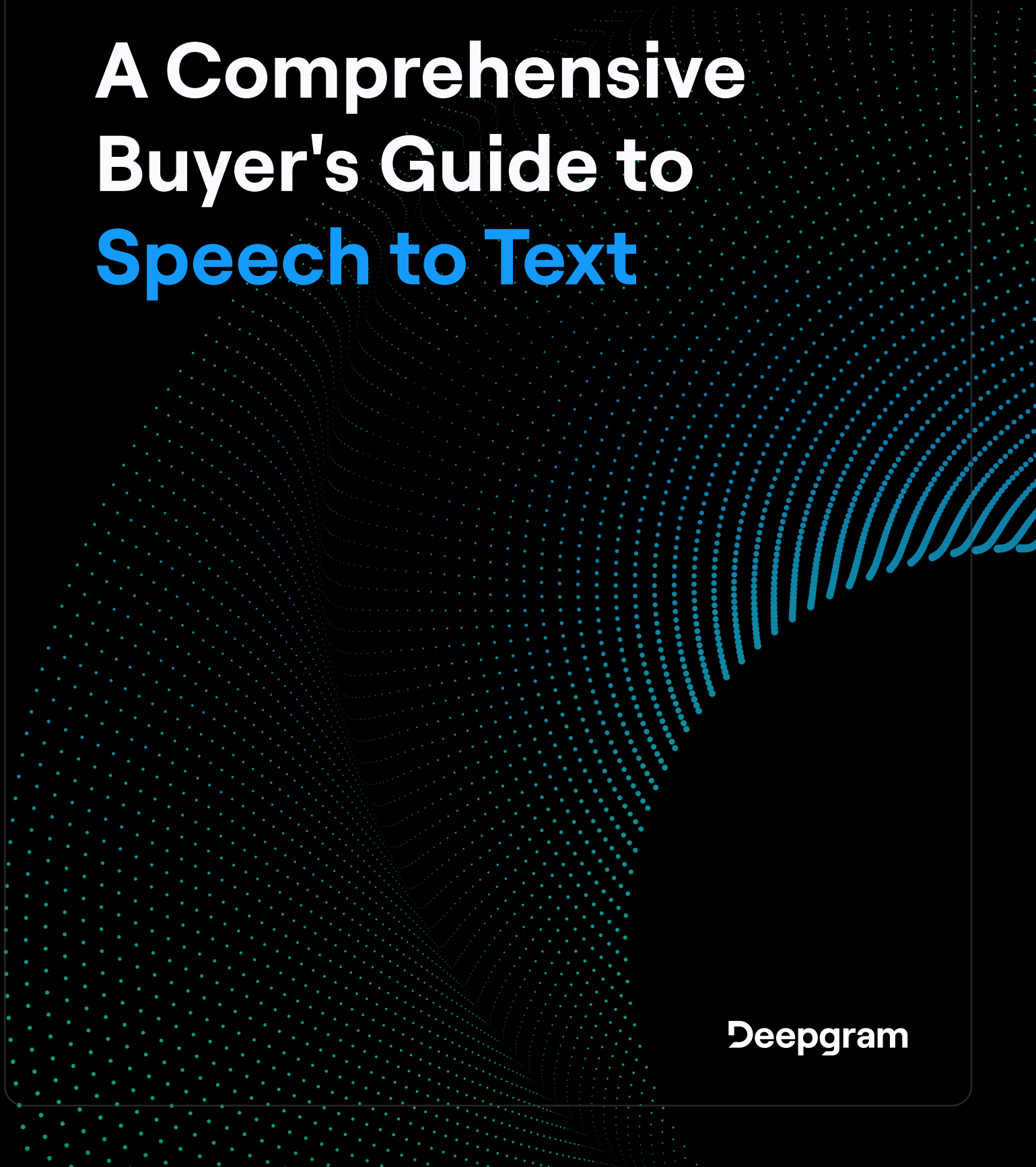
GUIDE

A Comprehensive Buyer's Guide to Speech to Text



Deepgram

Table of contents

Overview

The advent of automatic speech recognition (ASR), also known as speech-to-text (STT), has ushered in a new era of human-computer interaction. Speech recognition and understanding stand as critical components shaping the future of not only the digital interfaces we use on a daily basis but also fundamentally redefining the way businesses, customers, and individuals interact. This transformation transcends professional environments, influencing our personal spaces, and reshaping relationships with family, friends, and trusted advisors, including healthcare providers and financial consultants.

ASR technology, already a familiar tool in our everyday lives, is continuously evolving. Powered by deep neural networks, the next generation of ASR tech offers unrivaled processing speed and transcription accuracy. As an independent application developer or a corporate team, you may wonder: how can I harness this advanced ASR to enhance my product and refine my business strategy?

The potential benefits are significant, but selecting the ideal ASR solution presents a challenging conundrum. Should you partner with a seasoned external service provider, or is it

more advantageous to utilize an open-source solution, building transcription services in-house? Does it make sense to deploy considerable resources to install open-source software with limited capabilities? Or would an off-the-shelf solution, equipped to handle varied requirements from the outset, be a better fit?

A myriad of other considerations arise when contemplating this decision. Which alternative will deliver the most superior results? What are the costs associated with implementing ASR on a large scale, and how do they compare for both options? Which choice will better accommodate future scaling?

Beyond the basic functionality of speech-to-text, you also need to consider what additional features will enable you to resolve unique challenges and offer exceptional value to your customers. How do you balance between performance and cost? Accuracy and speed?

This guide seeks to address these concerns and more, offering comprehensive insight to those considering the integration of speech recognition and language understanding into an application, product, or service. Our goal is to arm you with the crucial information necessary to make the best decision for your team and maximize the benefits of this cutting-edge technology.

Executive Summary

If your application relies on speech recognition or natural language understanding (NLU) to deliver all or part of your customer experience, you've probably questioned whether it's a better long term decision to build in-house or rely on services from a third-party provider.

This question is especially pertinent today as new open source solutions like OpenAI Whisper, Deep-Speech, and wav2vec have emerged alongside leading speech-to-text and NLU providers like Deepgram, Google, and Microsoft.

This guide examines Deepgram's industry-leading model, Nova-2 and compares it against alternative solutions in the market today. Equipped with the information contained in this report, you'll be able to make a more informed decision about whether to build on an open source NLU framework (Whisper), or buy a more complete service offering (Deepgram), as well as the criteria to use to choose which service offering to adopt.

