Deepgram

REPORT

Benchmark Report: OpenAI Whisper vs. Deepgram

Table of contents

Overview

Speech recognition and understanding is a big part of the future of interaction: between humans and computers, between customers and vendors, and between you and your work—from your interactions with senior leadership all the way through the stack to direct reports. It may even shape and inform the relationships between you and those who are closest to you: the family, friends, healthcare providers, legal professionals, and financial advisors you entrust along the way.

You almost certainly use automatic speech recognition in your everyday life, and you may be wondering how to incorporate next-generation ASR—enabled by deep neural networks capable of unparalleled processing speed and transcription accuracy—into your product and business strategy.

The opportunity is clear: speech recognition and understanding can unlock enormous value for independent application developers and corporate teams alike. But, do you choose an established external service provider, or opt for an Open Source solution to build upon in-house?

As with any technology adoption decision, the question ultimately boils down to one of Build vs. Buy.

Is it worth spinning up a ton of resources internally to deploy open source software that only gets you so far, or is it easier to use a solution that's ready to handle anything right from the get-go? Which option is going to provide the best quality results? How much is my team going to pay to enable ASR at scale, and how do those costs differ? Which option scales better over the long term? For those who want to integrate speech recognition and understanding into an application, product, or service, this paper aims to give you the information you need to make the best decision for your team.

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 in 2024 as new open source solutions like OpenAI Whisper, DeepSpeech, and wav2vec have emerged alongside leading speech-to-text and NLU providers like Deepgram, Google, and Amazon.

This report compares Deepgram to one of the most recent open source projects to hit the NLU scene: Whisper from AI research company OpenAI. 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).

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

- High level overview of OpenAI Whisper and its model performance
- Overview of Deepgram Models and their performance
- Comparisons between OpenAI open source option and Deepgram
- Considerations for building speech technology or using a best of breed service
- Recommendations for testing model performance for your use case

Anti-Executive Summary / TL;DR

If you are short on time the quick highlights are:

- Deepgram and OpenAI have the same goal in mind: to enable developers to add voice interfaces to a much wider set of applications than is available today.
- Both Deepgram and OpenAI use end-to-end deep learning to create their models. This is the most effective method for accurate and cost effective speech model development.
- OpenAI Whisper intended users are AI researchers, whereas Deepgram is currently used by developers, product teams, and data scientists to build scalable, production-grade products.
- OpenAI models perform well when focused on 1 of 10 languages, low volume, pre-recorded audio
- Deepgram models perform exceptionally well in 30 languages, streaming or pre-recorded, high volume, require low latency / fast speeds
- Test out OpenAI model and Deepgram side by side in our console for free. Sign up here.

Framing The Debate: Build vs. Buy

Before selecting an underlying technology that will determine the performance or perceived quality of your application, there are a few ways to evaluate to build or to buy best of breed solutions. For example, you might ask your development and/or product leadership team:

- Are we looking to gain a competitive advantage? If so, where will that advantage come from? Will it come from advanced speech processing features, lower operating costs, faster processing speeds?
- Are we trying to expand into new markets or attract new audiences? What accent, language, or use case speech models will best set you up for success?
- How much does minimizing time-to-production matter to us?
- What are our requirements around accuracy? Does accuracy only matter overall, or is accuracy more important on key terms like branded products and industry-specific vocabulary?
- Do we have the knowledge and resources available in-house to build and maintain a homegrown speech product?
- Is building and maintaining a homegrown speech product the most highly leveraged use of our limited resources?

With a clear understanding of your opportunities and challenges, you're ready to assess your options. In the following report, we will look at six value areas that are relevant to any product assessment in the speech recognition and understanding market. The six areas are:

