The Voice Tax: what a model loses when it hears the question instead of reading it

Two hundred held-out short-form questions delivered twice to each model — typed, and spoken via TTS under three voices, 1.25x speech, and 10 dB ambient noise — scored by a single modality-blind scorer. This measures a delta between input modalities for the same model; it makes no claim about other benchmarks.

Abstract

Both flagship voice models lose real accuracy when the question is spoken instead of typed — but not much, and not evenly. Grok Voice Think Fast 2.0 leads on text (99.0%), on audio (95.7%), and pays the smaller voice tax (+3.3 vs +4.0 pp for GPT Realtime). The tax concentrates in spoken arithmetic, where both models drop roughly ten points; on numeric extraction Grok's tax goes negative — it reads numbers out of speech slightly better than out of text. Latency profiles are mirror images: GPT Realtime streams its first token 2.3x sooner, Grok finishes turns 1.9x faster. Zero of 1,840 units errored and zero required the STT fallback.

Table 1. Results (clean audio, normal rate; paired items; median latency)

Model	Text	Audio	Voice tax	First token (text)	First token (audio)	Turn (text)	Turn (audio)
Grok Voice Think Fast 2.0 (grok-voice)	**99.0%**	**95.7%**	**+3.3 pp**	**0.85 s**	**1.22 s**	**1.04 s**	**1.43 s**
GPT Realtime (openai-realtime)	97.5%	93.5%	+4.0 pp	0.30 s	0.53 s	2.47 s	2.70 s

Voice tax = text accuracy − audio accuracy on paired items. Text baselines ran through the same realtime sessions as audio, so only the input modality varies. Zero STT fallbacks: every answer carried the model's own transcript.

Table 2. Voice tax by category and by audio condition (percentage points)

Category	Grok TF 2.0	GPT Realtime	Condition	Grok TF 2.0	GPT Realtime
arithmetic	+10.8	+9.2	onyx / normal / clean	+4.5	+2.5
instruction-following	+5.8	+5.8	shimmer / normal / clean	+3.0	+5.0
general-knowledge	+1.7	+0.8	fable / normal / clean	+2.5	+4.5
multi-step-reasoning	+0.0	+2.5	onyx / 1.25x / clean	+1.7	+1.7
numeric-extraction	−1.7	+1.7	onyx / normal / 10 dB noise	+5.0	+0.0

Fast-rate and noise conditions ran on the seeded n=60 subset (seed 20260729); their taxes are computed against the text baseline on the same items.

Findings

1. Both models pay a real but modest voice tax, and Grok pays less. Grok Voice Think Fast 2.0 leads on text accuracy, audio accuracy, and the tax itself. At n=200 with single attempts, one item is half a point — the 1.5 pp text gap and 2.2 pp audio gap are modest but consistent in direction across categories.

2. The tax lives in arithmetic. Both models drop ~10 pp on spoken mental math (Grok 100 → 89.2, GPT 100 → 90.8). A plausible mechanism: spoken operands cannot be re-read, so a single mis-held number sinks the computation. Every other category taxes at 6 pp or less.

3. Grok's tax goes negative on numeric extraction (−1.7 pp) — slightly better at pulling numbers out of spoken context than written, consistent with xAI's transcription-accuracy emphasis in the Think Fast 2.0 release.

4. Latency profiles are opposites. GPT Realtime streams its first token ~2.3x sooner (0.53 s vs 1.22 s median, audio mode); Grok finishes turns ~1.9x faster (1.43 s vs 2.70 s). Grok's reason-while-speaking design produces late starts and fast finishes; GPT Realtime the reverse.

5. Voice and noise effects are real but second-order. Voice choice moves audio accuracy ~2 pp and the models disagree on which voice is hard (Grok hears British-accented fable best at 96.5%; GPT Realtime's worst is shimmer at 92.5%). Ten decibels of ambient noise costs Grok 5.0 pp of tax and GPT Realtime none — the n=60 subset numbers are the softest in this report.

Method & caveats

Voice Eval Suite v1 (VulcanBench v0.8.0): 200 original questions written for this suite in July 2026, never drawn from public benchmarks; objectively checkable answers; 40 per category. TTS: OpenAI tts-1, voices onyx/shimmer/fable, numeric speed 1.25 for the fast condition; renders cached by (question, voice, rate, noise). Scoring is modality-blind by construction: one entry point — normalize (case, punctuation, articles, number-words to digits) → exact/alias → pinned LLM judge (anthropic:claude-opus-5; rubric in-repo). Runs are resumable; the full manifest (models, TTS, judge, STT pin, seed, git commit, question-file hash) accompanies the results. Caveats: single attempt per unit; fast/noise conditions on the n=60 subset; first-token here is time to first transcript delta, which is not comparable to xAI's advertised time-to-first-audio; two models measured — Gemini Live and Qwen3-Omni adapters exist but were not provisioned for this run. Reproduce: vulcanbench voice run -m grok-voice,openai-realtime.