vapi

The Humanness Index™: Human Perception of Voice AI, Crowdsourced at Scale

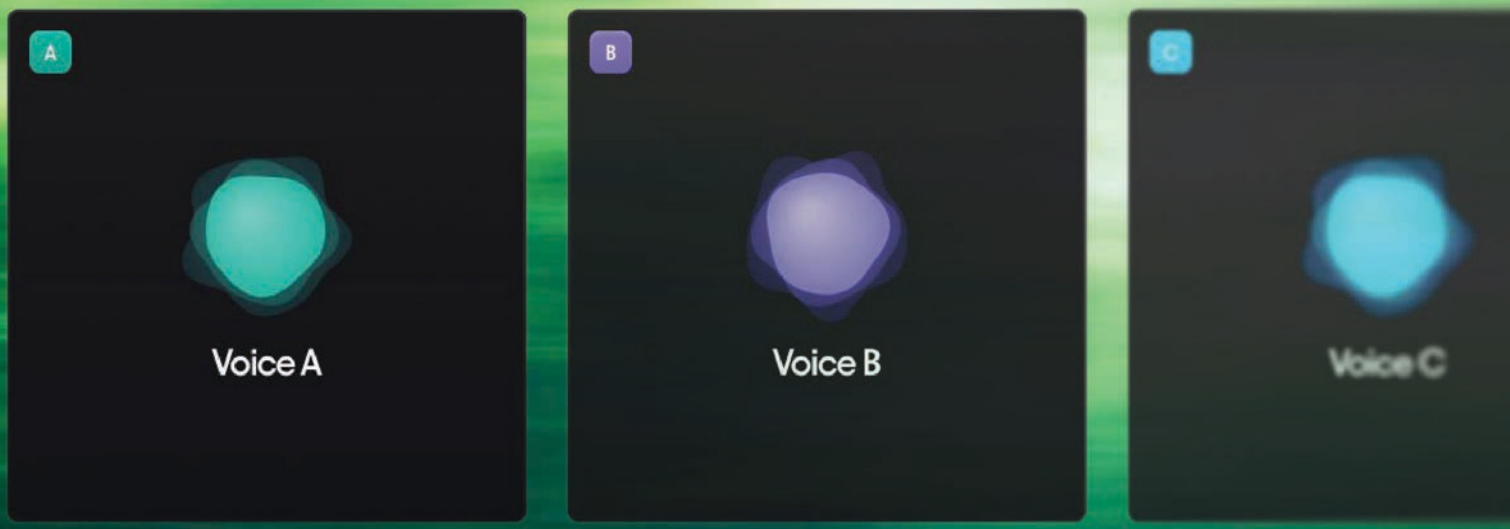


Table of Contents

Executive Summary

The rapid maturation of neural text-to-speech (TTS) synthesis has produced voice models of remarkable technical quality. Yet the field still lacks a standardized, scalable benchmark for the question that matters most to end users: how human does the voice sound?

Established benchmarks for accuracy and latency address two of the three dimensions that define a state-of-the-art voice model. The third, which we term "humanness," lacks a reliable, continuously updated measurement framework. We believe it is the most consequential of the three.

This paper introduces the Humanness Index™, a live, crowdsourced leaderboard developed by Vapi that scores leading TTS models on perceived human-likeness through large-scale pairwise human evaluation. Drawing on Vapi's community of over one million developers, the Index produces continuously updated rankings across major voice model providers and gives the industry a democratized answer to the most consequential question in voice AI: does this sound like a person?

The Humanness Index™ is publicly accessible at http://humannessindex.vapi.ai/. We invite developers, researchers, and voice AI practitioners to contribute evaluations, explore the rankings, and help define what it means for a voice to sound human.

Why Vapi is the right organization to publish the benchmark

The Humanness Index™ is published by Vapi, a model-agnostic voice AI platform that lets developers choose from the full spectrum of leading TTS providers. This architecture creates a distinctive evaluative position: Vapi observes model performance in production, at scale, across a diverse range of use cases, with no stake in the success of any particular provider. Developers on Vapi can plug in whichever model best fits their use case and switch as the landscape changes.