After reading this whitepaper we hope you will walk away with:

- High level overview of Deepgram and ASR use cases
- Comprehensive understanding of Deepgram's ASR platform and the key criteria to use in choosing a best of breed ASR offering
- Introduction to Deepgram's latest ASR model (Nova-2)
- Recommendations for testing model performance for your use case
- Comparisons (with benchmark data) between Deepgram Nova-2 and notable alternatives
- Considerations for building speech technology or using a best of breed service

Introduction to Deepgram

Founded in 2015, Deepgram is a leader in automatic speech recognition (ASR) and Natural Language Understanding. Our expert team, spanning Engineering, Research, Product Management, and Data Operations, has developed powerful models, a streamlined data labeling system, and robust GPU hosting. This approach empowers developers to create scalable, speech-enabled applications with cutting-edge capabilities.

To date, Deepgram has processed over two trillion words from production-grade audio for clients like NASA, Citibank, and Spotify. We provide diverse ASR models—Base, Enhanced, and Nova-2—alongside custom training for specific needs, including phone calls, meetings, finance, and more. Our models offer both pre-recorded and real-time streaming options, rapid processing speeds, and adaptable pricing.

Our ASR supports multiple languages and can be deployed flexibly—on-premises or in the cloud, public or private—making it ideal for various use cases. Unlike other ASR providers, Deepgram offers high accuracy, speed, and cost-efficiency without compromise, delivering a top-tier solution for voice technology.

For more detailed information about Deepgram's product offerings, we invite you to learn more here:

View all models

View all languages supported (over 36)

Explore our models in our API Playground

Sample Speech to text Use Cases

As an API-first platform meticulously engineered by developers for developers, Deepgram enjoys a panoramic view of the diverse ways our clients integrate our ASR capabilities into their products and workflows. Each client is unique, and their application of our technology paints a vibrant tapestry of use cases. Here, we delve into a few typical usage patterns.

Multi-language product development

Clients leverage Deepgram's capabilities to create demo products that offer transcription in one of 36 languages.

Expand use cases:

Some of the most exciting applications of Deepgram can be found in products such as:

- Contact Center Analytics
- Video captions and media transcription
- Conversational AI
- Medical scribes and clinical documentation
- Intelligent voice agents for customer support, order taking, and appointment scheduling
- IVR or Virtual Assistants

Add speech features to existing applications

Deepgram’s capabilities aren't just confined to new projects. We expand the functionality of existing applications by integrating:

- In-app call transcriptions segregated by speaker or condensed into a summary of the call
- In-app recommendations like categorizing and tagging audio based on identified topics, thereby enhancing product search and recommendation abilities.
- In-app analytics or conversational insights by various dimensions such as sentiment, location, topics, date, brand mentions, and speakers.

Enhance call center experiences

Call center experiences can be transformed with Deepgram. The oft-told horror stories of frustrating, time-consuming technical support calls can be rendered obsolete. Many large-scale call centers now use Deepgram to facilitate:

- Multilingual Call Support: Intelligent routing of calls based on the dominant language spoken.
- Article recommendations to agents based on detected topics, aiding in query resolution.
- Reduction in live call listening time, with Deepgram's sentiment analysis capabilities signaling if a call requires intervention from a senior support specialist.
- Automated Data Entry: Calls with prospects, customers, or candidates are summarized and added to the corresponding contact system of record (CRM, HRIS, etc.).

Improve employee productivity

Employee productivity is another area ripe for improvement. The wealth of actionable information contained in speech is no longer challenging to access with Deepgram’s ASR solution. Productivity enhancements include:

- Customer Insights: Extraction of significant, actionable insights from conversations and audio data based on discussed topics and recurring themes.
- Employee Sentiment Insights: A deeper understanding of factors influencing employee productivity and satisfaction, as they express it themselves.
- Quality Assurance: Analysis of conversations based on discussed topics to identify trends, patterns, and enhance the overall customer experience.
- Compliance Demonstration: Usage of speaker diarization and topic identification capabilities to ascertain who said what during discussions about regulated or sensitive subjects.

Deepgram's Features and Support

Developers and companies seeking speech processing solutions require more than a bare transcript; they need rich features that help them build scalable products with voice. Deepgram goes beyond providing just speech-to-text transcriptions. We offer enhanced transcripts and sophisticated speech understanding capabilities to empower the future of intelligent voice applications.

A crucial aspect of our service is the developer experience when integrating and operationalizing a third-party solution. For example, OpenAI Whisper is provided as code. This code must be hosted and maintained on the adopter's machines and is not made available by an application programming interface (API) or a graphical user interface (GUI). Deepgram not only provides an API and robust set of developer tools for its own models, but also makes Whisper available as a model option in its tools and the overall developer experience it provides.

Our focus extends to transcription features that enhance the usability of the raw transcript itself, including:

- Input modes for pre-recorded or real-time audio
- Formatting options like punctuation, numeral formatting, paragraphing, speaker labeling (or diarization), word-level timestamping, profanity filtering, enhance readability and utility for data science
- Tools like keyword boosting, deep search, and find & replace to increase the accuracy and navigability of the transcription output
- Support for languages beyond English
- Use-case specific models for common industry domains for improved accuracy

Deepgram also offers understanding features that analyze the data for context and insights, contributing to a deeper comprehension of the conversation, such as:

- **Speaker Diarization** (separates a transcript into sections based on speaker turns)
- **Sentiment Analysis**
- **Redaction**
- **Topic Detection**
- **Summarization**
- **Speaker Identification**
- **Translation**
- **Language Detection**
- **Entity Detection**

Our services encompass deployment options, including hosted, virtual private and public cloud, and on-prem choices. Furthermore, our partner integrations ensure seamless usage of your speech provider with other components of your stack.

Collectively, these functionalities, concerning developer experience, transcription and understanding, deployment, and integrations, create a tool designed not just for AI researchers but software developers building scalable products.