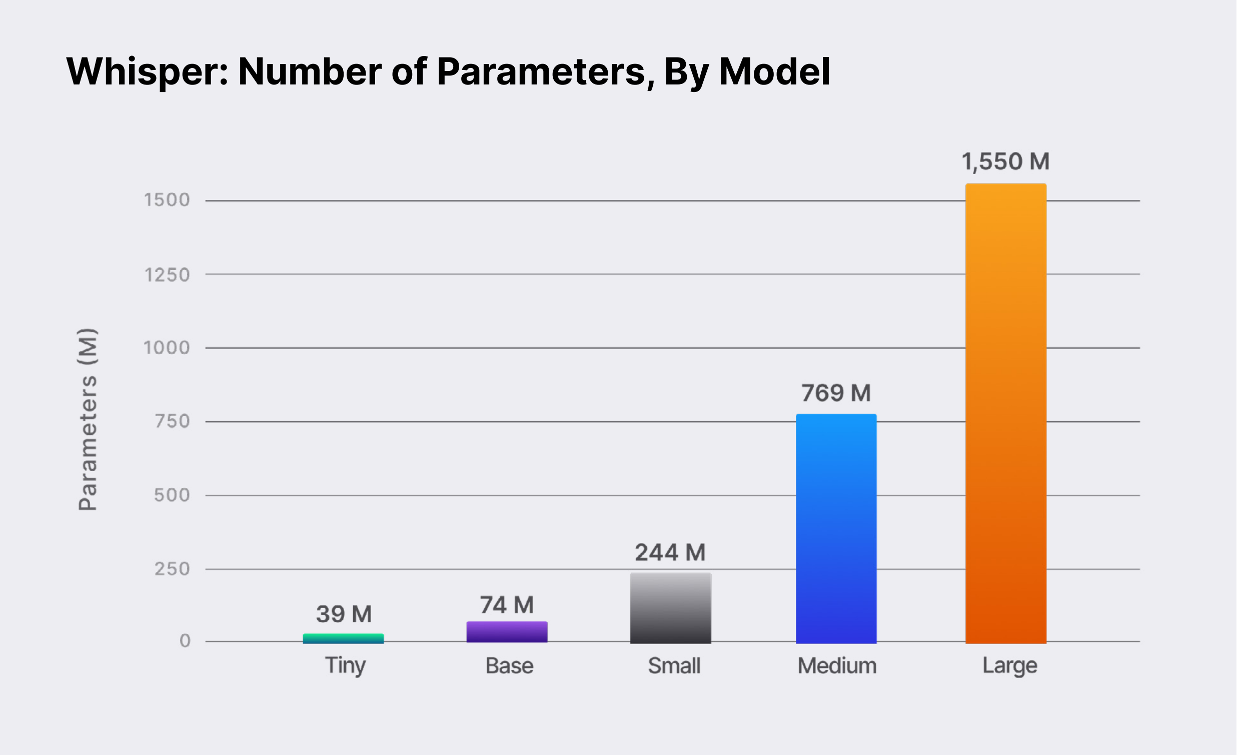
1. Accuracy
2. Cost
3. Speed And Latency
4. Features And Functionality
5. Support
6. Scale

Before getting into the details, though, we'll start with overviews of OpenAI Whisper and Deepgram.

Introduction to OpenAI Whisper

Released in September 2022, OpenAI Whisper models are trained for speech recognition and translation tasks. Researchers at OpenAI developed the models to study the robustness of speech processing systems trained under large-scale weak supervision. The primary intended users of these models are AI researchers studying robustness, generalization, capabilities, biases, and constraints of the current model.

OpenAI offers Whisper in five model sizes, ranging from a few tens of millions of parameters to over 1.5 billion parameters.



Larger models tend to provide higher accuracy at the tradeoff of increased processing speed and compute cost. To increase the processing speed for larger models additional computing resources are required.

Here's a little more information about the different models that are currently available.

Size	Parameters	English-only model	Multilingual model	Accuracy
Tiny	39 M	✓	✓	Lowest
Base	74 M	✓	✓	
Small	244 M	✓	✓	
Medium	769 M	✓	✓	
Large (v1, v2, v3)	1550 M		✓	Highest

Primary Use Cases for OpenAI Whisper

As an open source software package, Whisper can be a great choice for hobbyists and researchers. Here are some of the ways that Whisper can be used:

- Quickly develop a demo product which includes transcriptions of the ten languages that Whisper can currently process.
- Conduct product or technical research on AI speech recognition or non-English to English language translation.

As we'll discuss later, Whisper may not be the right choice for projects involving real-time processing of streaming voice data.