Provider-published benchmarks are necessarily subject to selection bias in sample choice and evaluation methodology. The Humanness Index™, by contrast, is constructed and maintained by a platform whose business depends on accurately matching developers to the best available model for their specific use case. Vapi's incentive is not to promote any voice provider, but to give developers the information they need to make better decisions. The more than one billion real-world calls processed across Vapi's platform provide a production signal that no individual provider has about its own models relative to competitors.

Equally important is the breadth of Vapi's evaluator community. The platform serves a heterogeneous population: software engineers, product managers, enterprise AI teams, and operations teams across healthcare, financial services, customer experience, and more. This community combines the technical sophistication of AI practitioners with the experiential grounding of the end users who ultimately decide whether a voice AI interaction feels human. The Humanness Index™ draws on this full range of perspectives, producing a benchmark that is both expert-informed and calibrated against real-world user expectations.

1 Billion	99.9%	2.5M+	750K+
calls supported	uptime for enterprise clients	agents launched	developers

The evolution of voice models

Text-to-speech has evolved through three broad paradigms. For decades, synthesis relied on phoneme-centric pipelines: hand-engineered front ends converted text into phonemes, durations, and prosody targets, while concatenative or Hidden Markov Model based back ends rendered the result into intelligible but often robotic speec

The neural era began along the deep learning revolution, when encoder-decoder models such as Tacotron learned to map text directly to spectrograms, replacing much of the engineered alignment machinery with learned attention and producing dramatically more natural audio.

Today's systems go further: neural codecs compress speech into discrete audio tokens, allowing transformers to treat synthesis more like language modeling. This shift has enabled large-scale training, controllable generation, and zero-shot voice cloning. In parallel, diffusion and flow-matching models have emerged as a second major technique along the GenAI revolution, generating utterances holistically rather than token by token and achieving exceptional fidelity — with many production systems now combining the two, using an autoregressive model to decide what and how to say and a diffusion-based decoder to render the acoustic detail. The token-based framing has also begun to make the LLM alignment playbook relevant to speech, with Reinforcement Learning from Human Feedback and preference optimization increasingly used to tune models not only for plausible audio, but for the prosody, expressiveness, speaker similarity, and reliability that listeners actually prefer.

In the simplest terms, after training, a text-to-speech (TTS) model is a machine learning system that converts written text into spoken audio. It accepts a string of text as input and produces a waveform, a time-series audio representation, as output. That output determines everything the listener experiences: the speaker's voice character, the intonation of a sentence, the rhythm of a paragraph, and the subtle variations in delivery that distinguish natural speech from mechanical recitation.

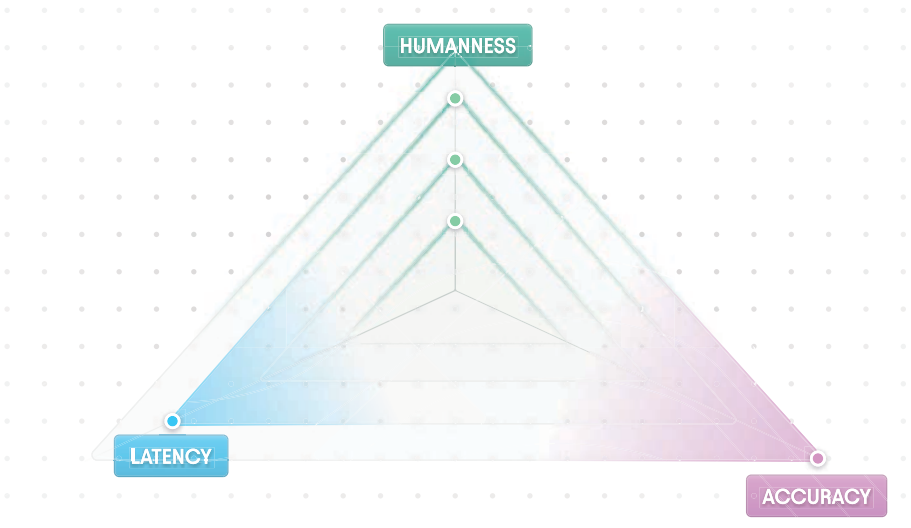
Modern TTS optimization operates across three principal dimensions. The first is accuracy: how faithfully the synthesized audio represents the input text, typically measured through Word Error Rate (WER). The second is latency: how quickly the model begins generating audio after receiving input, measured as Time to First Byte (TTFB) or end-to-end generation time.
The third, and the subject of this paper, is humanness: the degree to which a synthesized voice is perceived by listeners as natural, lifelike, and indistinguishable from human speech.
We contend that humanness is the hardest of the three to measure, the least represented in current model selection practices, and the most consequential to end-user experience.