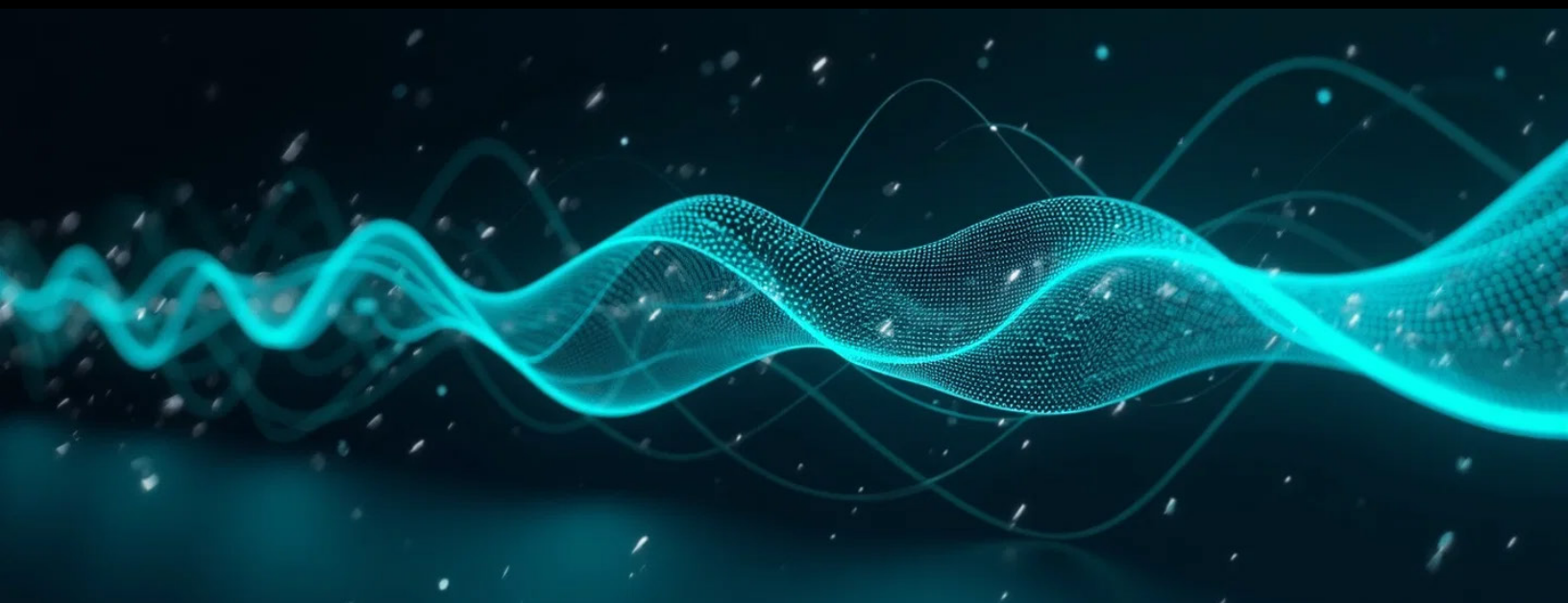
For an in-depth exploration of Deepgram's features, visit our website. Some of our most potent functionalities include:

Formatting	Replacement	Identification	Inference
Punctuation	Numerals	Deep Search	Topic Detection
Paragraphs	Profanity filtering	Keyword Boosting	Entity Detection
Utterances	Redaction	Speaker Diarization	Summarization
Smart Formatting		Language Detection	Sentiment Analysis

Deepgram is dedicated to providing support on all fronts. Whether you need technical assistance, help integrating our services, or maximizing the value out of Deepgram's platform for your application, we're here to assist. Some ways we do this include:

- Offering dedicated **enterprise support** through email and chat, setting up video calls to resolve challenging issues, and maintaining an active user community on Github for quick questions and DIY troubleshooting.
- Providing **model customization** to improve the accuracy of our models, tailoring them to your unique needs. Whether you need to transcribe highly technical medical interactions or work with messy audio data (e.g. lots of background noise, strong regional accents, etc.), with as little as 10 hours of sample data, we can develop a custom model for you that boosts accuracy and delivers superior value.
- Processing both pre-recorded and streaming data to build near real-time intelligence into high performance applications (e.g. a conversational AI, voice ordering system) with no asynchronous API calls.

- Offering **flexible deployment** options, meaning Deepgram models can be used anywhere, including in virtual private clouds or on-premises applications, vital for processing sensitive data and complying with rigorous regulatory or security requirements.
- Providing a **developer-friendly web UI**, with Deepgram Console offering account management and self-serve onboarding tools, which makes getting up to speed with Deepgram a breeze. Deepgram API Playground allows developers, AI engineers, and data scientists to quickly explore and test the Deepgram API – all models, languages, and various language understanding features – in an isolated environment.
- **Delivering APIs and SDKs** for an excellent developer experience in your choice of programming language with official SDKs for Node.js, Python, and .NET, plus community-developed SDKs for Deno and Go. Leverage powerful tools to build and receive context-rich JSON outputs from Deepgram's API.



Introducing Deepgram Nova-2

A Paradigm Shift in Speech-to-Text Technology

In September 2023, Deepgram released **Deepgram Nova-2**, the fastest, most accurate ASR model in the world and follow-up to their initial Nova model (Nova-1) released in early 2023. Nova-2 offers the following advantages:

- **30% improvement in word error rate (WER)**[1] on average over benchmarked competitors for both pre-recorded and real-time transcription
- **Overall WER of 8.4% for pre-recorded audio**, an 18.4% reduction from Nova-1 and 17% lower than the nearest competitor
- 36% relative WER improvement over OpenAI Whisper (large)
- **Overall WER of 10.7% for streaming audio**; 12% lower than nearest competitor
- Lightning-fast inference times that are **5-40x** faster than the competition
- A cost-effective solution, with prices starting from a mere $0.0043/min, making it **3-7x more affordable**

Since the launch of our initial Nova model earlier this year, we have been dedicated to delivering enhanced capabilities. These new features encompass improved speaker diarization, smart formatting, filler words support, and our inaugural domain-specific language model for summarization. These additions not only elevate the value we provide to our customers but also underline our commitment to advancing the forefront of language AI.

In addition, our model research team has maintained an exceptional level of productivity, upholding our longstanding tradition of relentless improvement in the quest for flawless speech-to-text accuracy and pursuit of superhuman transcription performance (refer to Fig. 1).

