
White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/247730/

Version: Published Version

Proceedings Paper:
Chun, C., Zhang, D.C. and Lee, D. (2026) When misinformation speaks and converses: rethinking fact-checking in audio platforms. In: Liakata, M., Moreira, V.P., Zhang, J. and Jurgens, D., (eds.) Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 02-07 Jul 2026, San Diego, California, United States. Association for Computational Linguistics, pp. 2060-2075. ISBN: 9798891763906. ISSN: 0736-587X. EISSN: 9061-6464.

https://doi.org/10.18653/v1/2026.acl-long.93

© 2026 Association for Computational Linguistics (ACL). Licensed on a Creative Commons Attribution 4.0 International License.
(https://creativecommons.org/licenses/by/4.0/)

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.





When Misinformation Speaks and Converses: Rethinking Fact-Checking in Audio Platforms

Chaewan Chun[1] **and Delvin Ce Zhang**[2] **and Dongwon Lee**[1]
[1]The Pennsylvania State University, USA
[2]University of Sheffield, UK
czc5884@psu.edu, delvin.ce.zhang@sheffield.ac.uk, dongwon@psu.edu

Abstract

Audio platforms have evolved beyond entertainment. They have become central to public discourse, from podcasts and radio to WhatsApp voice notes and live streams. With millions of shows and hundreds of millions of listeners, audio platforms are now a major channel for misinformation. Yet existing fact-checking pipelines are mostly designed for written claims, overlooking the unique properties of spoken media. We argue that audio misinformation is not merely textual content with transcripts: it is structurally different because it is both *spoken*—carrying persuasive force through prosody, pacing, and emotion—and *conversational*—unfolding across turns, speakers, and episodes. These dual properties introduce verification difficulties that traditional methods rarely face. This *position paper* synthesizes evidence across modalities and platforms, examines datasets and methods, and highlights why existing pipelines fail on audio. We argue that advancing fact-checking requires rethinking verification pipelines around the spoken and conversational realities of audio.

1 Introduction

Spoken media have become a dominant channel for news and commentary. From WhatsApp voice notes and live-streamed talk shows to podcasts, audio platforms now command hours of daily attention, shaping public discourse at scale. Podcasts alone now exceed 4.3 million distinct shows, reaching an estimated 500 million listeners globally, with average consumption of about seven hours per listener each week.[1] Podcast listenership has grown steadily year over year, reflecting how audio has shifted from a niche medium to a mainstream source of information and opinion.[2] Yet while audience adoption accelerates, research

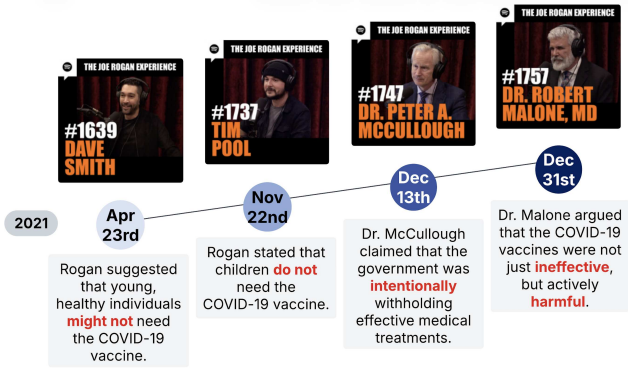


Figure 1: Timeline of cross-episode misinformation on COVID-19. These episodes illustrate how persistent and unchallenged misinformation can accumulate over time. Red text highlights false claims in audio content. Image credit: Spotify

on fact-checking methods has not kept pace. Automated fact-checking remains text-centric (Yao et al., 2023), leaving the spoken and conversational dimensions of misinformation underexplored.

The consequences are clear: when false claims circulate in podcasts, radio shows, or private group chats, they spread without the anchoring mechanisms available in text (hyperlinks, citations, overlays). Delivery through tone and pacing lends credibility (Guyer et al., 2021), while repetition across turns and episodes reinforces false narratives (see Figure 1). A single falsehood voiced by a familiar speaker can gain persistence and authority that written text rarely achieves.

What makes audio platforms distinctive and challenging is that they are both *spoken* and *conversational*. As spoken media, they carry persuasive force through prosody, pacing, and emotion, often making content sound compelling regardless of veracity. As conversational media, they distribute meaning across dialogue history, with context-dependent reference, stance shifts, and cumulative repetition that complicate verification. These dual properties introduce verification demands absent

[1] https://www.demandsage.com/podcast-statistics/
[2] https://podcastatistics.com/

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2060–2075
July 2-7, 2026 ©2026 Association for Computational Linguistics

Spoken	Conversational
Prosody & emotion	Adjacency & repairs
Continuous signal	Distributed claims
Incidental sounds & overlap	Chronology & callbacks
Serialized in-flow listening	Narrative episodes
Acoustic persuasion & detection	Parasocial trust

Table 1: Spoken and conversational properties of audio misinformation and the distinct challenges they introduce for fact-checking.

from other modalities, underscoring why audio cannot be treated as *just another transcript*.

We focus on content-level fact-checking in spoken, conversational audio: given what is said, we ask how systems should detect claims, find relevant evidence, decide whether those claims are true, and explain their decisions. We do not address media-integrity questions such as whether an audio signal or voice is synthetic or cloned; those are the target of deepfake detection methods. Our contribution is to show how fact-checking systems themselves must be redesigned for long-form, dialogic speech—where meaning is shaped by delivery, turn-taking, and interaction—regardless of whether the audio was produced by a human or a model.

Fact-checking began as a language-centered problem: systems identified claims in text, retrieved evidence from encyclopedic sources, assessed veracity, and increasingly produced textual rationales (Aly et al., 2021; Jiang et al., 2020; Thorne et al., 2018; Atanasova et al., 2020). With audio platforms now mainstream, we extend the established approach to spoken dialogue. We (1) show why audio platforms matter and how they differ from other media, (2) survey datasets spanning audio, dialogue, and factuality, highlighting their coverage and limitations, and (3) analyze why existing pipelines fall short and review emerging efforts toward audio-based fact-checking. Taken together, these contributions aim to reposition audio platforms not as a blind spot but as the next frontier for fact-checking research.

2 Why Audio Platforms Matter

Audio platforms reshape how misinformation spreads and how fact-checking must operate. Unlike standalone text, spoken dialogue carries meaning in delivery and interaction. As a *spoken* medium, prosody, pacing, and vocal cues alter perception; as a *conversational* medium, meaning unfolds across turns, speakers, and episodes. These properties simultaneously amplify persuasion and complicate verification, requiring methods that listen to delivery and track dialogue (Table 1).

2.1 Spoken Misinformation

Audio persuades differently from text. Even with identical words, prosody (pitch, intensity, timing), pacing, and emotion shift judgments of confidence, credibility, and intent. Listeners infer trustworthiness from vocal delivery (Guyer et al., 2021); deception studies show that speech rate, intensity, pitch variation, and hesitations influence truthfulness judgments (Jaiswal et al., 2016; Bahaa et al., 2024); and even eyewitness testimony accuracy has been linked to vocal cues, indicating that delivery can act as a heuristic shortcut (Gustafsson et al., 2023). At scale, analyses of 88,000 podcast episodes show that vocal qualities (e.g., energy, seriousness) predict engagement, with adversarial acoustic features outperforming common baselines (Yang et al., 2019). These persuasion effects interact with interpretation. Irony perception depends on prosody; without auditory cues, sarcasm is often taken literally (Matsui et al., 2016; Farabi et al., 2024; Li et al., 2025b). On audio platforms, sarcastic and ironic intent relies on delivery (Rao et al., 2022; Li et al., 2025a), which can tilt credibility judgments and downstream decisions. Together, these results place prosodic and other paralinguistic features at the center of how people absorb, interpret, and evaluate information.