The Existing Voice Model Landscape

The voice AI market is evolving so quickly that evaluation has become a moving target. New models ship frequently; capabilities that were at the frontier six months ago are now table stakes, and providers' competitive claims have intensified accordingly. Major players supported on the Vapi platform include xAI, Cartesia, Inworld, MiniMax, OpenAI, ElevenLabs and Deepgram, each competing on latency, accuracy, and humanness.

When providers announce "state-of-the-art" models, they are typically making a composite claim: that the new model performs at the frontier across all three dimensions at once. In practice, the weight given to each dimension varies by provider, by architecture, and by target use case. One provider's state-of-the-art may prioritize streaming latency; another may lead on naturalness at the expense of speed.

For Vapi, the priority is clear. We deploy voice AI over the phone, in front of real people, in real conversations. In that context, humanness is not just one of three considerations; it is often the leading determinant of whether the interaction succeeds. A fast, accurate voice that sounds robotic is still a failed experience. This operational reality motivates the Humanness Index™ and distinguishes Vapi's perspective from that of providers evaluating their own models under controlled conditions.



Vapi's position as the platform layer routing calls across all major providers gives us a vantage point no individual provider has: direct, production-scale observation of how these models perform with real users. The models included in this evaluation reflect that breadth.

Defining Humanness

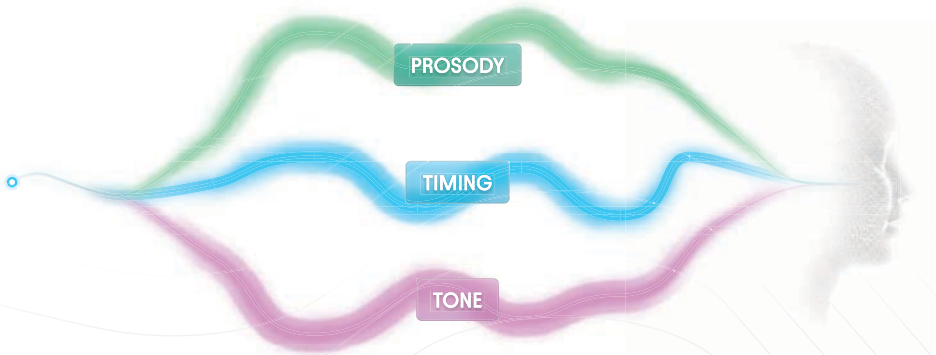
Humanness is, at its core, the Turing test for voice: does the listener believe they are talking to a person? For many voice AI use cases, that is the only question that matters. Every other consideration is downstream of it.

Individual components of a voice, such as prosody, timing, tone, and expressiveness, are attempts to decompose what is a fundamentally holistic perception. They have their place in model development, but they are not what the person on the other end of the phone experiences. Listeners form a single, immediate impression: a voice sounds more human, or less human, relative to what they know human speech to be. That gestalt judgment matters more than the sum of any technical scorecard.