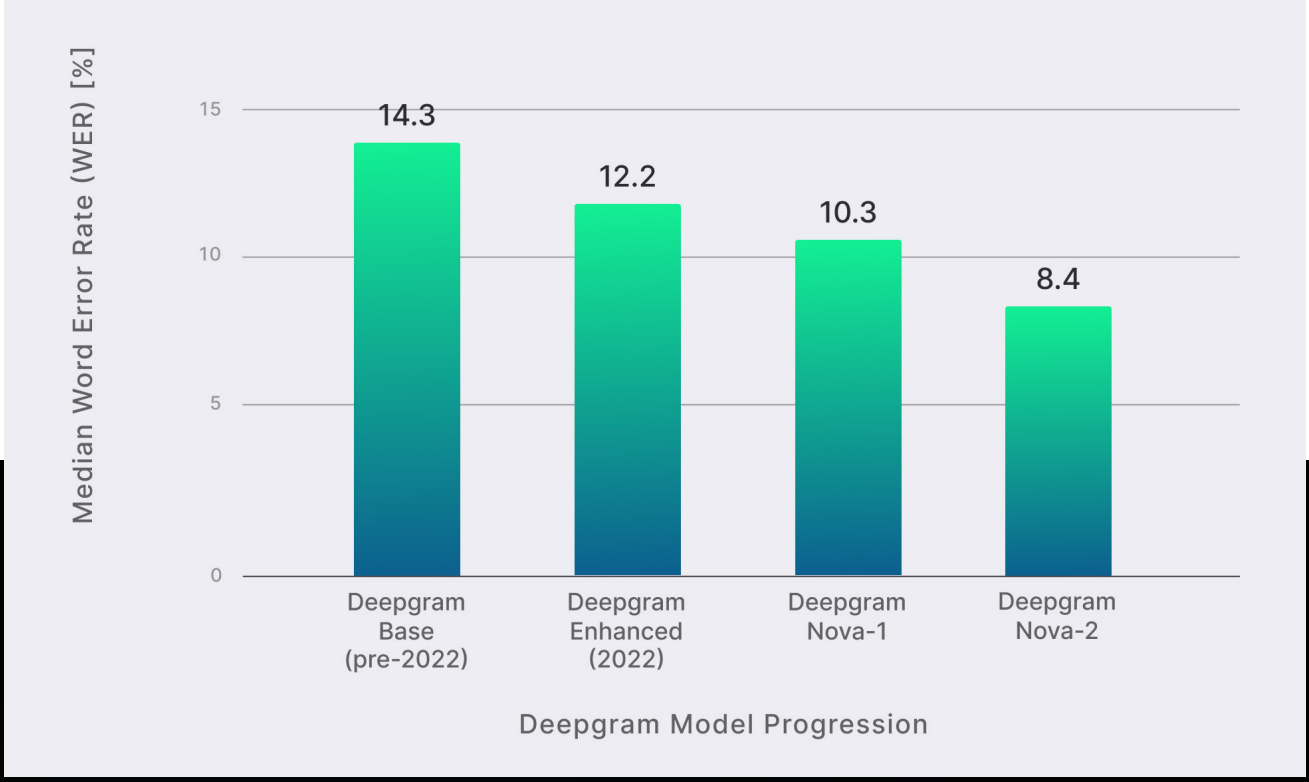


Figure 1: Deepgram model evolution over time. Over the last two years, Deepgram's word error rate has decreased by more than 40%.

Our Approach

While Nova-1 resulted in a significant improvement in accuracy from prior model architectures, Nova-2 extends those advancements even further across diverse audio domains due to three key factors:

1. Speech-specific optimizations to the underlying Transformer architecture.
2. Utilization of advanced data curation techniques, skillfully executed by Deepgram's in-house DataOps team.
3. Rigorous implementation of a multi-stage training methodology, leveraging a substantial real-world conversational audio dataset.

Our model optimization is anchored on an abstraction of the renowned transformer architecture, dividing it into two generalized components:

1. **Acoustic Transformer:** Encodes input audio waveforms into a sequence of audio embeddings
2. **Language Transformer:** Decodes the audio embeddings into text, given some initial context from an input prompt

Information flow between these sub-networks happens through a universal attention mechanism that amalgamates and integrates useful representations from different depths in the Acoustic Transformer network.

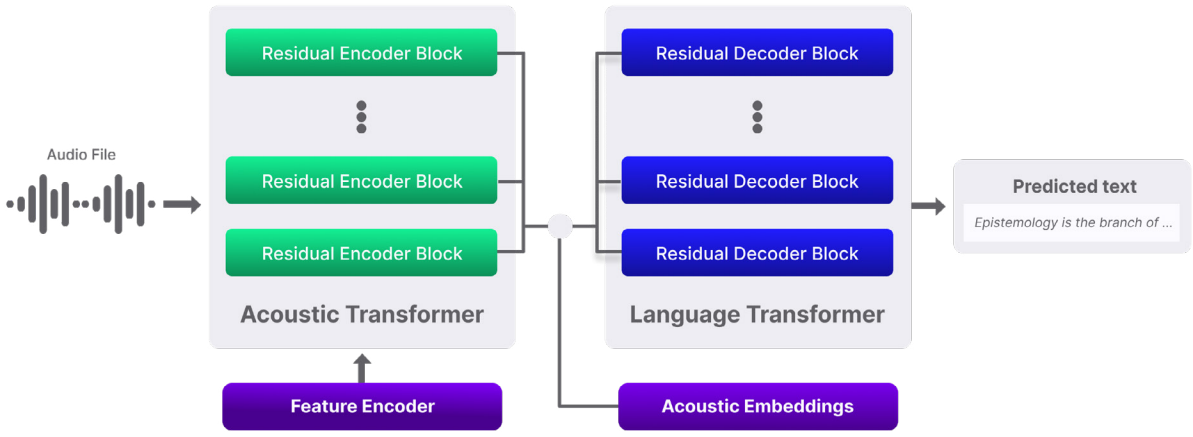


Figure 2: Deepgram Nova-2 High-Level Architecture

Employing proprietary architecture characterization algorithms, we identified weak spots in the transformer architecture that resulted in less-than-ideal performance in both accuracy and speed for audio transcription.

Our team then harnessed our large-scale distributed computing infrastructure to explore the architecture space and identify dozens of architectural innovations. We optimized the Acoustic and Language components separately, designing the arrangement of layer types within the transformer blocks and distributing capacity across the network. These enhancements led to substantial breakthroughs in model accuracy without compromising speed.

The result is a novel Transformer-based architecture that offers significant improvements over its predecessor, as evidenced by an 18.4% reduction in word error rate (WER) from Nova-1. Additionally, Nova-2's architectural enhancements significantly boost accuracy for both pre-recorded and streaming transcription of entities (i.e. proper nouns, alphanumerics, etc.), punctuation, and capitalization.

