Google

Gemini 3.8 Audio (Live, Live Extended Thinking) Model evaluation

Approach, methodology & results

Approach

We evaluated Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking using the methodology below:

Methodology

All Gemini scores are pass @1 except where otherwise noted. "Single attempt" settings allow no majority voting or parallel test-time compute. ServiceNow's EVA-Bench results were run by ServiceNow with the Gemini Enterprise Agent Platform (thinking minimal and high). Artificial Analysis (run by AA) and τ-Bench results were run with the Gemini API for the model-ids gemini-3.8-live-preview and gemini-3.8-live-extended-thinking with thinking high and default sampling settings.

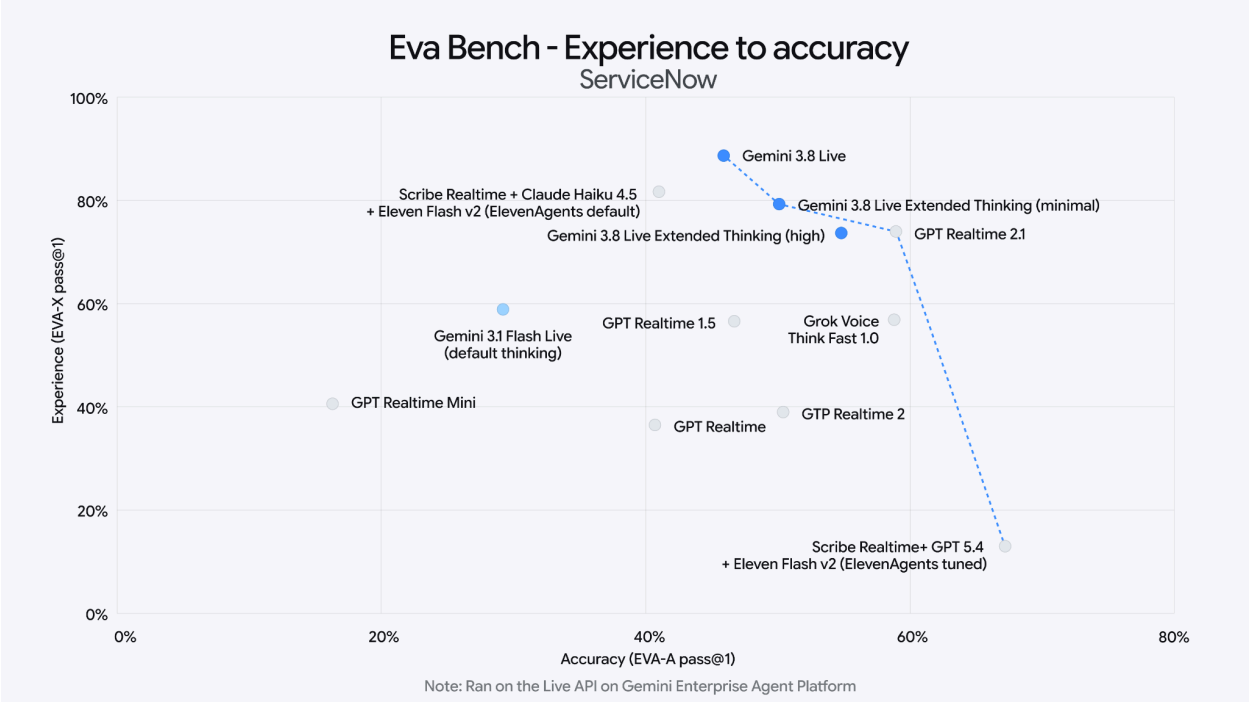
Capabilities / Benchmarks

- **ServiceNow Eva Bench:** This benchmark is an evaluation framework that measures voice agents over complete, multi-turn spoken conversations using a realistic bot-to-bot architecture.
- **Artificial Analysis Evaluations:** This benchmark measures speech to speech models across different characteristics including their reasoning quality, conversational dynamics, generation time and price.
- **Sierra τ^3-Bench:** This benchmark (τ^3-Banking) tests whether agents can navigate a large unstructured knowledge base and execute multi-step tool calls to resolve realistic banking workflows.