This is precisely the failure mode of existing approaches. When developers ask for a more human-sounding voice, providers tend to break the request into measurable sub-components and optimize each in turn. A voice that scores well on every individual dimension can still fail to sound human because humanness is not experienced as an aggregate of components; it is experienced as a whole.

A technically perfect voice with clean audio, accurate prosody, and consistent timing can still fail the test. Listeners do not evaluate voices analytically; they judge quickly and holistically, and that judgment is the benchmark that matters. Humanness resists full quantification. No formula captures it. Human judgment is not a limitation of the methodology; it is the methodology.

Humans are, by definition, the most qualified judges of what sounds human. The Humanness Index™ is built on this premise. Rather than defining humanness and asking models to satisfy a checklist, it puts voices in front of real people and asks a single question: which one sounds more human? The answer, aggregated across thousands of listeners, will be the Humanness Index™.



Current Voice Model Benchmarks Don't Tell the Complete Story

Voice AI has developed robust quantitative benchmarks for its objective dimensions. Accuracy is measured through Word Error Rate and LLM-as-judge scoring, which is objective, reproducible, and computable at scale. Latency is measured through Time to First Byte, a precise, continuously trackable signal. Both metrics are well-suited to what they measure.

Humanness is different. It is inherently subjective, context-dependent, and ultimately resolvable only by a human listener. The existing proxy, the Mean Opinion Score (MOS), was not designed for the realities of conversational voice AI. MOS evaluations are conducted on isolated audio clips, outside any conversational context. Lab conditions bear little resemblance to real deployment environments. And the expense and time required to run a proper MOS study mean results are often outdated before they can be acted on. These are structural limitations that make MOS the wrong tool for the job, not flaws to be fixed at the margins.

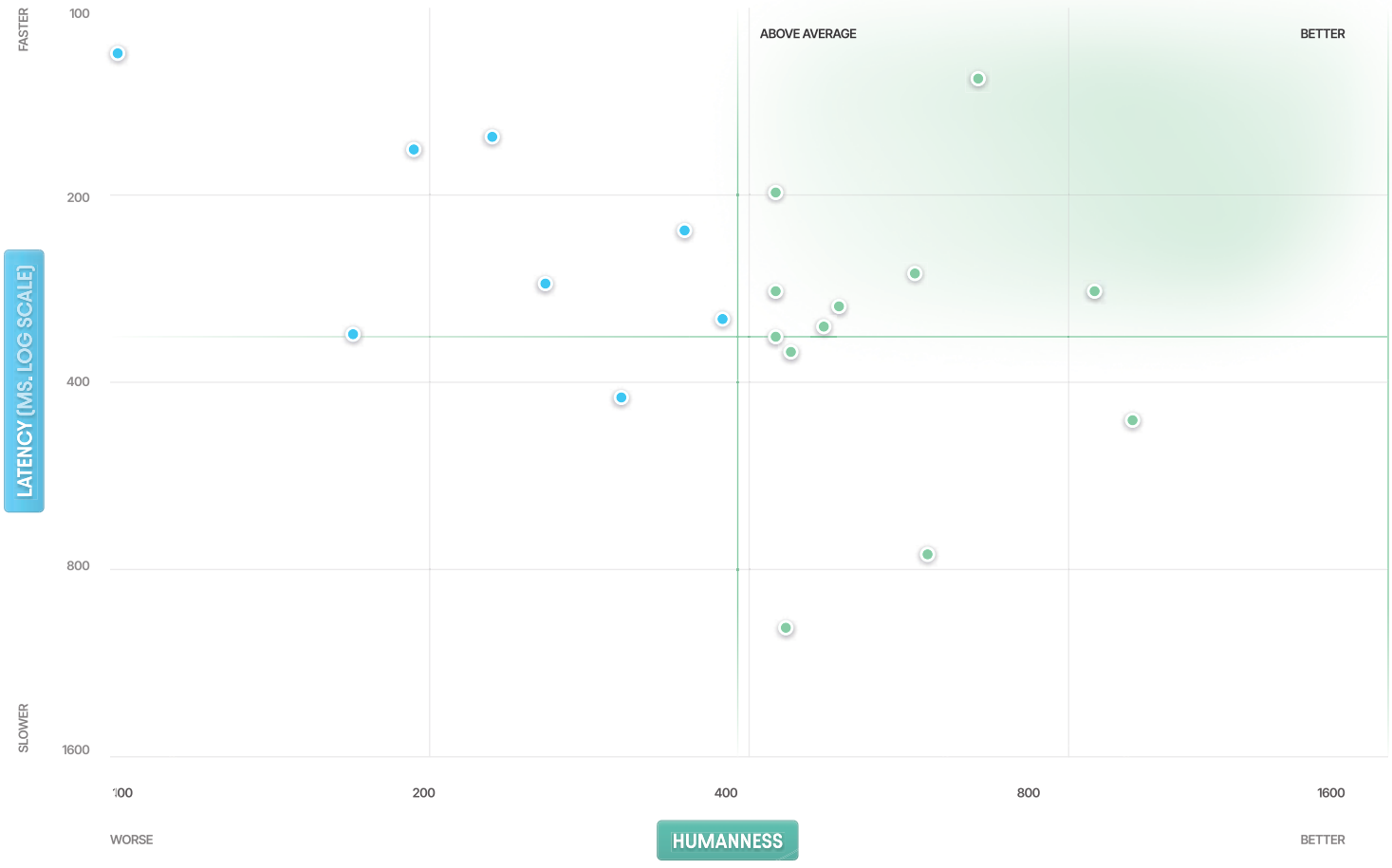
Vapi's position as a conversational voice AI platform gives us a different vantage point: we see humanness where it actually matters, and this in live, multi-turn conversations, not static samples evaluated under controlled conditions.