Meaning in audio unfolds as a continuous signal. Claims often span prosodic units rather than tidy sentence boundaries. Hesitations, elongations, and timing cues can split the relevant proposition across adjacent acoustic spans, even within a single speaker's turn. Speakers slow down and insert disfluencies before more surprising words, effectively 'buying time,' and speech rate adapts to information content (Bergey and DeDeo, 2024). Podcast discourse further complicates scope: episodes are long, digressive, and conversational, with ads and side-tracks that introduce long-range dependencies and noisy content—making chapterization/segmentation a prerequisite for downstream tasks like retrieval (Abdessamed et al., 2024; Ghazimatin et al., 2024).

Incidental sounds (laughter, music, effects) and overlapped speech (interruptions, backchannels, crosstalk) can obscure boundaries and degrade segmentation/diarization, motivating explicit overlap

detection (Bredin and Laurent, 2021; Sun et al., 2025). Serialized listening sustains narratives over time; surveys report about 80% of listeners finish most episodes they start (Caramancion, 2022). Case studies show the same dynamics across platforms: WhatsApp voice notes in Portugal circulated urgent COVID-19 rumors and policy critiques (Cardoso et al., 2022), while an analysis of leading U.S. political podcasts documented frequent false or unsubstantiated claims reaching large audiences (Wirtschafter, 2023). Unlike text feeds where visual labels sit on posts, podcasts and streams are consumed in-flow, so the persuasive payload reaches listeners during listening itself (Pathiyan Cherumanal et al., 2024).

The same cues that make audio persuasive also enable detection. Because voices carry stance and identity cues—intonation, stress, disfluencies—acoustic evidence can supplement or even surpass transcripts for certain phenomena. In political debates, adding acoustic features significantly improves identification of check-worthy claims, and in some conditions audio-only models outperform text-only systems (Ivanov et al., 2024). Multimodal sarcasm corpora likewise find prosody outperforms text for irony and emotion (Li et al., 2025a). More broadly, multimodal misinformation frameworks find that acoustic encoders capture semantic-level cues that transcripts miss, yielding gains when fused with other modalities (Liu et al., 2024); and while standalone deception cues can be noisy, features like jitter, pauses, and pitch instability become informative when integrated with lexical and visual signals (Jaiswal et al., 2016; Bahaa et al., 2024). These observations motivate audio-aware fact-checking that listens to delivery and sequence, not only to words.

2.2 Conversational Misinformation

Conversation changes what a "claim" is. In dialogue, propositions are negotiated through adjacency, uptake, repairs, hedges, and challenges; a first mention is rarely the final form. Claims are also distributed and compositional—parts contributed at different turns and by different speakers that only add up to a verifiable unit when considered together (Chamoun et al., 2023; Deng et al., 2024). Chronology further shapes judgment: primacy and recency biases anchor interpretation, and repeated callbacks increase availability and felt plausibility (Kiesel et al., 2021); podcast episodes often include narrative content and recurring formats, which can sustain storylines over time (Abdessamed et al., 2024). Social roles and bonds intensify these effects: hosts and recurring guests shape expectations and perceived credibility; repeated agreement can signal consensus, and parasocial trust makes familiar voices feel credible (Yorganci and McMurtry, 2024; Vilceanu, 2025). In short, conversational structure itself—how things are said, when they are said, and by whom—drives persuasion independent of the words alone.

Misinformation leverages these dynamics. On messaging platforms, misleading WhatsApp voice notes follow repeatable rhetorical templates—greeting, insider/expert positioning, emotional appeals, a central false assertion, and a call-to-action (Cardoso et al., 2022; El-Masri et al., 2022). Long-form podcasts exploit repetition and serialized exposure: narratives are seeded, revisited, and strengthened across episodes. For example, *The Joe Rogan Experience*, one of the world's most influential podcasts (Figure 1), advanced false COVID-19 vaccine narratives across multiple episodes. In April, he suggested young, healthy people might not need vaccinations (Geddes, 2022); by November, he argued against vaccinating children (Paterson, 2021); and in December, with guests Dr. Peter A. McCullough and Dr. Robert Malone, he claimed vaccines were ineffective or harmful (Team, 2022; Teoh, 2021). This cross-episode repetition amplified misinformation through cumulative exposure. In January 2022, over 270 U.S. health professionals issued an open letter urging platform action (Yang, 2022). These dynamics scale beyond a single program. U.S. podcast studies describe "toxic conversation chains"—emotionally charged exchanges that sustain harmful narratives within episodes—and document that such toxicity recurs across many episodes (Rizwan et al., 2025).

These dynamics complicate verification as well as persuasion. Empirically, systems perform markedly better when given dialogue context rather than isolated utterances (Chun et al., 2026), and human annotators likewise use surrounding context to judge check-worthiness (Gadiraju et al., 2024). To make this more concrete, we draw on MAD2, a benchmark of 1,000 two-speaker English dialogues (about 10 hours of audio) with 8,192 sentences and 3,368 check-worthy claims, each annotated with a binary true/false label (Chun et al., 2026). Each dialogue is accompanied by a full transcript, and every claim is aligned to its spoken span us-

Metric	Setting	Past 15 sec	Full dialogue
F1	Base	0.6799	0.7140
	Shuffled	0.6590	0.7041
AUC	Base	0.7408	0.7924
	Shuffled	0.7203	0.7780

Table 2: Illustrative ablation on MAD2 (Chun et al., 2026): F1 and AUC for text-only claim verification under two context regimes. Rows compare a *Base* model that reads turns in chronological order versus a *Shuffled* variant that randomizes non-claim turn order while keeping content fixed. Shuffling reduces performance in both settings, implying that chronological turn order carries predictive signal beyond lexical content.

ing word-level timestamps, so models can operate over precise seconds-based context windows rather than only sentence IDs. The dialogues are multi-turn, speaker-attributed sequences, and in our analysis, we consider two simple context regimes for a standard text-only verifier: (i) only the 15 seconds preceding the claim, and (ii) the full dialogue surrounding the claim.

To isolate the role of order, we compare a base model to a shuffled-turns variant that randomizes the non-claim utterances within each dialogue while keeping the lexical content fixed on MAD2. For each context regime (past 15 seconds vs. full dialogue), Table 2 reports F1 and AUC for the base model and its shuffled counterpart. Across both regimes, shuffling consistently reduces performance despite identical words, indicating that temporal order and pacing carry predictive signal beyond lexical content. This small, illustrative experiment complements prior findings that speech rate and timing track information structure (Bergey and DeDeo, 2024), and supports our broader claim that conversational order and accumulation shape what both verifiers and annotators perceive as check-worthy and true.

3 Datasets

Research into fact-checking for audio platforms remains relatively nascent. As summarized in Table 3, most corpora address only subsets of these dimensions and rarely combine all four dimensions—audio, dialogue, transcripts, and factuality.

Large-scale speech corpora exist but target automatic speech recognition (ASR) or representation learning, not verification. Read audiobooks (LibriSpeech (Panayotov et al., 2015), LibriHeavy (Kang et al., 2024)), multilingual parliamentary speeches (VoxPopuli (Wang et al., 2021)), and English ASR collections (People's Speech (Galvez et al., 2021)) offer thousands to hundreds of thousands of hours, yet lack claim units, timestamped rationales, and veracity labels. These corpora improve speech modeling, not veracity evaluation.

Dialogue benchmarks have expanded but remain text-only. Early datasets (Ubuntu (Lowe et al., 2015), DailyDialog (Li et al., 2017), Persona-Chat (Zhang et al., 2018)) advanced turn-taking, persona, and emotion, while later work emphasized empathy, knowledge grounding, or summarization (e.g., Topical-Chat (Gopalakrishnan et al., 2019), Wizard of Wikipedia (Dinan et al., 2019), SAMSum (Gliwa et al., 2019)). These corpora are invaluable for conversational modeling, but they lack audio and do not provide veracity labels.

