

From "Which podcast production steps can AI do well?"

Where AI slips, step by step

Working links, fluent summaries, good signal and word scores, clean AI reviews: reassuring signs can hide failures.

01 RESEARCH

More searching did not make the facts more accurate.

Onweller and colleagues benchmarked 14 models. Even the strongest frontier models had link validity above 94%, but only 39-77% factual accuracy. As tool calls rose from 2 to 150, fact-check accuracy fell approximately 42% on average across two frontier models.

R1 · Cited but Not Verified · May 2026

A · Preprint; abstract only · Read 2026-10-14

→ Ask for each claim's supporting passage, not a link-check score or search count.

02 SUMMARY DRAFTING

The summary left out the main event.

In Spring 2026, Sim, Eiger and Kohno had two AI assistants summarize two uploaded 25-second animated accident videos. Across 20 summaries, 51.6% of central details were omitted on average; 95% omitted the car hitting the pedestrian.

R3 · AI-Enabled Human Memory Manipulation · Published Sep 2026

A · Small primary audit · Read 2026-10-14

→ Name the recording's main event before reading the summary. Check that it survived.

03 AUDIO CLEANING

A better signal score hid more word errors.

In 824 noisy English clips, Islam, Nahar and Hamid found processing raised average signal quality from 32.28 to 35.99 dB, yet one recognizer's average word error rate rose from 10.53% to 21.66%. The signal measure is peak signal-to-noise ratio, not listener-rated clarity. One cleaning tool.

R2 · When Audio Separation Hurts Zero-Shot ASR · Mar 2026, revised Jul 2026

A · Empirical preprint · Read 2026-10-14

→ Check raw and processed transcripts against the audio before choosing a file.

Five different tests, none on podcasts. The checks are ours, built from each study; none has a measured catch rate. Platform actions follow documented controls.

Overleaf: synthetic voice, AI review pass and Spotify correction controls. · 1 of 2

04 SYNTHETIC VOICE

The best word score conveyed key facts worst.

In a test published October 2025, Tee and colleagues had listeners answer 2,688 key-fact questions on English NPR news read by four open research voice models from 2024 and 2025. The authors single out names, numbers and abbreviations as problem areas.

The test turns a year old on 30 October 2026. We keep it as the only test we found that sets word scores against what listeners took in: a measuring gap, not a rate for any tool you use.

AUDIO	WORD ERROR RATE	LISTENER ACCURACY
Fewest word errors (synthetic)	5.7%	81.2%
Most facts conveyed (synthetic)	9.0%	90.4%
Study's cut of original NPR audio	8.1%	92.0%

Rounded to one decimal.

R7 · SP-MCQA · Oct 2025
A · Preprint; open research models · Read 2026-10-14

→ Ask about a name, number or abbreviation in the audio. Check the answer against the text.

05 AI REVIEW PASS

Different reviewers found different errors.

Nguyen and colleagues tested six models on research papers containing known errors. The best configuration caught 71.6%; combining all six models' detections caught 83.3%. The authors suggest different models catch different errors. This counts known errors found, not how often flags are right.

R5 · Benchmarking Agentic Review Systems · Jun 2026
A · Preprint; abstract only · Read 2026-10-14

→ Have a second model review the draft too. A clean AI review is not clearance: a person still does the final listen.

Spotify corrections: eligible shows, web only

Transcripts

Download, correct, upload and save. Enabling auto-generated transcripts replaces your uploaded file.

Chapter hooks

Short AI rewordings of chapter titles for listeners discovering shows. Different listeners may see different types. Cannot be edited; switch off for an episode or the whole show.

→ Keep auto-generated transcripts off after uploading a fix. Switch off unwanted hooks.

A · Spotify official help, undated; read 2026-10-14. Documented controls, not account-tested. This is not legal advice.