Improved entity accuracy with a **15%** relative reduction in entity error rate from Nova-1 on entities overall	**22.6%** relative improvement in punctuation accuracy over Nova-1	**31.4%** relative improvement in capitalization error rate compared to Nova-1

Extending upon Nova's groundbreaking training, which spanned over 100 domains and 47 billion tokens, Nova-2 continues to be the deepest-trained automatic speech recognition (ASR) model in the market.

Nova-2 was trained in a 2-stage curriculum starting from the largest, most diverse dataset in Deepgram's history, curated from nearly 6 million resources and incorporating an extensive library of high quality human transcriptions. This extensive and diverse training has given birth to a model that outclasses all other ASR models across various datasets (refer to benchmarks below).

The model was first trained using a substantial dataset of unique audio data in a weakly supervised manner. It then underwent a series of fine-tuning stages with carefully selected, high-quality, domain-specific data. Whereas other models depend on smaller, tightly correlated audio-text training datasets or unsupervised audio pre-training, Nova-2 sets itself apart through its wide

and diverse training data. This meticulous process equipped the model to attain unparalleled accuracy in areas significant to our clients. As a result, Nova-2 doesn't just perform well in a single specific domain; it's your go-to model for versatility and adaptability.

In essence, Nova-2 resets the bar in ASR technology. These major advances are the outcome of years of devoted research, large-scale distributed training, and in-house data labeling expertise. Its thorough training on diverse data establishes it as the most reliable and adaptable model available, perfect for numerous voice applications requiring high accuracy across different contexts. Furthermore, Nova-2 provides an excellent foundation for fine-tuning in specific domains and handling enterprise use cases, regardless of the field.

Detailed Performance Comparison

Accuracy: Undeniably the Industry's Most Accurate Speech-to-Text Model

All speech-to-text solutions aim to produce highly accurate transcripts in a user-friendly format. We advise performing side-by-side accuracy testing using files that resemble the audio you will be processing in production to determine the best speech solution for your needs.

The generally accepted industry metric for measuring transcription quality is Word Error Rate (WER). Consider WER in relation to the following equation:

WER + Accuracy Rate = 100%

Thus, an 80% accurate transcript corresponds to a WER of 20% WER is an industry standard focusing on error rate rather than accuracy as the error rate can be subdivided into distinct error categories. These categories provide valuable insights into the nature of errors present in a transcript. Consequently, WER can also be defined using the formula:

WER = (# of words inserted + # of words deleted + # of words substituted) / total # of words

We suggest a degree of skepticism towards vendor claims about accuracy. This includes the qualitative claim that OpenAI's model "approaches human level robustness on accuracy in English," and the WER statistics published in Whisper's documentation.

One limitation of WER as a benchmarking tool is its high sensitivity to the difficulty of the audio data it measures. For example, testing our product using two different audio files—

one with “easy” audio (i.e., slowly-spoken, simple vocabulary, and good diction, recorded with high-quality equipment in a quiet environment), and another with challenging real-world audio (i.e., a fast-paced conversation full of industry jargon, where the speakers are far from the microphone in a noisy environment and frequently speak over each other)—can result in significant variance in WER from a single model. **The self-reported WER figures from other vendors represent easy audio.**

Our benchmarking methodology for Nova-2 builds upon our previous test suite for Nova. In this iteration, we expanded our assessment by pitting Nova-2 against the latest models for a broader spectrum of competitors, utilizing over 50 hours of human-annotated audio across more than 250 files extracted from real-life scenarios. This encompassed a wide range of audio lengths, diverse accents, varying environments, and subjects.

This stands in contrast to other benchmarks that rely on meticulously curated, pre-cleaned audio data from a limited set of sources, carefully controlled to narrow the scope of test results. This distinction is of paramount importance because the true value of a competitive benchmark lies in its ability to accurately mirror real-world performance, making it more relevant and reliable for assessing practical applications.

Using these datasets, we calculated Nova-2's Word Error Rate (WER) in pre-recorded inference mode and compared it to other prominent models. The results show Nova-2 achieves a median **WER of 8.4%** for all domains/files tested overall, representing a 16.8% relative error rate improvement over the nearest provider (see Fig. 3) and besting the performance of all tested competitors by an average of 30%. Of particular note is Nova-2’s sizable relative WER improvement (36.4%) over OpenAI’s popular, and most performant, Whisper (large) model.

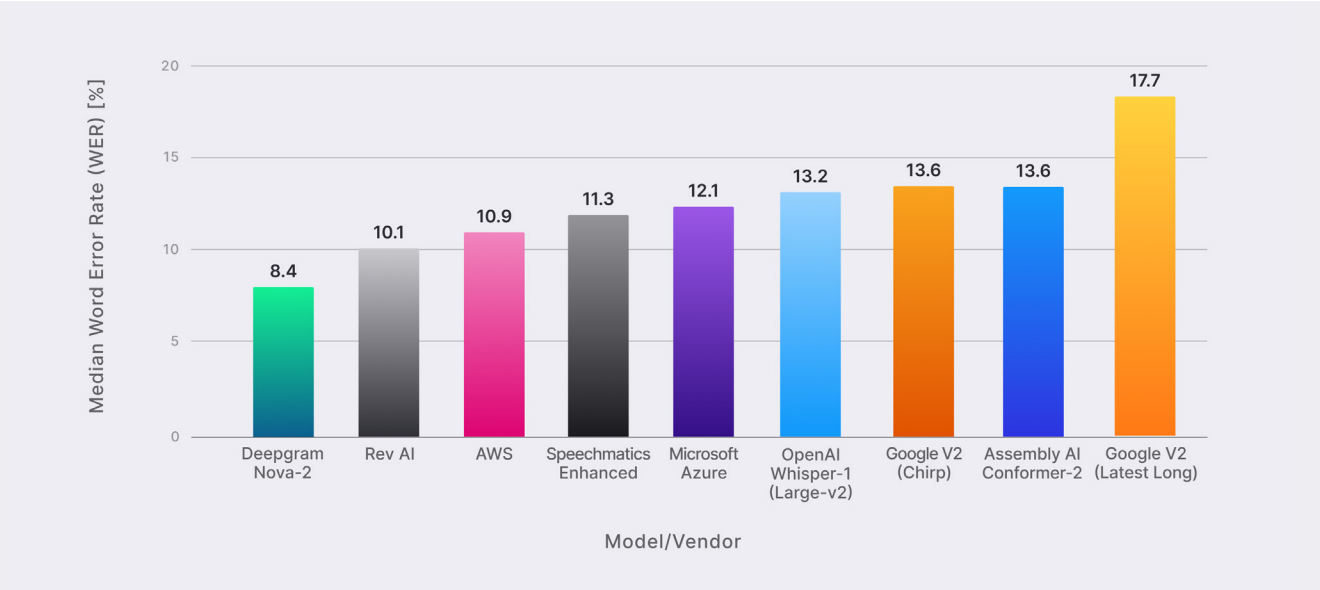


Figure 3: Median word error rate (WER) across all audio domains shows that Nova-2 outperforms commercial ASR models and open-source alternatives like Whisper on real-world data, highlighting its superior accuracy for diverse speech recognition tasks.