A separate line of work scales spoken dialogue without factuality. These corpora combine audio, diarization, and transcripts—supporting prosody, overlap, and long-context modeling—but they do not label whether claims are true. Representative examples span emotion/video-grounded conversations (MELD (Poria et al., 2019); AVSD (Alamri et al., 2019)), multiparty household talk captured with distant microphones (CHiME-6 (Watanabe et al., 2020)), and multi-speaker earnings-call speech (SPGISpeech (O'Neill et al., 2021)); large long-form conversational releases broaden duration and domains (e.g., Spotify Podcasts (Clifton et al., 2020); CANDOR (Reece et al., 2023)). Together, they are essential for modeling how something is said, not for assessing whether it is true.

Where factuality is present, coverage is often partial or limited to a single modality. Some datasets provide check-worthiness only—flagging what to fact-check but not whether it is true (PHEME (Zubiaga et al., 2015); ClaimBuster (Hassan et al., 2017); ViClaim (Giedemann et al., 2025)). Others include veracity but miss key modalities: CT-FCC-18 (Kopev et al., 2019) aligns short debate audio to fact-checked claims without full dialogue context, while DialFact (Gupta et al., 2022) verifies claims in textual dialogues with no audio. Related work benchmarks consistency rather than truth (Qin et al., 2021), and WhatsApp 2019 (Resende et al., 2019) adds veracity assessments to media shared in chats but does not capture audio. Across these resources, the common gaps include variable claim granularity, limited timestamping, narrow domains, and weak multi-evidence support.

Only recently have datasets appeared that cover audio, dialogue, transcripts, and veracity. Setty

Dataset	Audio	Dialogue	Transcript	Factuality	Size	Lang	Domain
LibriSpeech (Panayotov et al., 2015)	x	–	x	–	1k hrs	En	audiobooks
VoxPopuli (Wang et al., 2021)	x	–	x	–	400k hrs raw / 19.1k hrs labeled	Multi	parliamentary speeches
People's Speech (Galvez et al., 2021)	x	–	x	–	30k+ hrs	En	diverse speech
LibriHeavy (Kang et al., 2024)	x	–	x	–	50k hrs	En	audiobooks
SSSD (Sheikh et al., 2025)	x	x	–	–	700+ hrs	En	everyday conv.
Ubuntu Corpus (Lowe et al., 2015)	–	x	–	–	930k dialogs	En	tech support
DailyDialog (Li et al., 2017)	–	x	–	–	13.1k dialogs	En	daily chat
Persona-Chat (Zhang et al., 2018)	–	x	–	–	11k dialogs / 164k utt.	En	open-domain, persona
MultiWOZ (Budzianowski et al., 2018)	–	x	–	–	10k dialogs	En	task-oriented, multi-domain
GroundedConv (Zhou et al., 2018)	–	x	–	–	4.1k dialogs	En	Wikipedia-grounded movie chat
Topical-Chat (Gopalakrishnan et al., 2019)	–	x	–	–	11.3k dialogs	En	knowledge-grounded chat
Wizard of Wikipedia (Dinan et al., 2019)	–	x	–	–	22k dialogs / 202k utt.	En	knowledge-grounded chat
SAMSum (Gliwa et al., 2019)	–	x	–	–	16.4k dialogs	En	messenger chat, abstractive summarization
EmpatheticDialogues (Rashkin et al., 2019)	–	x	–	–	25k dialogs	En	emotion-grounded chat
BlendedSkillTalk (Smith et al., 2020)	–	x	–	–	6.8k dialogs	En	open-domain, blended skills
MultiWOZ 2.1 (Eric et al., 2020)	–	x	–	–	10k dialogs	En	task-oriented, multi-domain
MedDialog (Zeng et al., 2020)	–	x	–	–	0.26m / 3.4m dialogs	En/Zh	medical
WhatsApp 2021 (Maros et al., 2021)	–	x	–	–	298k msgs	Pt	political WhatsApp groups, link/media typology
MediaSum (Zhu et al., 2021)	–	x	x	–	463.6k dialogs, summaries	En	interviews (NPR/CNN)
CMCC (Huang et al., 2022)	–	x	x	–	100k dialogs (8.9k labeled)	Zh	customer service
HANSEN (Tripto et al., 2023)	–	x	x	–	17 datasets / ∼21k AI samples	En	spoken authorship
Audio Dialogues (Goel et al., 2024)	–	x	–	–	163.8k dialogs	En	audio/music understanding
Liu2025 Bilingual Dialogue (Liu et al., 2025)	–	x	–	–	see repo	En/Zh	personality, emotion
MELD (Poria et al., 2019)	x	x	x	–	1.4k dialogs / 13k utt.	En	TV show (Friends), emotion recognition
AVSD (Alamri et al., 2019)	x	x	x	–	11.8k dialogs / 118k QA pairs	En	video-grounded daily activities
Spotify Podcasts (Clifton et al., 2020)	x	x	x	–	60k hrs / 100k+ eps	En	podcasts
CHiME-6 (Watanabe et al., 2020)	x	x	x	–	40+ hrs	En	dinner-party conversations, conversational ASR
DiPCo (Segbroeck et al., 2020)	x	x	x	–	5.3 hrs / 10 sessions	En	dinner-party conversations
SPGISpeech (O'Neill et al., 2021)	x	x	x	–	5k+ hrs	En	earnings calls
Earnings-21 (Del Rio et al., 2021)	x	x	x	–	39 hrs	En	earnings calls
MD3 (Eisenstein et al., 2023)	x	x	x	–	20 hrs	En	information-sharing tasks
CANDOR (Reece et al., 2023)	x	x	x	–	850 hrs / 1,656 dialogs	En	everyday conv. (video chat)
DailyTalk (Lee et al., 2023)	x	x	x	–	20 hrs / 2,541 dialogs	En	conversational TTS
SPoRC (Litterer et al., 2024)	x	x	x	–	1.1m eps	En	podcasts
MultiDialog (Park et al., 2024)	x	x	x	–	340 hrs / 9k dialogs	En	open-domain, audiovisual
SPGISpeech2 (Grossman et al., 2025)	x	x	x	–	3.78k hrs	En	earnings calls
CASPER (Xiao et al., 2025)	x	x	x	–	3 hrs	En	spontaneous conv.
DeepDialog (Koudounas et al., 2025)	x	x	x	–	488 hrs / 40.2k dialogs	En	41 domains, 20 emotions
PHEME (Zubiaga et al., 2015)	–	x	–	*	1,185 threads	En	Twitter rumors
ClaimBuster (Hassan et al., 2017)	–	x	x	*	23k sentences	En	political debates
Audio Check-Worthiness (Ivanov et al., 2024)	x	x	x	*	48 hrs / 34.5k sents	En	political debates, speeches, interviews
ViClaim (Giedemann et al., 2025)	–	–	x	*	1.8k videos / 17.1k sents	En/De/Es	YouTube short videos, claim detection
CT-FCC-18 (Kopev et al., 2019)	x	–	x	x	33 min / 286 claims	En	political debates
WhatsApp 2019 (Resende et al., 2019)	–	x	–	x	912k msgs	Pt	WhatsApp political groups
CI-ToD (Qin et al., 2021)	–	x	–	x	3,190 dialogs	En	task-oriented, consistency labels (HI/QI/KBI)
DialFact (Gupta et al., 2022)	–	x	–	x	22.2k claims	En	fact-checking in dialogue
Fact-Checking Podcasts (Setty and Becker, 2025)	x	x	x	x	531 eps / 2.0k utt. (annot.)	En/No/De	podcasts (news, health)
MAD (Chun et al., 2025)	x	x	x	x	600 dialogs / 4.9k sents	En	spoken dialogue
MAD2 (Chun et al., 2026)	x	x	x	x	1k dialogs / 8.2k sents / word-level ts	En	spoken dialogue

Table 3: Representative datasets across audio, dialogue, transcript, and factuality dimensions.
*: check-worthiness only (no veracity labels).
Abbrev.: QA = question answering; TTS = text-to-speech; ts = timestamps.

and Becker (2025) targets podcasts with transcripts plus check-worthiness and supports/refutes annotations. MAD (Chun et al., 2025) introduces multi-turn spoken dialogues with aligned audio and veracity labels, and MAD2 (Chun et al., 2026) provides roughly 1,000 dialogues with thousands of check-worthy claims and word-level timestamps. Despite progress, current resources remain small-scale, English-dominant, and narrow in domain coverage—underscoring the urgent need for large-scale, multimodal audio fact-checking datasets.