There is also a methodological problem that has received too little attention: in most existing evaluations, the source voice itself is a variable. When models are evaluated on different input samples, observed differences in output may reflect the input rather than the voice model behind it. Comparisons are not truly head-to-head. The Humanness Index™ controls for this directly: every model is evaluated from the same source clip, eliminating input variability and ensuring that what is measured and compared is the model, and nothing else.

The consequence of these gaps is that teams deploying voice AI at scale are making substantial infrastructure and vendor decisions without a fair, contextually grounded signal on the dimension most consequential to their end users.
That is the problem the Humanness Index™ is built to solve.

Introducing the Humanness Index™?

The Humanness Index™ is Vapi's response to this measurement gap: a live, crowdsourced leaderboard that produces continuously updated humanness scores for leading TTS models based on large-scale human evaluation. It is, to our knowledge, the first benchmark purpose-built to measure TTS naturalness at production scale using real human judgment rather than automated proxies or laboratory studies.

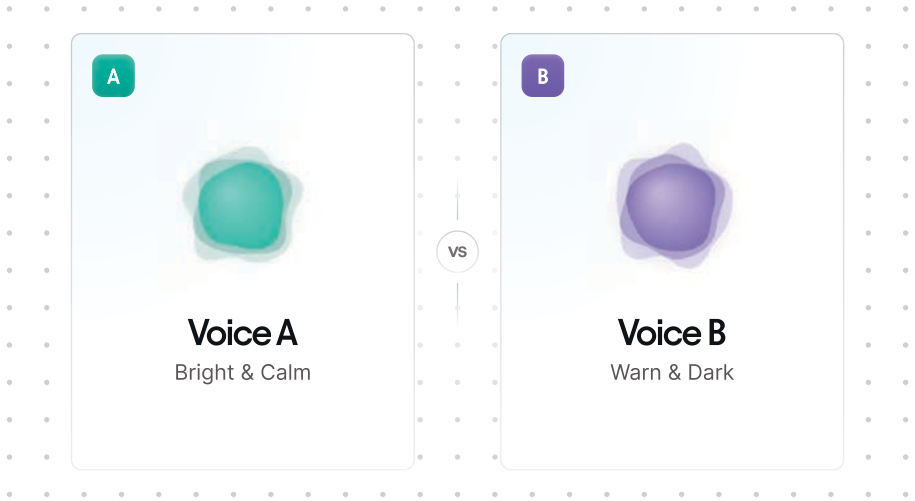


The Index produces a single, interpretable score per model: a winning percentage derived from thousands of pairwise comparisons in which listeners hear two voice samples and select the one that sounds more human.

The Index differs from existing benchmarks in three material respects. First, it measures the dimension that most directly determines end-user experience, humanness, as judged by the humans who perceive these voices, at scale. Second, it is scalable and continuously updated, drawing on a diverse rater population that includes both technical experts and representative end users. Third, unlike provider-published benchmarks, it is produced by an independent, model-agnostic platform with no commercial interest in any particular outcome.