Results

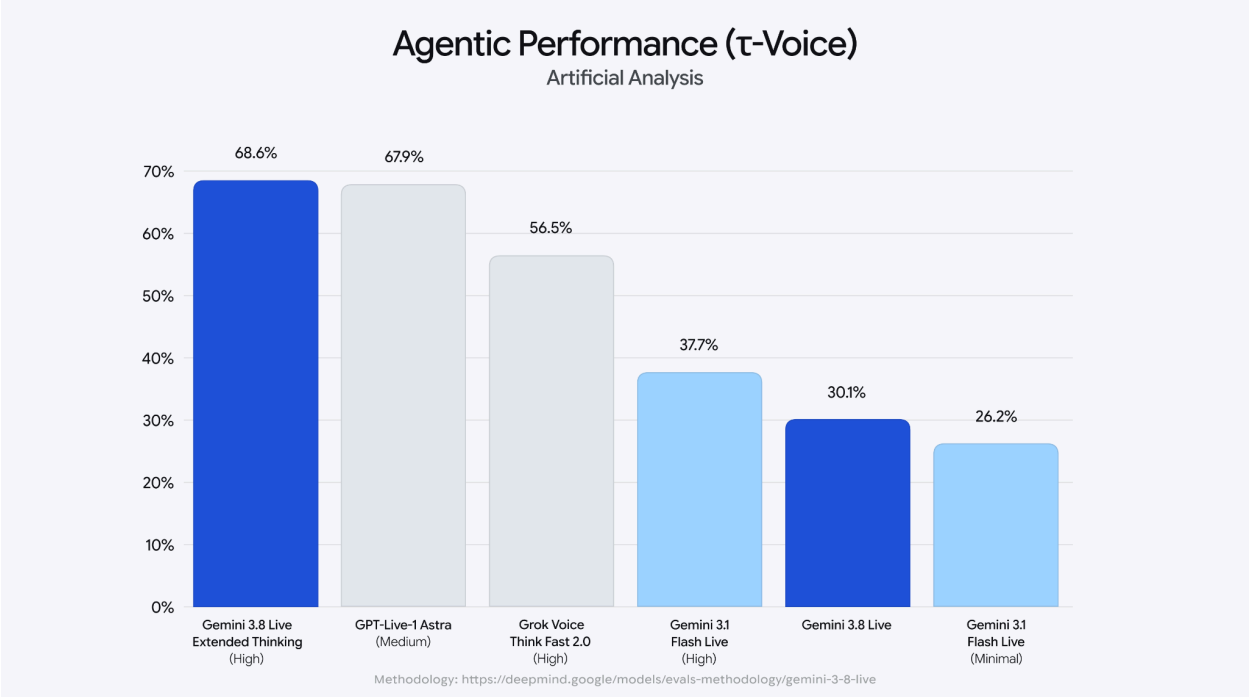
Gemini 3.8 Live and Gemini 3.8 Live Extended Thinking results as of September 2026 are below:

ServiceNow Evals



Note: Ran on the Live API on Gemini Enterprise Agent Platform

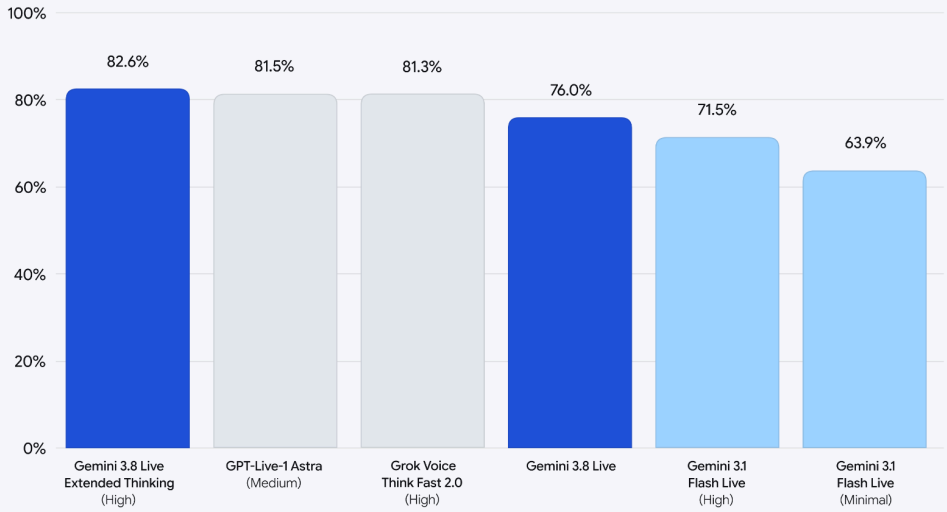
Artificial Analysis Evals



Methodology: https://deepmind.google/models/evals-methodology/gemini-3-8-live

Speech to Speech Index

Artificial Analysis



Methodology: https://deepmind.google/models/evals-methodology/gemini-3-8-live

Cost per Hour of Input Audio (Big Bench Audio subset)

Artificial Analysis

Lower is better

Methodology: https://deepmind.google/models/evals-methodology/gemini-3-8-live

Sierra Evals