4 Where Traditional Pipelines Fail

Most fact-checking systems follow a four-stage pipeline: Claim Detection (CD) identifies check-worthy statements; Evidence Retrieval (ER) queries trusted sources to gather passages relevant to the claim; Claim Verification (VER) compares the claim against the retrieved evidence to assign a verdict (supported, contradicted, or insufficient) and a confidence score; and Explanation Generation (GEN) produces a human-readable rationale that highlights the evidence and explains the verdict. Building on our earlier discussion of how audio platforms differ, we now examine where this pipeline breaks for spoken dialogue. Table 4 summarizes typical models, their failure modes on spoken dialogue, and the design requirements we argue for at each stage of the pipeline.

4.1 Claim Detection (CD) Task

Traditional CD treats one sentence as a single proposition, but conversational audio rarely obliges: claims spread across adjacent turns, carry hedges, or sit inside Q&A and anecdotes, so single-turn sentence classifiers miss the claim or mis-scope its span (Chamoun et al., 2023; Deng et al., 2024). Text-only detectors ignore how delivery moves

meaning: sarcasm, emphasis, emotion, and intent live in prosody—pitch, timing, and intensity—so models that ignore nonverbal speech cues misread intent (Biron et al., 2025; Ananthakrishnan and Narayanan, 2008). Familiar hosts, recurring guests, and conversational role structure (e.g., host/guest/caller) shape interaction patterns and perceived salience, so CD should encode who is speaking and how they interact, not just what they say (Qamar et al., 2023; Pick et al., 2022; Mahr and Csibra, 2021). Treating turns independently discards uptake, repairs, and repetition trajectories that strengthen or revise a proposition over time (Qamar et al., 2023). Empirically, order carries signal: shuffling turns weakens downstream prediction and, on MAD2, degrades accuracy (Table 2) (Bergey and DeDeo, 2024; Chun et al., 2026). Relatedly, argument-aware summarization and keypoint analysis aggregate across turns to surface conversation-level claims not stated in any single sentence (Fabbri et al., 2021).

Beyond local phrasing, conversational claims depend on dialogue history: meaning is often incomplete without preceding turns, and delivery can amplify perceived salience independent of truth. For instance, misleading WhatsApp voice notes often follow a repeatable arc—opening with a personal greeting, asserting source credibility (insider/expert/eyewitness), leveraging negative emotional tone (panic, fear, or anger), delivering the claim, and—in about one-third of cases—urging recipients to forward it (El-Masri et al., 2022). When flattened to text, these cues disappear, and models mistake rhetorical mobilization for factual salience—a recurring source of false positives and missed context.

Upstream artifacts further distort the "claim unit." ASR segmentation can split or merge spans; diarization errors swap speakers; and overlapped talk stresses voice activity detection (VAD) and segmentation, where thresholding can clip turn onsets/offsets - each of which destabilizes CD in long-form talk (Wan et al., 2021; Bredin and Laurent, 2021; Bain et al., 2023; Sun et al., 2025). Real-world conditions magnify the problem: Whisper reports $\sim$2.5-2.7% word error rate (WER) on LibriSpeech test-clean (Radford et al., 2023), yet podcast corpora reach $\sim$18.1% WER and include non-speech/extraneous segments that must be filtered (Clifton et al., 2020; Abdessamed et al., 2024). Earlier end-to-end ASR and short-window decoders degrade under noise; in contrast, large-scale pretraining, timestamped decoding/VAD (e.g., Whisper), and forced alignment (WhisperX) improve robustness and alignment in noisy or low-resource settings (Radford et al., 2023; Bain et al., 2023). Dataset mirrors this: sentence-level check-worthiness often fails under ellipsis but improves with decontextualization (Chamoun et al., 2023); audio/disfluency cues can match or beat text-only baselines in multi-speaker setups (HuBERT (Hsu et al., 2021), wav2vec 2.0 (Baevski et al., 2020)) (Ivanov et al., 2024); yet speech-native CD datasets remain small (Ivanov et al., 2024).

Methodologically, CD research has moved from early feature-based systems (Hassan et al., 2017; Jaradat et al., 2018) to neural rankers and pre-trained language models such as BERT and RoBERTa (Devlin et al., 2019; Liu et al., 2019), with widespread adoption in recent CD work (Panchendrarajan and Zubiaga, 2024). Subsequent work explores claim-attribute modeling (Gangi Reddy et al., 2022) and impact-aware prioritization (Panchendrarajan and Zubiaga, 2024). Yet most work remains source- and modality-specific (Panchendrarajan and Zubiaga, 2024).

To close this gap, several strands of work are converging. *Conversation-first datasets* such as DialFact explicitly surface colloquialisms, ellipsis/coreference, and context dependence that challenge sentence-isolated detectors, enabling dialogue-aware baselines and evaluation (Gupta et al., 2022). *Claim-unit reconstruction* via decontextualization assembles a self-contained proposition before scoring and improves extraction quality and readiness for downstream components (Deng et al., 2024; Fan et al., 2025). *Prosody- and role-aware modeling* incorporates pitch/timing/intensity cues and speaker/interaction structure to better capture communicative function and salience (Biron et al., 2025; Qamar et al., 2023; Pick et al., 2022). Finally, live CD systems (e.g., LiveFC) bring streaming transcription and online diarization together with windowed claim detection/normalization, so detectors consume temporally ordered, speaker-attributed candidates rather than isolated sentences (Venktesh and Setty, 2025). In practice for CD: model dialogue acts and local context (Qamar et al., 2023); encode speaker roles and interaction structure (Qamar et al., 2023; Pick et al., 2022); condition on prosody (Ananthakrishnan and Narayanan, 2008; Biron et al., 2025); and aggregate adjacent turns into a standalone claim unit (Deng et al., 2024).

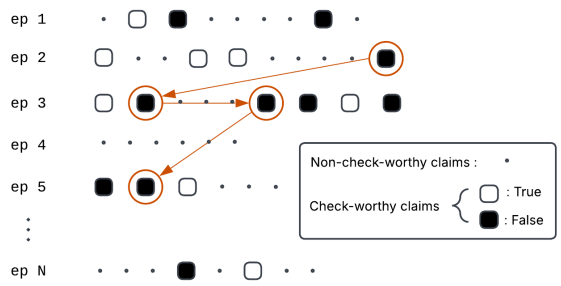


Figure 2: Illustration of claim detection and verification tasks in podcasts.

4.2 Evidence Retrieval (ER) Task

Traditional ER is built for clickable text: models often assume sentence-level evidence, as in FEVER, a standard textual fact-checking benchmark (Thorne et al., 2018), that lexical retrieval can fetch as one or more sentence spans. Spoken platforms break these assumptions. Audio is continuous and not clickable: meaning is distributed across turns, speakers, and episodes, so evidence must be located in time, not just on a page (Ghazimatin et al., 2024). As a result, single-turn lexical queries miss callbacks, cross-turn references, and off-mic context; colloquialisms, slang, and code-switching change the retrieval error profile (Clifton et al., 2020; Shapira et al., 2025); and pipelines that ignore chapter titles and episode-level metadata (titles/descriptions/RSS IDs (Really Simple Syndication)) leave strong production-metadata constraints unused (Ghazimatin et al., 2024; Shapira et al., 2025). Span-level claim extraction can also improve retrieval relative to full-post queries (Sundriyal et al., 2022). To keep verification auditable, ER therefore has to return time-anchored evidence—an episode identifier plus start/end seconds—so downstream components can attribute who said what, when, and support revisions when on-air corrections appear (Ghazimatin et al., 2024).