More specifically, our benchmarking test set was equally split across four predominant audio domains in the transcription world: podcast, video/media, meeting, and phone call. In domain-specific tests, Nova-2 outperformed all commercial ASR vendors and open-source alternatives, demonstrating exceptional overall accuracy and emerging as the winner within each individual audio domain.

As seen in Fig. 4, Nova-2 displays lower variance than competing offerings as well, with a tighter spread in test results that corresponds with more consistently accurate transcripts in real-world implementations.

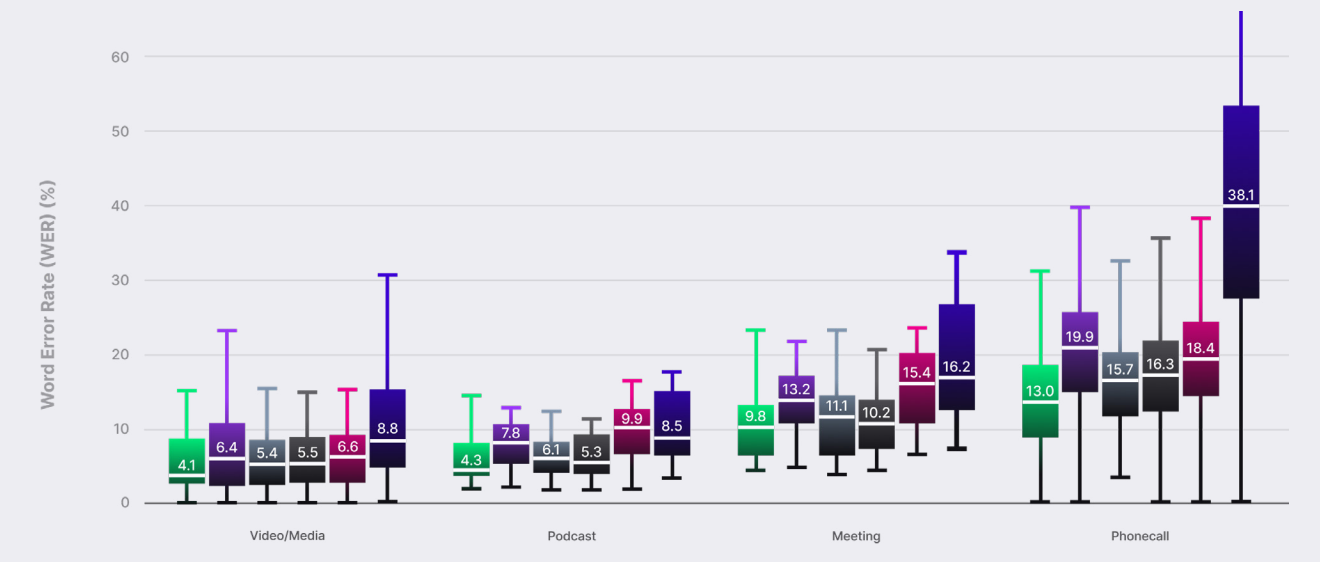


Figure 4: The figure above compares the average Word Error Rate (WER) of our Nova-2 model with other popular models across four audio domains: video/media, podcast, meeting, and phone call. It uses a boxplot chart, which is a type of chart often used to visually show the distribution of numerical data and skewness. The chart displays the five-number summary of each dataset, including the minimum value, first quartile (median of the lower half), median, third quartile (median of the upper half), and maximum value.

In addition to its out-of-the-box accuracy, Deepgram also offers customers the option to train a model on their specific data. This allows for higher accuracy on the accents, keywords, and acoustic environments unique to the customer's use case. Deepgram also continues to improve its use case and language models over time. Customers who choose to have a model trained on their unique data can work with their Deepgram contact to determine the ideal update frequency.

Accuracy

The undisputed leader in real-time accuracy

Modern speech applications–such as real-time agent assist, live captioning of streaming video, and automated food ordering systems–depend on real-time transcriptions to automate interactions with end users and deliver a good customer experience. However, limited options exist for true real-time speech-to-text.

Several providers included in our pre-recorded audio tests lack a native streaming model, such as OpenAI Whisper, so our real-time benchmarking was restricted to a smaller set of alternatives. As the results of our evaluation show in Fig. 5, Nova-2 handily outperforms the field with a median **WER of 10.7%** for all domains/files tested, achieving an average relative reduction in WER of 30% vs. benchmarked competitors overall and 12% lower error rate than the nearest competitor.

