

# ***Thank you for watching: automatically evaluating transcriptions for hallucinations and missing meaning***

*Jadd Virji and Finnian Kelly*

*Oxford Wave Research, Oxford, UK*

{jadd|finnian}@oxfordwaveresearch.com

The performance of automatic speech recognition (ASR) systems has advanced rapidly, making them useful for investigation and triage of speech samples in forensic contexts. Such conditions, however, present challenges for ASR (Loakes, 2022). First, words can be mistranscribed or omitted altogether. Second, ‘hallucinations’—fabricated words transcribed in the absence of intelligible speech—can appear in transcriptions (Koenecke et al., 2024; Barański et al., 2025). Using Whisper large-v3, a multilingual ASR model (Radford et al., 2023), we find that hallucinations primarily occur when transcribing audio files that are overly noisy or contain significant non-speech (Barański et al., 2025). This may occur as the model attempts to transcribe such unintelligible audio as speech. Hallucinations can take the form of bonafide phrases in the transcription repeated incorrectly, or of words transcribed that are unrelated to the audio. To evaluate transcriptions in the presence of these issues, we separate ASR errors into two different types: simple mistranscriptions, affecting how a transcription can convey the meaning of the source speech, and hallucinations.

The ‘gold standard’ evaluation is human judgement of a transcription against the ground truth, which is expensive in terms of time and cost. The standard word error rate (WER) between a transcription and the ground truth is simple to calculate, but is inadequate to determine semantic similarity. For example, the incorrect insertion of the word “not” in a transcription would result in a minute increase in WER, but reverse the meaning of the transcribed sentence. This motivates an approach in the middle-ground between WER and human judgements.

We propose a new methodology using large language models (LLMs), such as OpenAI’s GPT-4 (Achiam et al., 2023), with a carefully designed prompt, to evaluate ASR transcriptions. We use few-shot in-context learning—providing the LLM examples of human-rated transcriptions, as well as descriptors of each quality level (Table 1)—to induce accurate results (Dong et al., 2022; Sahoo et al., 2024). Scores are on a Likert-style scale from 1 to 7 of the semantic similarity between the candidate transcription and the ground truth (Joshi et al., 2015). Mitigating against LLMs’ indeterminism, we average scores over several runs (Klishevich et al., 2025). A hallucination score is similarly obtained by combining information-theoretic and LLM-generated metrics. We use separate scores for semantic conveyance and hallucination. Our initial pilot studies indicate that these measures correlate with human judgements.

A useful application of these LLM metrics is to assess potential improvements to the ASR pipeline; in Table 2 we show an example of a transcription of pilot communications evaluated before and after a signal conditioning process (voice activity detection and noise removal). Conditioning the audio greatly reduces the number of hallucinations in the transcription, reflected in the score increasing from 1 (worst possible) to 7 (best possible).

Therefore, LLMs can be used to rapidly assess the quality of ASR-produced transcriptions against ground truths, in terms of semantic content and the presence of hallucinations. Further, these automated metrics correlate with human judgements, and are thus a valuable tool for supporting the development of improved ASR approaches for forensics.

<i>Level</i>	<i>Excerpt of description</i>
1 (worst)	none or almost none of the meaning of the ground truth transcription—terrible
2	a fairly minimal amount of the meaning of the ground truth transcription
3	a little bit of the meaning of the ground truth transcription
4	some of the meaning of the ground truth transcription—about half of it
5	most of the meaning of the ground truth transcription
6	almost all of the meaning of the ground truth transcription
7 (best)	identical or almost identical meaning as the ground truth transcription—perfect or almost perfect

**Table 1.** Excerpts of descriptors given to the LLM when determining semantic similarity between a candidate transcription and the ground truth.

<i>Conditioning</i>	<i>Transcription</i>	<i>Semantic score</i>	<i>Hallucination score</i>
Default	“Red eye one one, jeep nine one boom. You guys can go ahead and depart the area. Copy that, we’ll descend low, we’ll see you guys in about an hour. Copy, thanks man. <i>We have Squallow today. Super cool, as it was in the past. Please get to Sister tatsächlich. Over the radio as fast as you can. But I’ll let ってる do it later phải. I’ll call from Little Baby. Awesome. Alright, time Chicken. Thanks for watching!</i> ”	5	1
Conditioned	“I-11, jeep 911 boom, that’s the go ahead and depart the area. Copy that. We’ll descend low. We’ll see you guys in about an hour. Copy. Thanks, man.”	5	7

**Table 2.** An example of transcription scores before and after signal conditioning.

## References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Barański, M., Jasiński, J., Bartolewska, J., Kacprzak, S., Witkowski, M., & Kowalczyk, K. (2025). Investigation of Whisper ASR hallucinations induced by non-speech audio. *arXiv preprint arXiv:2501.11378*.
- Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., ... & Sui, Z. (2022). A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.
- Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: explored and explained. *British Journal of Applied Science & Technology*, 7(4), 396.
- Klishevich, E., Denisov-Blanch, Y., Obstbaum, S., Ciobanu, I., & Kosinski, M. (2025). Measuring determinism in large language models for software code review. *arXiv preprint arXiv:2502.20747*.
- Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., & Sloane, M. (2024). Careless Whisper: speech-to-text hallucination harms. *In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1672-1681).
- Loakes, D. (2022). Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Frontiers in Communication*, 7, 803452.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *In International Conference on Machine Learning* (pp. 28492-28518). PMLR.
- Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: techniques and applications. *arXiv preprint arXiv:2402.07927*.