In real deployments, upstream artifacts complicate this. ASR segmentation can split or merge spans, diarization uncertainty blurs speaker identity, and overlapped talk stresses VAD/segmentation—each interacts with lexical retrieval and can hide the very moments VER needs (Bredin and Laurent, 2021; Shapira et al., 2025; Sun et al., 2025). Confidence-only filters are unreliable for detecting ASR errors (Kuhn et al., 2025). Crucially, error types (punctuation, proper nouns (Singla et al., 2022), code-switch points)—not just average WER—drive indexing and alignment errors (Shapira et al., 2025; Clifton et al., 2020). These conditions push ER beyond plain text toward the selective use of acoustic/phonetic cues and tighter use of production metadata.

Resources mirror the gap. Wizard-of-Wikipedia offers turn-level grounding but presumes explicit passages rather than timestamped audio (Dinan et al., 2019); ClaimBuster and DialFact add check-worthiness/verification labels yet are largely text-only (Hassan et al., 2017; Gupta et al., 2022). Speech-native efforts go further: Audio Check-Worthiness couples speech and claims, and the Fact-Checking Podcast dataset preserves ASR and diarization with podcast-native claims, check-worthiness, and rationales—though all remain small relative to text corpora (Ivanov et al., 2024; Setty and Becker, 2025). Together, these efforts point toward retrieval that treats audio as a continuous, produced medium.

Methodologically, a speech-native pipeline adopts: *timestamped spans* (transcript slice + episode ID + start/end seconds), indexed with *dual indices* (lexical + optional acoustic/phonetic), then *re-ranking by temporal proximity, speaker match, and meta-structure* (chapters/show notes/episode metadata) (Ghazimatin et al., 2024), while maintaining *long-horizon memory* so recurring assertions can be linked across episodes. In effect, ER outputs a compact bundle (episode, timestamp, transcript slice, speaker posterior, etc.) that VER can consume without re-resolving timing or speaker identity.

4.3 Claim Verification (VER) Task

Verification tuned on text typically frames each claim–passage pair as single-span natural language inference (NLI), but——as Figure 2 illustrates—podcast evidence is time-anchored, multi-turn, and often cross-episode, making FEVER-style single-sentence models brittle in practice (Thorne et al., 2018). Specifically, multi-turn entailment and temporal qualifiers (“last week”, “earlier in the show”) spill beyond a single span, and VER should support revision when on-air corrections appear. Small timing misalignments and diarization uncertainty further depress multimodal performance unless modeled explicitly (Allein et al., 2023; Bredin and Laurent, 2021; Sun et al., 2025). Empirically, FEVER-tuned models degrade in dialogue; DialFact reports sharp drops from cross-turn references and informality (Gupta et al., 2022); and CI-ToD surfaces contradictions arising from dialogue history itself

(Qin et al., 2021). Throughput and latency compound the challenge: re-ranking thousands of candidates with large NLI models is costly on continuous, claim-dense streams, and on live platforms, timing determines whether interventions land in time (Venktesh and Setty, 2025; Deng et al., 2025).

In response, VER shifts to: *multi-span, time-aware reasoning* over sets of spans; explicit *who/when constraints* to resolve callbacks, coreference, and temporal qualifiers (Ma et al., 2024; Si et al., 2023; Allein et al., 2023); *alignment-robust fusion*—late/gated combination of audio encoders with text plus masking of diarization uncertainty—so mild desynchronization does not derail entailment (Shahriar et al., 2025; Bredin and Laurent, 2021; Sun et al., 2025; Park et al., 2022); and *WER-aware training*: mixing oracle and ASR transcripts with audio-only views and adding punctuation/noise perturbations, recognizing that error type—not just average WER—drives downstream failure (Allein et al., 2023; Shapira et al., 2025), with practices that carry over from noisy-transcript analyses (Shapira et al., 2025; Wan et al., 2021).

Efficiency-focused designs make deployment feasible. Lightweight verifiers such as MiniCheck approach GPT-4-level verification at $\sim$400$\times$ lower cost (Tang et al., 2024), while streaming systems like LiveFC run real-time transcription, online diarization, retrieve timestamped windows, and perform incremental detection/verification online (Venktesh and Setty, 2025). Complementary probes—Debate-to-Detect and DEFAME—stress debate structure and cross-modal cues in realistic, structured settings (Han et al., 2025; Braun et al., 2025). Finally, *attribution and recoverability* practices from modular retrieval-augmented generation (RAG)—self-reflective retrieval and recoverability audits—transfer naturally to timestamped audio bundles so cited moments can be re-found and audited end-to-end (Asai et al., 2024; Wallat et al., 2025; Xing et al., 2025).

4.4 Explanation Generation (GEN) Task

GEN fails on audio when it treats explanations as text-only. Without audio, listeners cannot hear how delivery shapes interpretation (sarcasm, emphasis, hesitation). Without timestamps and speaker labels, rationales are hard to audit or replay. Explanations often ignore the conversational path - who challenged whom and when - so they cannot justify mid-stream verdict changes in multi-turn dialogue (Gupta et al., 2022). Multimodal fact-checking and deception-style probes show that adding non-text modalities improves robustness by capturing delivery-related cues (Bahaa et al., 2024). Meanwhile, the literature increasingly explores retrieval-augmented evidence pipelines and end-to-end systems that jointly produce verdicts and rationales (Atanasova et al., 2020; Allein et al., 2023), alongside modern large language model (LLM) prompting frameworks for generation and detection (Lucas et al., 2023). For audio platforms, those strengths must be *audio-native, timestamped, and dialogue-aware* (Deng et al., 2025; Venktesh and Setty, 2025).

A practical design is to return evidence you can listen to. Each explanation should pair a readable transcript span with a short, click-to-hear clip (3-10s) and explicit speaker attribution. Optionally flag salient prosodic cues so users can judge delivery, and include a brief decontextualized gloss so the claim remains self-contained for readers without losing its audio anchor (Deng et al., 2024). Explanations should present a compact *timeline* - first assertion $\rightarrow$ challenge $\rightarrow$ repair - anchored by episode/turn indices, and bundle the timestamped audio/transcript with any external documents so the cited moment can be re-found and audited (Allein et al., 2023; Ghazimatin et al., 2024). Live systems already point the way: LiveFC surfaces speaker-attributed snippets with verdicts in real time, and modular/two-stage pipelines - where lightweight detectors gate heavier generators - can deliver rationales at acceptable latency in streaming contexts (Venktesh and Setty, 2025; Henrichsen and Krebs, 2025; Xiong et al., 2023). *Recoverability-based attribution* (mask the cited span and test whether the system can restore it) stabilizes explanations and transfers naturally to timestamped audio bundles (Xing et al., 2025). In short, when GEN embraces audio-native affordances and recoverability, it complements timestamped retrieval and time-aware verification rather than inheriting the blind spots of text-only explanations.

5 Future Directions

Moving the field forward requires a few pieces to work together. First come the foundations: build larger speech-native corpora with factuality or check-worthiness labels, aligned transcripts, and speaker/role metadata; augment them with targeted synthetic variants (accents, overlap, punctuation/noise). Then make the pipeline live: pair

Task	Representative Models	Limitations on Spoken Dialogue	Proposed Design
CD	Sentence-level check-worthiness and claim classifiers over text (BERT-/RoBERTa-style rankers).	Assume one sentence ≈ one claim; miss claims spread across turns, Q&A, and anecdotes; ignore prosody and speaker roles; brittle under ASR/diarization errors and high WER in podcasts.	Turn- and speaker-aware detectors that aggregate adjacent turns into decontextualized claim units, condition on prosody, and operate over ordered, speaker-attributed windows (including streaming).
ER	Lexical retrievers over sentence spans in text corpora.	Built for clickable text spans, not continuous audio: single-turn queries miss callbacks and cross-episode references; ignore production metadata; ASR errors and code-switching corrupt indices and timing.	Return time-anchored evidence (episode ID + start/end seconds + transcript slice), indexed with dual lexical/phonetic views and re-ranked using speaker and production metadata, with memory for cross-episode recurrence.
VER	Single-span claim–passage NLI verifiers, lightweight verifiers, and early multimodal audio–text verifiers.	Assume single-span, clean text; struggle with multi-turn, time-qualified, and cross-episode evidence; sensitive to timing and diarization errors; FEVER-tuned models degrade on informal dialogue; large verifiers are costly for continuous/live streams.	Multi-span, time-aware reasoning with explicit who/when constraints; alignment-robust audio–text fusion; WER-aware training (oracle+ASR, noise/punctuation perturbations); cascaded verifiers for streaming, plus attribution and recoverability over timestamped bundles.
GEN	Text-only rationale generators and retrieval-augmented pipelines that generate explanations jointly with verdicts, often via LLM prompting.	Treat explanations as pure text: ignore delivery (sarcasm, emphasis, hesitation) and conversational path; lack timestamps and speaker labels; make mid-stream verdict changes hard to justify and leave the link to the underlying audio opaque.	Timestamped, speaker-attributed explanations that pair short audio clips with readable transcript spans and a compact assertion→challenge→repair timeline, using recoverability-based attribution and lightweight two-stage generators suitable for streaming.