OpenAI Whisper's Features and Support

For an open source project, Whisper's models are certainly better than other, more spartan open source models that came before it. Whisper goes beyond mere speech-to-text transcription, which is unique in the open source landscape, but Whisper does not offer features which haven't already been available in production environments. Again, accuracy varies as a function of the model's size, but in general, Whisper offers:

- Punctuation
- Numeral formatting
- Profanity filtering (English only)
- Voice activity detection
- Language detection
- Language translation

There are, however, some needs which OpenAI Whisper doesn't address:

- **Whisper can't be used for real-time transcription out of the box.** Whisper processes audio in batches, which may disqualify it for certain applications that require real-time transcriptions. Now, there are some clever ways to package and pass brief snippets of audio to Whisper and get results back quickly, but it's still an asynchronous process with additional lag. Whisper does not readily support connections to audio streams.
- **Whisper does not offer model customization.** Some applications require accurate transcription of highly specialized vocabulary (like brands or pharmaceutical names) or deal with audio that would be difficult for a model to accurately process off the shelf. There is no way to train Whisper on your own data. Without the ability to customize a model with additional training, users can do little more than hope their speech data plays well with Whisper.

- **Whisper does not offer keyword detection.** Sometimes, you might want to specify certain words that the model should listen for extra closely, either to highlight their presence (or absence) in a particular piece of speech data, or to boost (or suppress) certain keywords in a transcript. Whisper is currently not capable of keyword detection.
- **Whisper is slow and expensive.** OpenAI Whisper requires extensive computing resources you'll need to requisition yourself. Whether that's a bank of high-end GPUs for a local deployment, or costly cloud computing credits, there are significant costs to running OpenAI Whisper at any kind of production scale. Unlike Deepgram, which can offer flexible pricing as customer needs grow, there are no such economies of scale with running Whisper.

Let's talk about support. The short answer is that OpenAI does not offer ongoing support, integration assistance, or other help with getting the most out of Whisper. There are some user forums for Whisper, including the Community page on OpenAI's website and the discussion section of Whisper's Github page, but otherwise, users are mostly on their own if they run into issues or have questions.

Other Factors to Consider

As with any open source project, it is up to the user to host, develop, and maintain any solution that incorporates Whisper

- Given a limited feature set, users must be prepared to divert engineering and research resources to build and maintain additional functionality.
- Predictions may include texts that are not actually spoken in the audio input (i.e. hallucination). These errors can be catastrophic in certain ASR use cases, such as compliance, finance, healthcare, and legal services.

For those interested in testing the model without needing to host it themselves with the Whisper model, we've made it available through the Deepgram API. Further instructions are available on our blog. And if you really want to get in the weeds, you can follow our tutorial explaining how to build an OpenAI Whisper API yourself.

Introduction to Deepgram

Launched in 2015, Deepgram models are trained for speech recognition and Natural Language Understanding tasks. The product team—which includes Engineering, Research, Product Management, and Data Operations—developed the models and supporting data labeling operations and GPU hosting infrastructure, to enable developers to quickly and easily build highly scalable applications with speech.

Since launch, Deepgram has processed trillions of words from production audio data for customers like NASA, Citibank, and Spotify. We offer our models with rich feature sets, dedicated support, batch and real-time streaming options, fast processing speeds, model customization, and scalable pricing.

Deepgram offers several classes of ASR models—Base, Enhanced, and our most recently released model, Deepgram Nova-2—and optionally offers additional training to customize a model for a specific use case. These use cases include phone call, voicemail, meeting, finance, conversational AI, video, medical, and general purpose speech.

Our ASR technology is polyglot-friendly, supporting multiple languages, and is designed for flexible deployment—either on-premises or cloud-based, public or private. This versatility allows the processing of batches of pre-recorded audio and real-time streams from an array of sources. Unlike OpenAI, Deepgram models do not require users to trade accuracy for speed or computing cost.

