Summaries	Huang et al. (2023)	
	Not Specified	Not Specified	Not Specified	Eom et al. (2025)	
	Crowdsourced Domain Experts	Readability, Informativeness	Output Summaries	Feigenblat et al. (2021)	
QA-Based Eval	Crowdsourced	Podcast Specifics, Language, Redundancy	Transcript, Output Summary	Song et al. (2022b)	6
	Graduate Students	Challenges in Transcript	Transcripts, Output Summary Reference	Kirstein et al. (2024b)	
	System Users	Comprehension	Audio, Output Summary	Koumpis and Renals (2005)	
	Not Specified	Informativeness, Factual Accuracy	Transcripts or Output Summary	Zechner and Waibel (2000a)	
	Graduate Students	Grammatical Correctness, Semantic Comprehensibility	Audio, Transcript, Output Summary	Li et al. (2021)	
	Crowdsourced Domain Experts	Informativeness, Saliency, Readability	Transcripts, Output Summary	Feigenblat et al. (2021)	
MOS Score	System Users	Not Specified	Subset of Transcript, Output Summary	Koumpis and Renals (2005)	2
	Not Specified	Relevance	Transcript, Output Summarys	Chowdhury et al. (2024)	
Accuracy	Not Specified	Readability	Sentences of Output Summary	Banerjee et al. (2015)	2
	Domain Experts	Factual Accuracy	Transcript, Sentences of Output Summary	Tang et al. (2024)	
Absolute Score	Domain Experts	Relevance, Completeness	Transcript, (Topic,) Output Summary	Tang et al. (2024)	2
	Not Specified	Discourse Relations, Intent, Coreference	Source (Dialog), Output Summary	Chen et al. (2021)	

Table 7: Different ways of human evaluation for SSum, based on the SSum papers in this survey.

C Supplementary Statistics

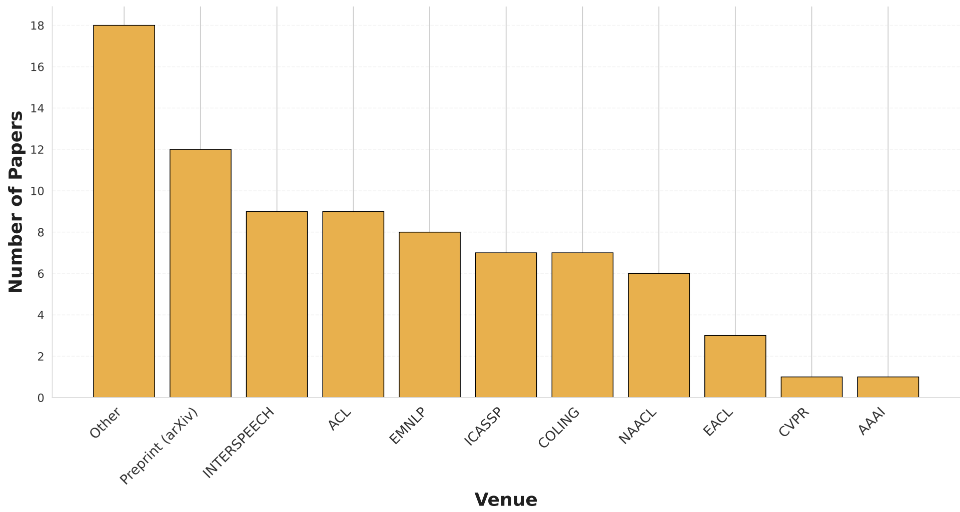

Figure 4: Total number of SSum papers published in different venues, based on the Ssum papers included in this survey.