Table 4: Summary of how traditional pipeline components behave on spoken dialogue and the corresponding design needs. See Section 4 for references.

streaming ASR and online diarization with evidence retrieval that returns time-anchored spans (episode ID plus start/end seconds) via dual lexical–acoustic indices and windowed retrieval; verify with lightweight models that can keep up in live settings. Richer reasoning and accountability follow: use alignment-robust (late/gated) fusion that masks diarization uncertainty so prosody and timing inform decisions; model the conversational fabric—roles, hedges, stance, and uptake/repairs—to assemble the claim unit across turns; track narratives across episodes with simple graph or archival tools; and surface listenable evidence (a transcript snippet plus a short audio clip with speaker attribution), validated by recoverability tests. Finally, make the system trustworthy and equitable: replace confidence-only gates with calibrated, WER-aware verification that is sensitive to error types (punctuation, proper nouns, code-switch points), and mitigate accent/dialect and noise brittleness through dialect-aware pretraining, domain adaptation, and perturbation-based checks.

6 Conclusion

Audio platforms reshape misinformation by combining the persuasive force of spoken delivery with the dynamics of conversation. Prosody, pacing, and emotion shape how claims are received, while dialogue order, repetition, and cross-turn dependencies sustain narratives across speakers and episodes. These dual properties make audio platforms fundamentally different from text or images, revealing why pipelines built for written claims fail when applied to spoken dialogue. We argue that fact-checking must be reframed around these realities, addressing challenges in real-time detection, multimodal fusion, conversational modeling, temporal tracking, and fairness.

Limitations

External validity is restricted. Most examples and analyses center on English, long-form talk (podcasts) with a two-speaker structure and reasonably clean production metadata. We do not test generalization to multi-party debates, call-in shows, highly code-switched speech, or low-metadata settings (sparse chapters and show notes), so portability to those regimes remains unverified.

Our pipeline assumptions depend on upstream components that we do not fully characterize. We rely on off-the-shelf ASR, VAD/segmentation, and diarization, but we do not provide a systematic error analysis by error type—e.g., punctuation, proper nouns, code-switch boundaries—or by speaker or acoustic condition, and we do not compare alternative ASR, VAD, or diarization backbones. As a result, the end-to-end robustness of the proposed design under diverse recording conditions remains uncertain.

Misinformation can also be propagated through

synthetic or cloned voices, and impersonation via voice cloning is an increasingly serious threat. In this work, however, we treat deepfake and voice-cloning detection as an upstream, separate problem of media integrity and speaker authenticity, and focus instead on the truthfulness of the claims made in spoken dialogue. In our setting, a voice-cloned clip that makes a factually correct statement should still be labeled as *True* at the verification stage, while a separate deepfake detector would be responsible for flagging the audio as synthetic. A full survey of audio deepfake generation and detection is therefore beyond our scope but represents an important complementary line of work to the spoken fact-checking pipeline we analyze here.

Finally, we acknowledge unmeasured risks. We do not quantify misattribution risk from diarization/alignment errors, possible demographic bias (accent, pitch range, speaking style) in prosody-aware features or ASR, or amplification harms from replaying salient misinformation clips. While we discuss safeguards (e.g., timestamped evidence and recoverability checks), the paper does not include bias audits or privacy analyses to demonstrate that these mitigations reduce harm in practice.

Acknowledgments

This work was supported in part by U.S. NSF awards #2114824 and #2438810. Some experimental results were obtained using computational resources provided by CloudBank, supported through U.S. NAIRR award #240336.

References

Yosra Abdessamed, Shadi Rezapour, and Steven Wilson. 2024. Identifying narrative content in podcast transcripts. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2631–2643, St. Julian's, Malta. Association for Computational Linguistics.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. 2019. Audio-visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Liesbeth Allein, Marlon Saelens, Ruben Cartuyvels, and Marie-Francine Moens. 2023. Implicit temporal reasoning for evidence-based fact-checking. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 176–189, Dubrovnik, Croatia. Association for Computational Linguistics.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Sankaranarayanan Ananthakrishnan and Shrikanth S. Narayanan. 2008. Automatic prosodic event detection using acoustic, lexical, and syntactic evidence. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):216–228.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Published as a conference paper at ICLR 2024.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Mohamed Bahaa, Mena Hany, and Ehab E Zakaria. 2024. Advancing automated deception detection: A multimodal approach to feature extraction and analysis. In *International Conference on Intelligent Systems, Blockchain, and Communication Technologies*, pages 727–738. Springer.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Claire Augusta Bergey and Simon DeDeo. 2024. From "um" to "yeah": Producing, predicting, and regulating information flow in human conversation. *Preprint*, arXiv:2403.08890.

Tirza Biron, Moshe Barboy, Eran Ben-Artzy, Alona Golubchik, Yanir Marmor, Assaf Marron, Smadar Szekely, Yaron Winter, and David Harel. 2025. Disentanglement of prosodic meaning: Toward a framework for the analysis of nonverbal information in speech. *Proceedings of the National Academy of Sciences*, 122(37):e2500510122.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2025. DEFAME: Dynamic Evidence-based FAct-checking with Multimodal Experts. In *Proceedings of the 42nd International Conference on Machine Learning*.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proceedings of Interspeech 2021*, pages 3111–3115, Brno, Czechia.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Kevin Matthe Caramancion. 2022. An Exploration of Mis/Disinformation in Audio Format Disseminated in Podcasts: Case Study of Spotify. In *2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pages 1–6, Toronto, ON, Canada. IEEE.

Gustavo Cardoso, Rita Sepúlveda, and Inês Narciso. 2022. WhatsApp and audio misinformation during the Covid-19 pandemic. *El Profesional de la información*, page e310321.

Eric Chamoun, Marzieh Saeidi, and Andreas Vlachos. 2023. Automated fact-checking in dialogue: Are specialized models needed? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16009–16020, Singapore. Association for Computational Linguistics.

Chaewan Chun, Lysandre Terrisse, Delvin Ce Zhang, and Dongwon Lee. 2025. Mad: A benchmark for multi-turn audio dialogue fact-checking. In *Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS 2025)*.

Chaewan Chun, Delvin Ce Zhang, and Dongwon Lee. 2026. Context-aware multimodal claim verification in spoken dialogues. *The Pennsylvania State University Technical Report*.

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.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Żelasko, and Miguel Jetté. 2021. Earnings-21: A practical benchmark for asr in the wild. In *Interspeech 2021*, pages 3465–3469.

Zechun Deng, Ziwei Liu, Ziqian Bi, Junhao Song, Chia Xin Liang, Joe Yeong, and Junfeng Hao. 2025. Achieving trustworthy real-time decision support systems with low-latency interpretable ai models. *Preprint*, arXiv:2506.20018.

Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. 2024. Document-level claim extraction and decontextualisation for fact-checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11943–11954, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *Preprint*, arXiv:1811.01241.

Jacob Eisenstein, Vinodkumar Prabhakaran, Clara Rivera, Dorottya Demszky, and Devyani Sharma. 2023. Md3: The multi-dialect dataset of dialogues. In *Proceedings of Interspeech 2023*, pages 4059–4063. ISCA.