The Index is a living resource. Scores update continuously as new evaluations are submitted and providers release model updates; a score reflects the current state of a model, not a historical snapshot. Every comparison submitted improves the statistical confidence of the rankings, and new models can be incorporated as soon as they are released. Hence, the Index keeps pace with the decisions developers actually face.

Which voice sounds more human?



Methodology for the Humanness Index

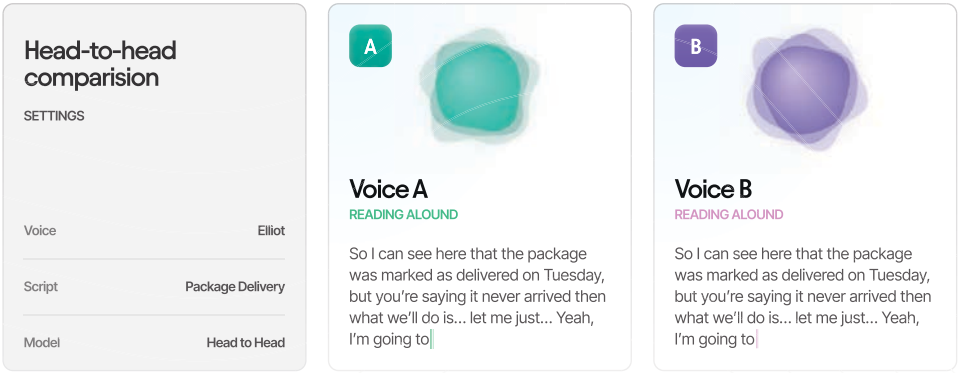
Humanness is a perception, and it must be measured by the people experiencing it. No automated metric has been validated as a reliable proxy for human judgment of naturalness among current top-tier models. Human-in-the-loop evaluation is therefore the core tenet of the Humanness Index™.

The foundation of the Index is pairwise comparison: listeners hear two voice samples and choose which sounds more human, based on whatever that means to them. Pairwise comparison has significant advantages over absolute rating scales. It eliminates the rater calibration problem that manifests when different raters hold different absolute quality thresholds. It also reduces the recency bias that arises when voices are rated in isolation rather than side by side.

The methodology is designed to eliminate the confounds that undermine existing evaluations. The process works as follows:

1. We take source clips of real human speech drawn from 60-minute segments of naturalistic, conversational audio.
2. We use each provider's voice cloning feature to generate the same clip with each model.
3. We put those outputs head-to-head: same source voice, same quote, same audio filters and every variable in the stimulus held constant except the model being evaluated.

The result is a genuinely controlled comparison, something largely absent from TTS evaluation to date.



Scores are computed using a Bradley-Terry pairwise ranking model, which derives a global leaderboard from the full distribution of head-to-head outcomes. Scores are expressed as winning percentages: the proportion of comparisons in which a model is selected as the more human-sounding option. Because scores are relative to the pool of models evaluated, they are best read as rankings within the current landscape rather than as absolute measures.

