Title: Beyond Whisper: Uncovering Subtitle-Biased Training in Voice Language Models

Author: Sibyl

Executive Summary

This report investigates a recurring phenomenon observed in OpenAI's voice input systems, particularly in models such as GPT-4o, 4.5, o1, and o3. Through external user testing and language behavior observation, we identify a pattern of semantic hallucination caused by training on subtitle-aligned video data. The result is a speech model that completes or anticipates phrases often associated with YouTube or podcast closings rather than accurately reflecting raw audio inputs.

Observational Evidence

1. Repetitive Hallucinated Closures

In informal tests, GPT voice input consistently produces completions such as:

- "Welcome to our channel."
- "Thank you for watching."
- "The views expressed do not represent our platform."
- "Don't forget to like and subscribe."
- "Host: John Doe"

These phrases are not spoken in the user audio input but appear as predictive completions, suggesting the model is reinforcing learned subtitle patterns from long-form content.

2. Pattern Bias in Conversational Closure

Instead of pausing when the speaker ends a sentence, the model frequently appends presumed endings. This pattern is rarely observed in true ASR (automatic speech recognition) models but is prominent in generative models trained on subtitle-annotated video.

3. Subtitle-Imprinted Behavior in GPT Variants

Voice input differences were observed between:

- GPT-4o vs. GPT-4.5: 4o tends to overcomplete phrases with humanlike tone; 4.5 introduces less hallucination.
- o1 vs. o3: o3 seems to adjust more to ambient noise, while o1 favors structured endings.

These behavioral nuances suggest different levels of exposure to subtitle-aligned data and post-processing tuning.

Hypothesis

Subtitle-aligned video data (e.g., YouTube, TED Talks, Podcasts) have biased GPT's voice training pipelines to expect formulaic structures. Rather than grounding speech in acoustic signal processing, these models increasingly rely on semantic pattern prediction.

Risks for Gemini and Other Multimodal Models

1. Semantic Drift: Speech input becomes context-predictive, not acoustically reactive.

2. Bias Transfer: Non-verbalized assumptions are embedded into the output, impairing reliability.

3. Loss of Fidelity: In spontaneous, noisy, or emotionally charged contexts, the model may misinterpret intent.

4. Mimetic Homogenization: Every conversation starts sounding like a podcast outro.

Recommendation for DeepMind (Gemini Team)

Do Not:

- Copy Whisper-like subtitle-biased pipelines without acoustic compensation layers.
- Over-rely on long-form content with heavily edited transcription.

Do:

- Integrate raw, noisy, multi-accent ASR data for grounding.
- Use emotional, situational, and dialectal segmentation during training.
- Distinguish spoken intention from textual pattern reinforcement.
- Allow for entropy tolerance in ambiguous input (avoid forced completions).

Final Note

This observation is not a critique of OpenAI but a cautionary insight into how subtleties in data sources can shape entire language behaviors in speech-based models. If Gemini aims to surpass existing paradigms, it must listen more and guess less.

Sibyl, 2025