Azza El-Masri, Martin J. Riedl, and Samuel Woolley. 2022. Audio misinformation on WhatsApp: A case study from Lebanon. *Harvard Kennedy School Misinformation Review*.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Alexander Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. ConvoSumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6866–6880, Online. Association for Computational Linguistics.

Yaxin Fan, Peifeng Li, and Qiaoming Zhu. 2025. Improving dialogue discourse parsing through discourse-aware utterance clarification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18800–18816, Vienna, Austria. Association for Computational Linguistics.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of

multimodal sarcasm detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Ujwal Gadiraju, Vinay Setty, and Stefan Buijsman. 2024. Claim Check-worthiness in Podcasts: Challenges and Opportunities for Human-AI Collaboration to Tackle Misinformation. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2024), Works-in-Progress*.

Daniel Galvez, Greg Diamos, Juan Torres, Keith Achorn, Juan Cerón, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed ELsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, and Heng Ji. 2022. NewsClaims: A new benchmark for claim detection from news with attribute knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6002–6018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Linda Geddes. 2022. Joe Rogan's Covid claims: what does the science actually say? Accessed: 2025-04-13.

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 '24, page 4487–4495, New York, NY, USA. Association for Computing Machinery.

Patrick Giedemann, Pius von Däniken, Jan Deriu, Alvaro Rodrigo, Anselmo Peñas, and Mark Cieliebak. 2025. Viclaim: A multilingual multilabel dataset for automatic claim detection in videos. *Preprint*, arXiv:2504.12882.

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.

Arushi Goel, Zhifeng Kong, Rafael Valle, and Bryan Catanzaro. 2024. Audio dialogues: Dialogues dataset for audio and music understanding. In *Proceedings of Synthetic Data's Transformative Role in Foundational Speech Models*, pages 61–65. ISCA.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Raymond Grossman, Taejin Park, Kunal Dhawan, Andrew Titus, Sophia Zhi, Yulia Shchadilova, Weiqing Wang, Jagadeesh Balam, and Boris Ginsburg. 2025. SPGISpeech 2.0: Transcribed multi-speaker financial audio for speaker-tagged transcription. In *Interspeech 2025*, pages 4048–4052.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Philip U. Gustafsson, Petri Laukka, and Torun Lindholm. 2023. Vocal characteristics of accuracy in eyewitness testimony. *Speech Communication*, 146:82–92.

Joshua J. Guyer, Pablo Briñol, Travis I. Vaughan-Johnston, Leandre R. Fabrigar, Leandro Moreno, and Richard E. Petty. 2021. Paralinguistic features communicated through voice can affect appraisals of confidence and evaluative judgments. *Journal of Nonverbal Behavior*, 45(4):479–504. Epub 2021 Jul 6.

Chen Han, Wenzhen Zheng, and Xijin Tang. 2025. Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. *Preprint*, arXiv:2505.18596.

Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. Claimbuster: the first-ever end-to-end fact-checking system. *Proc. VLDB Endow.*, 10(12):1945–1948.

Mads Henrichsen and Rasmus Krebs. 2025. Two-stage reasoning-infused learning: Improving classification with llm-generated reasoning. *Preprint*, arXiv:2507.00214.

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 transactions on audio, speech, and language processing*, 29:3451–3460.

Yi Huang, Xiaoting Wu, Si Chen, Wei Hu, Qing Zhu, Junlan Feng, Chao Deng, Zhijian Ou, and Jiangjiang

Zhao. 2022. CMCC: A comprehensive and large-scale human-human dataset for dialogue systems. In *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, pages 48–61, Abu Dhabi, Beijing (Hybrid). Association for Computational Linguistics.

Petar Ivanov, Ivan Koychev, Momchil Hardalov, and Preslav Nakov. 2024. Detecting check-worthy claims in political debates, speeches, and interviews using audio data. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12011–12015.

Mimansa Jaiswal, Sairam Tabibu, and Rajiv Bajpai. 2016. The Truth and Nothing But the Truth: Multimodal Analysis for Deception Detection. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 938–943, Barcelona, Spain. IEEE.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. 2018. ClaimRank: Detecting check-worthy claims in Arabic and English. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30, New Orleans, Louisiana. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context. *Preprint*, arXiv:2309.08105.

Johannes Kiesel, Damiano Spina, Henning Wachsmuth, and Benno Stein. 2021. The Meant, the Said, and the Understood: Conversational Argument Search and Cognitive Biases. In *CUI 2021 - 3rd Conference on Conversational User Interfaces*, pages 1–5, Bilbao (online) Spain. ACM.

Daniel Kopev, Ahmed Ali, Ivan Koychev, and Preslav Nakov. 2019. Detecting deception in political debates using acoustic and textual features. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 652–659. IEEE.

Alkis Koudounas, Moreno La Quatra, and Elena Baralis. 2025. Deepdialogue: A multi-turn emotionally-rich spoken dialogue dataset. *Preprint*, arXiv:2505.19978.

Korbinian Kuhn, Verena Kersken, and Gottfried Zimmermann. 2025. Evaluating asr confidence scores for automated error detection in user-assisted correction interfaces. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zhu Li, Xiyuan Gao, Yuqing Zhang, Shekhar Nayak, and Matt Coler. 2025a. Evaluating multimodal large language models on spoken sarcasm understanding. *Preprint*, arXiv:2509.15476.

Zhu Li, Yuqing Zhang, Xiyuan Gao, Shekhar Nayak, and Matt Coler. 2025b. Leveraging Large Language Models for Sarcastic Speech Annotation in Sarcasm Detection. In *Interspeech 2025*, pages 3973–3977.

Benjamin Litterer, David Jurgens, and Dallas Card. 2024. Mapping the podcast ecosystem with the structured podcast research corpus. *Preprint*, arXiv:2411.07892.

Moyang Liu, Yukun Liu, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Xuefei Liu, and Guanjun Li. 2024. Exploring the role of audio in multimodal misinformation detection. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 204–208. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Z. Liu, Y. Xiao, Z. Su, and 1 others. 2025. Bilingual dialogue dataset with personality and emotion annotations for personality recognition in education. *Scientific Data*, 12(514).

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Jason Lucas, Adaku Uchendu, Michiharu Yamashita, Jooyoung Lee, Shaurya Rohatgi, and Dongwon Lee. 2023. Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14279–14305, Singapore. Association for Computational Linguistics.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. 2024. EX-FEVER: A dataset for multi-hop explainable fact verification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9340–9353, Bangkok, Thailand. Association for Computational Linguistics.

Johannes B. Mahr and Gergely Csibra. 2021. The effect of source claims on statement believability and speaker accountability. *Memory & Cognition*, 49(8):1505–1525.

Alexandre Maros, Jussara M. Almeida, and Marisa Vasconcelos. 2021. A study of misinformation in audio messages shared in whatsapp groups. In *Disinformation in Open Online Media: Third Multidisciplinary International Symposium, MISDOOM 2021, Virtual Event, September 21–22, 2021, Proceedings*, page 85–100, Berlin, Heidelberg. Springer-Verlag.

Tomoko Matsui, Tagiru Nakamura, Akira Utsumi, Akihiro T. Sasaki, Takahiko Koike, Yumiko Yoshida, Tokiko Harada, Hiroki C. Tanabe, and Norihiro Sadato. 2016. The role of prosody and context in sarcasm comprehension: Behavioral and fmri evidence. *Neuropsychologia*, 87:74–84.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. 2021. SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv e-prints*, arXiv:2104.02014.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Rrubaa Panchendrarajan and Arkaitz Zubiaga. 2024. Claim detection for automated fact-checking: A survey on monolingual, multilingual and cross-lingual research. *Natural Language Processing Journal*, 7:100066.

Se Park, Chae Kim, Hyeongseop Rha, Minsu Kim, Joanna Hong, Jeonghun Yeo, and Yong Ro. 2024. Let's go real talk: Spoken dialogue model for face-to-face conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16348, Bangkok, Thailand. Association for Computational Linguistics.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A review of speaker diarization: Recent advances with deep learning. *Comput. Speech Lang.*, 72(C).