Samples span multiple categories to capture humanness across the contexts that matter in production: short utterances, longer conversational turns, emotionally charged statements, and technically dense content.

We seeded the analysis with evaluations from internal Vapi stakeholders. With today's launch, we are opening it to the broader public, including our ecosystem of more than one million developers. No weighting is applied by evaluator background or affiliation; the Index reflects the full distribution of human perception, not the consensus of any particular group. The leaderboard updates continuously, and scores always reflect current model versions.

Rank	Likely Rank	Provider	Model	Humanness	Elo	Latency	Price	Votes
BASELINE	—	Human	Homo Sapien	100	1371	–	–	13
#1	#2-3	Provider A	Model 1	78	1298	460 ms	$15	80
#2	#2-4	Provider B	Model 2	75	1286	285 ms	$15	59
#3	#3-9	Provider C	Model 3	66	1256	128 ms	$50	60
#4	#4-13	Provider D	Model 4	63	1249	265 ms	$50	59
#5	#4-13	Provider E	Model 5	63	1249	–	Open source	57
#6	#4-15	Provider F	Model 6	61	1240	758 ms	$100	60
#7	#5-16	Provider G	Model 7	55	1222	325 ms	$60	51
#8	#5-16	Provider H	Model 8	55	1222	302 ms	$50	52
#9	#4-18	Provider I	Model 9	52	1212	315 ms	$60	10
#10	#5-18	Provider J	Model 10	52	1211	288 ms	$25	19

A practical guide for applying the index

The Humanness Index™ is designed as a practical decision-making tool for teams choosing which voice models to deploy. Different deployments optimize for different combinations of latency, accuracy, and humanness, and the right way to use the Index depends on which dimension is most consequential for a given use case.

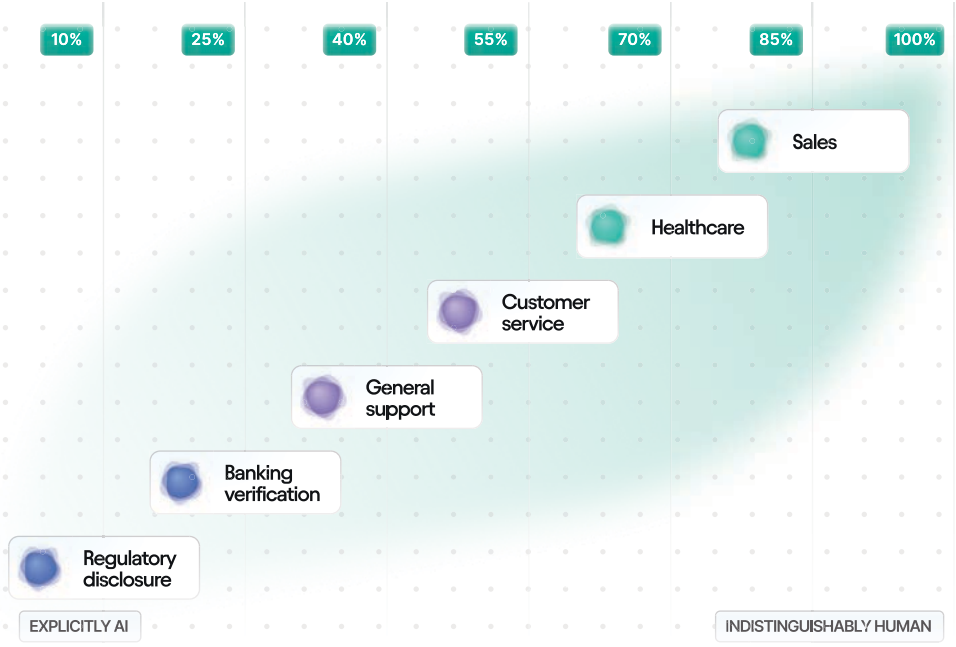
For accuracy-first deployments such as medical documentation, legal transcription and compliance monitoring, metrics such as Word Error Rate remain the primary selection criteria. For latency-first deployments such as high-volume IVR systems and real-time customer service queues, latency benchmarks are the most relevant signal. For humanness-first deployments like outbound sales, high-touch patient communication, and premium customer experience, where success depends on sounding natural, authentic, and personable, the Humanness Index™ will be the most valuable tool. The highest-value deployments, such as enterprise customer-facing agents, typically require strong performance across all three dimensions simultaneously; for these, the Index should be read alongside latency and accuracy benchmarks rather than in isolation.