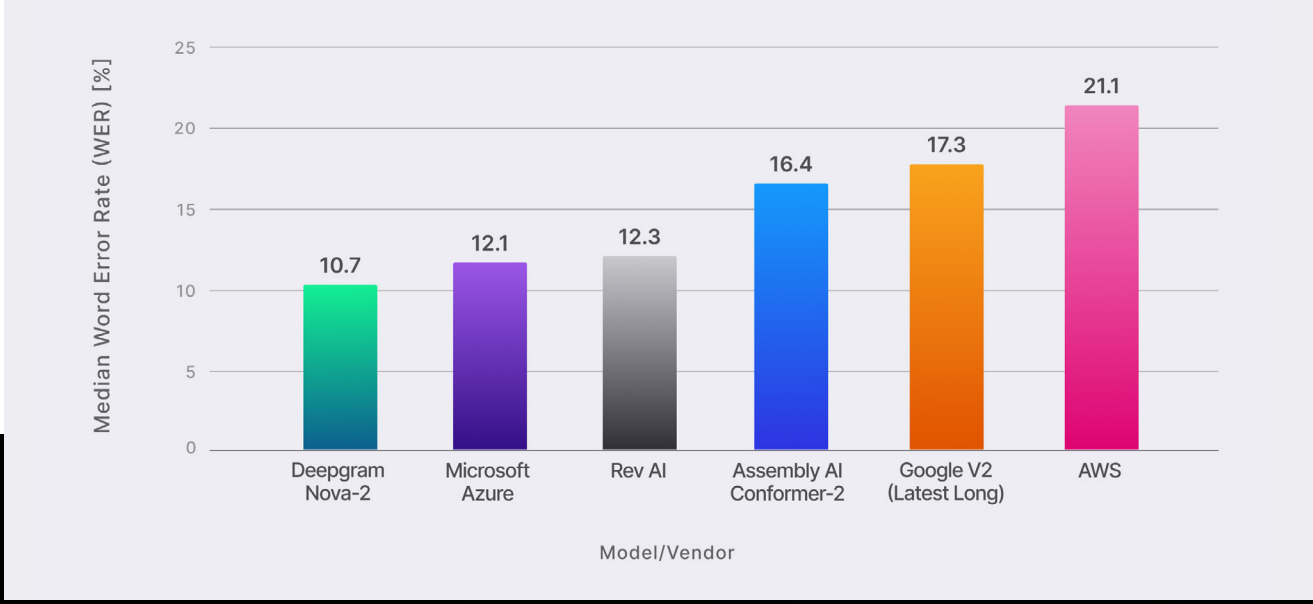


Figure 5: Median file word error rate (WER) for overall aggregate real-time transcription across all audio domains.

> *"When we switched from a cloud vendor's transcription service to Deepgram's Nova-2, we saw a notable leap in **transcription accuracy and responsiveness**"*
>
> **—Tim Houlne, CEO at Humach**

Speed

Hands-Down the Fastest Model

Transcript generation speed is critical for success, especially when fast inference is essential. For example, an IVR system requires responses in milliseconds, making rapid transcription crucial. Deepgram supports both batch processing for pre-recorded audio and real-time processing for streaming, unlike OpenAI Whisper, which only offers batch processing and requires significant effort to enable real-time use.

To evaluate Nova-2's competitive performance, we conducted multiple inference trials across each model/vendor for comparison, measuring total turn-around time (TAT) from request to response using a 15-minute 10 times through each model/vendor. with speaker diarization[3] included, to ensure a fair and direct comparison.

Our findings reveal that Nova-2 surpasses all other speech-to-text models in speed performance, achieving an impressive median inference time of 29.8 seconds per hour of diarized audio. This represents a significant speed advantage, ranging from 5 to 40 times faster than comparable vendors offering diarization.

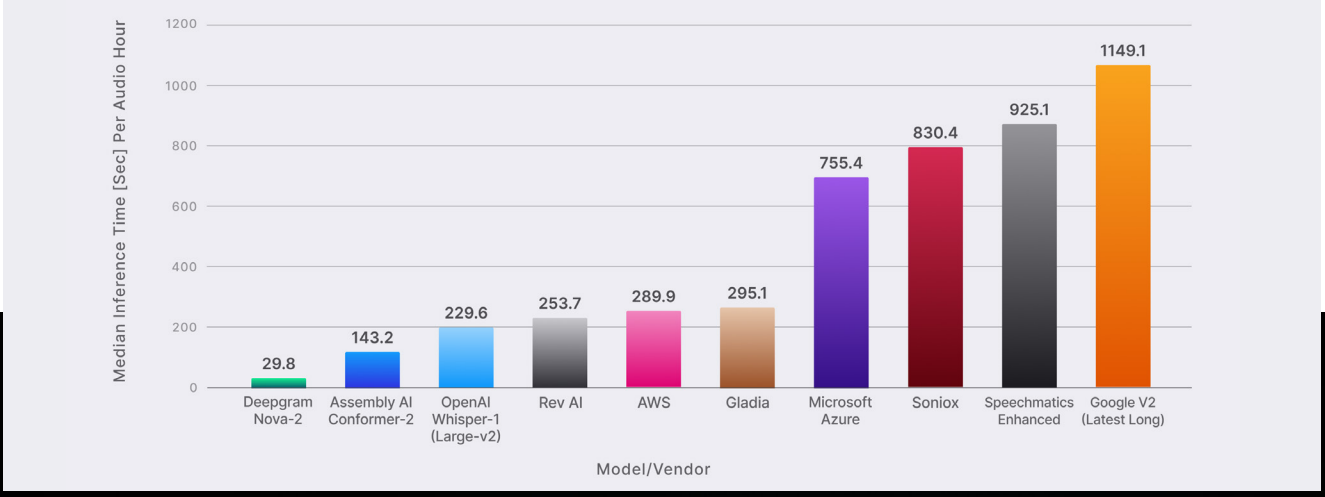


Figure 6: The median inference time per audio hour was measured using longer files, including diarization when applicable. Results are based on multiple trials per vendor for a fair performance comparison.

> *"Deepgram's Nova-2 medical model can transcribe entire phone conversations with precision and speed that rivals human transcribers. Thanks to Nova-2, we've been able to deliver unparalleled ROI to our customers in the healthcare sector."*
>
> **—Rosinol, CEO & Co-founder at Stack AI**

Cost

The Most Affordable Speech to Text Model

Deepgram vs. Open Source

At Deepgram, we're committed to consistently pushing the boundaries in voice AI. Nova-2 is not only ultra-fast and highly accurate but also helps reduce your costs. Our next-generation models achieve new levels of efficiency, which translates into cost savings for you.

With open source models, like Whisper, it's tempting to think their use is free. While there may not be a licensing fee, the costs accumulate quickly when it comes to implementation at almost any production scale.

Even for small-scale tinkering and research use cases, running Whisper's Large model—which delivers the best accuracy but is also the slowest—requires an expensive GPU to produce transcripts on a local installation, and even then the processing times can be frustratingly long.

The cost of purchasing or renting high-end hardware to run Whisper as an in-house speech recognition solution is, if anything, its smallest cost factor. Building on top of open source software requires significant dedicated engineering and research time to integrate Whisper into existing systems, and to continuously optimize system performance to improve accuracy and efficiency as inference requirements scale up. ML engineers and researchers are well-compensated for the value they provide. In other words, this is a case where the "free" option can turn out to be the most expensive. And compared to OpenAI's hosted Whisper API, our own hosted Deepgram Whisper "Large" model (OpenAI's large-v2) starts at only $0.0048/minute, making it about 20% more affordable than OpenAI's native offering.