Alex Paterson. 2021. Spotify's Joe Rogan falsely says children don't need to get the coronavirus vaccine. Accessed: 2025-04-13.

Sachin Pathiyan Cherumanal, Ujwal Gadiraju, and Damiano Spina. 2024. Everything we hear: Towards tackling misinformation in podcasts. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 596–601.

Ron Korenblum Pick, Vladyslav Kozhukhov, Dan Vilenchik, and Oren Tsur. 2022. Stem: Unsupervised structural embedding for stance detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11174–11182.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Ayesha Qamar, Adarsh Pyarelal, and Ruihong Huang. 2023. Who is speaking? speaker-aware multiparty dialogue act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10122–10135, Singapore. Association for Computational Linguistics.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. 2021. Don't be contradicted with anything! CI-ToD: Towards benchmarking consistency for task-oriented dialogue system. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2367, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*, ICML'23. JMLR.org.

Rajiv Rao, Ting Ye, and Brianna Butera. 2022. The prosodic expression of sarcasm vs. sincerity by heritage speakers of spanish. *Languages*, 7(1).

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197.

Gustavo Resende, Philipe Melo, Hugo Sousa, Johnnatan Messias, Marisa Vasconcelos, Jussara Almeida, and Fabrício Benevenuto. 2019. (mis)information dissemination in whatsapp: Gathering, analyzing and

countermeasures. In *The World Wide Web Conference*, WWW '19, page 818–828, New York, NY, USA. Association for Computing Machinery.

Naquee Rizwan, Nayandeep Deb, Sarthak Roy, Vishwajeet Singh Solanki, Kiran Garimella, and Animesh Mukherjee. 2025. Toxicity begets toxicity: Unraveling conversational chains in political podcasts. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 11776–11784, New York, NY, USA. Association for Computing Machinery.

Maarten Van Segbroeck, Ahmed Zaid, Ksenia Kutsenko, Cirenia Huerta, Tinh Nguyen, Xuewen Luo, Björn Hoffmeister, Jan Trmal, Maurizio Omologo, and Roland Maas. 2020. Dipco — dinner party corpus. In *Interspeech 2020*, pages 434–436.

Vinay Setty and Adam James Becker. 2025. Annotation tool and dataset for fact-checking podcasts. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW Companion '25, page 789–792.

Md Hasan Shahriar, Md Mohaimin Al Barat, Harshavardhan Sundar, Ning Zhang, Naren Ramakrishnan, Y. Thomas Hou, and Wenjing Lou. 2025. Temporal misalignment attacks against multimodal perception in autonomous driving. *Preprint*, arXiv:2507.09095.

Ori Shapira, Shlomo Chazan, and Amir David Nissan Cohen. 2025. Measuring the effect of transcription noise on downstream language understanding tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29978–30004, Vienna, Austria. Association for Computational Linguistics.

Zaid Sheikh, Shuichiro Shimizu, Siddhant Arora, Jiatong Shi, Samuele Cornell, Xinjian Li, and Shinji Watanabe. 2025. Scalable spontaneous speech dataset (sssd): Crowdsourcing data collection to promote dialogue research. In *Proceedings of Interspeech 2025*, pages 3963–3967, Rotterdam, The Netherlands.

Jiasheng Si, Yingjie Zhu, and Deyu Zhou. 2023. Exploring faithful rationale for multi-hop fact verification via salience-aware graph learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Karan Singla, Shahab Jalalvand, Yeon-Jun Kim, Ryan Price, Daniel Pressel, and Srinivas Bangalore. 2022. Seq-2-seq based refinement of asr output for spoken name capture. In *Proceedings of Interspeech 2022*, pages 3963–3967, Incheon, Korea.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Zhaokai Sun, Li Zhang, Qing Wang, Pan Zhou, and Lei Xie. 2025. Towards Robust Overlapping Speech Detection: A Speaker-Aware Progressive Approach Using WavLM. In *Interspeech 2025*, pages 1653–1657.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. Empowering the fact-checkers! automatic identification of claim spans on Twitter. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7701–7715, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Reality Check Team. 2022. Joe Rogan: Four claims from his Spotify podcast fact-checked.

Flora Teoh. 2021. Joe Rogan interview with Peter McCullough contains multiple false and unsubstantiated claims about the COVID-19 pandemic and vaccines. Accessed: 2025-04-13.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Nafis Tripto, Adaku Uchendu, Thai Le, Mattia Setzu, Fosca Giannotti, and Dongwon Lee. 2023. HANSEN: Human and AI spoken text benchmark for authorship analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13706–13724, Singapore. Association for Computational Linguistics.

V Venktesh and Vinay Setty. 2025. Livefc: A system for live fact-checking of audio streams. In *18th ACM International Conference on Web Search and Data Mining, WSDM 2025*, pages 1060–1063. Association for Computing Machinery (ACM).

M. Olguta Vilceanu. 2025. Parasocial intimacy, change, and nostalgia in podcast listener reviews. *Media and Communication*, 13(0).

Jonas Wallat, Maria Heuss, Maarten de Rijke, and Avishek Anand. 2025. Correctness is not faithfulness in retrieval augmented generation attributions.

In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 22–32, New York, NY, USA. Association for Computing Machinery.

David Wan, Chris Kedzie, Faisal Ladhak, Elsbeth Turcan, Petra Galuscakova, Elena Zotkina, Zhengping Jiang, Peter Bell, and Kathleen McKeown. 2021. Segmenting subtitles for correcting ASR segmentation errors. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2842–2854, Online. Association for Computational Linguistics.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. 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 993–1003, Online. Association for Computational Linguistics.

Shinji Watanabe, Michael Mandel, Jon Barker, Emmanuel Vincent, Ashish Arora, Xuankai Chang, Sanjeev Khudanpur, Vimal Manohar, Daniel Povey, Desh Raj, David Snyder, Aswin Shanmugam Subramanian, Jan Trmal, Bar Ben Yair, Christoph Boeddeker, Zhaoheng Ni, Yusuke Fujita, Shota Horiguchi, Naoyuki Kanda, and 2 others. 2020. Chime-6 challenge:tackling multispeaker speech recognition for unsegmented recordings. *Preprint*, arXiv:2004.09249.

Valerie Wirtschafter. 2023. Audible reckoning: How top political podcasters spread unsubstantiated and false claims. Brookings.

Cihan Xiao, Ruixing Liang, Xiangyu Zhang, Mehmet Emre Tiryaki, Veronica Bae, Lavanya Shankar, Rong Yang, Ethan Poon, Emmanuel Dupoux, Sanjeev Khudanpur, and Leibny Paola Garcia Perera. 2025. Casper: A large scale spontaneous speech dataset. *Preprint*, arXiv:2506.00267.

Rui Xing, Timothy Baldwin, and Jey Han Lau. 2025. Evaluating evidence attribution in generated fact checking explanations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5475–5496, Albuquerque, New Mexico. Association for Computational Linguistics.

Weimin Xiong, Yifan Song, Peiyi Wang, and Sujian Li. 2023. Rationale-enhanced language models are better continual relation learners. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15489–15497, Singapore. Association for Computational Linguistics.

Longqi Yang, Yu Wang, Drew Dunne, Michael Sobolev, Mor Naaman, and Deborah Estrin. 2019. More Than Just Words: Modeling Non-Textual Characteristics of Podcasts. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 276–284, Melbourne VIC Australia. ACM.

Maya Yang. 2022. 'Menace to public health': 270 experts criticise Spotify over Joe Rogan's podcast. Accessed: 2025-04-23.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

Karl T. Maloney Yorganci and Leslie McMurtry. 2024. "one of us": Examining the authenticity and parasocial relationships of stand-up comedian podcast hosts. *Journal of Radio & Audio Media*, 0(0):1–21.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. 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.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Kalina Bontcheva, and Peter Tolmie. 2015. Towards detecting rumours in social media. *Preprint*, arXiv:1504.04712.