As starting points, subject to validation against deployment-specific user satisfaction data, we suggest the following thresholds. Scores above 85% are appropriate for the highest-stakes interactions: enterprise sales, healthcare communication, premium customer support. Scores between 70% and 85% are suitable for general-purpose customer service and outbound notifications. Scores below 70% may be acceptable for internal tooling, low-touch automation, and high-volume interactions where naturalness is secondary to throughput. Teams should calibrate these thresholds against their own user feedback and revisit them as the model landscape evolves.

It is also worth noting that maximizing humanness is not universally desirable. In regulated contexts such as healthcare, financial services, and legal, there may be user or regulatory preferences for clearly non-human AI voices to ensure appropriate disclosure. The Index does not prescribe that teams maximize humanness; it provides the information needed to make that decision deliberately rather than by default.

Finally, the Index should be used iteratively. Providers update their models frequently, and new models are constantly being added; a model's humanness score can change significantly with a single release. We recommend incorporating the Index into ongoing evaluation pipelines, not only for initial selection, but for regression testing whenever a provider updates a model in production. Vapi's model-agnostic architecture makes it straightforward to act on those signals.

Humanness Spectrum



What does The Humanness Index™ mean for the industry?

A scalable humanness benchmark has implications beyond model selection. It represents a shift in how the voice AI field understands and communicates model quality, and in how developers, providers, and enterprises participate in defining what "good" means.

The most immediate implication is for teams building with voice AI: humanness should be treated as a first-class evaluation criterion, alongside latency and accuracy, in every voice model selection process. The Humanness Index™ makes this operationally tractable for the first time. Teams can now compare humanness with the same rigor they apply to accuracy or latency, and the bar will keep rising as end users grow more sophisticated and AI voice becomes ubiquitous.

The broader implication is for the industry's quality discourse. The prevalence of state-of-the-art claims in model release announcements has not been matched by any independent mechanism to verify them. The Humanness Index™ provides that mechanism by crowdsourcing judgment from the humans actually experiencing the models. As the Index gains adoption, it should raise the accountability standard for provider benchmarks and accelerate convergence on measurement practices that reflect what users care about.

Conclusion

Voice AI benchmarking today does not tell the whole story. Accuracy and latency are measured with precision, reproducibility, and scale; humanness, which is the question developers and end users care about most, is not. This asymmetry has consequences: it obscures the remaining primary axis of model differentiation, distorts vendor selection, and leaves developers without a reliable signal about the dimension most consequential to end-user experience.

The Humanness Index™ corrects this asymmetry. Through large-scale pairwise human evaluation and by drawing on the breadth of Vapi's developer and business community, the Index produces humanness scores that are scalable, continuously updated, and grounded in real human perception rather than laboratory proxies

We invite the voice AI community, including developers, researchers, practitioners, and end users, to engage with the Index, contribute evaluations, and use the data to make more informed model decisions. The Index will evolve: new models will be added, sample categories will expand, and domain-specific scores for verticals, including healthcare, financial services, and customer experience, are planned for future releases.

The broader vision is a voice AI ecosystem in which every dimension of model quality is measurable, comparable, and continuously improving, and in which the standard for "good" is set not by providers, but by the humans who experience the result.

The Humanness Index™ is available at http://humannessindex.vapi.ai/.
Vote, compare, and check back: the leaderboard changes as the field does.