Deepgram vs. Commercial Competitors

Deepgram's ASR platform has been highly optimized for universal deployment (public/private cloud and on-prem) in a cost-efficient and scalable manner using accelerated hardware. Our scalable infrastructure enables developers to handle high-traffic usage, providing high throughput performance with consistent reliability at any scale. Deepgram offers transparent

pricing that scales with your needs: from multiple self-serve plans to enterprise plans priced to serve virtually any scale of voice processing workload.

For all its advantages in speed and accuracy, Deepgram Nova-2 starts at just $0.0043 per minute and is 3-5x more affordable than any other full-functionality provider (based on currently listed pricing).

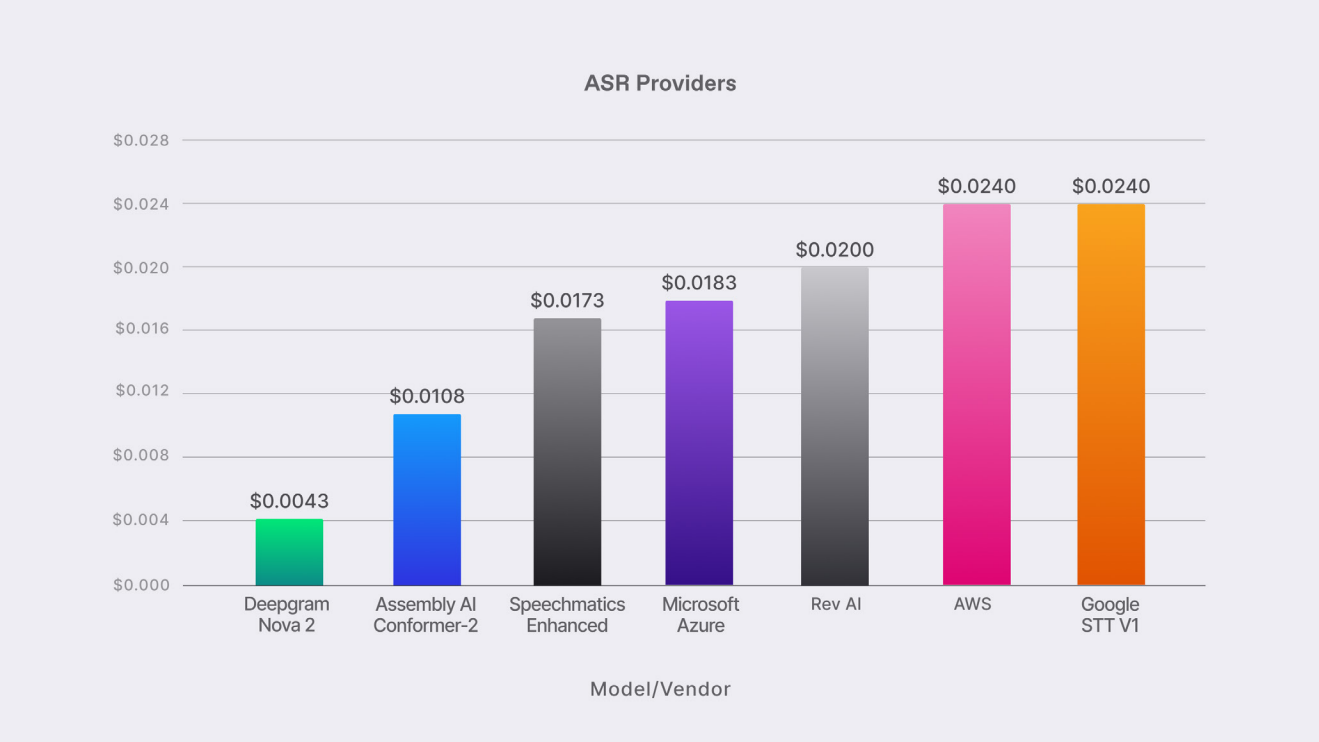


Figure 7: Pre-recorded transcription pricing.

Conclusion

The state of the art of speech recognition is constantly evolving. At Deepgram, we are excited to see advancement to technologies that make speech recognition more useful for solving real problems. Our goal is to make it easy to integrate speech recognition technology into your applications.

We encourage you to test any of the models mentioned in this guide on your own and evaluate performance (accuracy and speed) for yourself while also carefully weighing the cost considerations for your application needs. What tradeoffs are you willing to make across accuracy, speed, and cost? And do you have to make any tradeoffs at all?

At Deepgram, we believe that speech is the hidden treasure within enterprise data, waiting to be discovered. Our mission is to make world-class voice AI not just a possibility but a reality for every developer through a simple API call.

We've processed nearly 10,000 years of audio for trailblazing customers like Citi, Twilio, and Spotify, even transcribing NASA's radio communications between the ISS and Mission Control. We've served nearly 2 trillion enriched, computable words to our customers.

As pioneers in voice AI, Deepgram is committed to transforming how we connect with technology and each other. We firmly believe that language is the key to unlocking AI's full potential, shaping a future where natural language is the backbone of human-computer interaction. With industry-leading ASR provided by models like Deepgram Nova-2, we're one step closer to realizing that future.

Footnotes

- [1] To compare speech-to-text models accurately and obtain an objective score that corresponds to people's preferences, we first ensured that the output of each model followed the same formatting rules by running it through Deepgram's text and style guide normalizer.
- [2] We attempted to standardize speed tests for all Automatic Speech Recognition (ASR) vendors by using files with durations longer than 10 minutes. However, OpenAI's file size limitations prevented us from submitting longer files. Therefore, the turnaround times shown for OpenAI in this analysis are based on the types of shorter audio files that can be successfully submitted to their API.
- [3] Features like diarization were enabled where possible for consistency across vendors under evaluation when making pre-recorded audio transcription calls, however neither OpenAI Whisper nor Google V2 API support diarization. This gives those vendors an advantage in the pre-recorded speed comparisons with results that are inflated compared to those that would occur with such feature support added.

About Deepgram

Deepgram is a foundational AI company on a mission to transform human-machine interaction using natural language. Deepgram gives any developer access to the fastest, most powerful voice AI platform, including models for speech-to-text, text-to-speech and spoken language understanding with just an API call. From transcription to sentiment analysis to voice synthesis and AI agents, Deepgram is the preferred partner for builders of innovative conversational AI applications. Contact us to learn more at Contact Us | Deepgram.

Deepgram

Essential Building Blocks for **Voice AI**