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

- View models by tier, pre-recorded, streaming
- View all languages supported (30+ as of publishing this paper).
- View all tailored use case models
- Watch a video tutorial showing how to transcribe podcasts from your terminal
- Explore our models and features with Deepgram API Playground

Size	Parameters	English-only model	Multilingual model	Accuracy	Processing Speed
Base	General Phonecall Voicemail Meeting Finance Conversational AI Video	✓	✓	High	Fast
Enhanced	General Phonecall Meeting Finance	✓	✓	Higher	Fast
Nova-2	General Phonecall Voicemail Meeting Finance Conversational AI Video Medical Drivethru Automotive	✓	✓	Highest	Fast
Trained	All	✓	✓	Highest	Fast

Sample Use Cases For Deepgram

As an API-first platform built by developers, for developers, we get a front row seat to see all the ways people use Deepgram to add Speech capabilities to their products and workflows. Every customer is unique, but here are some of the general usage patterns we've observed.

Multi-language product development

Use Deepgram to create a demo product which includes transcription in one of **30** languages.

Build new voice-enabled applications

Some of the most exciting applications of Deepgram are in products such as:

- Conversational AI
- IVR or Virtual Assistants
- Intelligent agents to assist sales or call center staff

Add speech features to existing applications

Deepgram isn't just a part of new projects; it can help expand the functionality of existing applications by offering:

- **In app call transcriptions** separated by speaker and/or summarize the contents of the call in a short paragraph
- **In app recommendations** such as categorizing and tagging audio based on identified topics to enhance product search and recommendation capabilities.
- **In app Analytics or Conversational Insights** by various dimensions such as Sentiment, Location, Topics, Date, Brand Mentions, and Speakers.

Enhance call center experiences

Many folks have a horror story about that one time they called a company for technical support or customer service, were given the run-around, and eventually gave up after spending 90 minutes on the phone. It can be a day-ruining event that forever damages a customer's relationship with that company.

Eventually, such experiences will be a thing of the past, since more large-scale call center operations use Deepgram to provide:

- **Multilingual call support:** Route calls based on the dominant language spoken
- **Providing recommended articles** to agents based on topics detected to help answer questions or resolve issues.
- **Reduce time spent listening into calls live.** Instead, use Deepgram's sentiment analysis capabilities to be notified if a call goes off the rails and requires intervention from a senior support specialist.

- **Automate Data Entry.** Summarize calls with prospects, customers, or candidates so they can be added to the related contact system of record (CRM, HRIS, etc.)

Improve employee productivity

Speech contains an ocean of actionable information that used to be difficult—if not impossible—to navigate before Deepgram's automatic speech recognition hit the market. Our ASR capabilities can improve productivity by providing:

- **Customer Insights.** Extract meaningful and actionable insights from conversations and audio data based on discussed topics and recurring themes
- **Employee Sentiment Insights.** Get a better understanding of what is contributing to or detracting from employee productivity and happiness. After all, they say it best themselves.
- **Quality Assurance.** Analyze conversations based on the topics discussed, identify trends and patterns, and improve overall customer experience.
- **Demonstrate compliance.** Use speaker diarization and topic identification capabilities to know who said what during conversations about regulated or sensitive subject areas.

Deepgram's Features and Support

Deepgram offers its customers way more than just speech-to-text transcriptions. We've built out the most robust offering on the market: offering enhanced transcripts and speech understanding capabilities that can help enable the future of intelligent voice applications.

If you'll forgive the pun, there's a more full-featured explanation of Deepgram's features on our website, but here's a quick selection of some of our most powerful functionality.

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

When it comes to support, Deepgram has you covered on all fronts. From technical assistance and support with integrating and using Deepgram, to ensuring you get the most value out of Deepgram for your application (whatever it is), we're here to help in whatever way we can. Here are some of those ways:

- **Dedicated enterprise support** is available by email and chat, and we're happy to set up a video call to hash out more challenging issues with you. We also maintain an active user community on Github for quick questions and do-it-yourself troubleshooting.

- **Model customization** enhances the accuracy of our models (which are pretty darned accurate to begin with) to fit your unique needs. Whether you need to capture highly technical medical interactions, or are working with messy audio data (e.g. with lots of background noise, strong regional accents, etc.), so long as you can provide at least 10 hours of sample data, we can spin up a custom model for you.
- **We can work with both pre-recorded and streaming data** so you can build near real-time intelligence into a conversational AI, voice ordering system, and more. There's not a single janky asynchronous API call in sight.
- **Flexible deployment options** mean that you can use Deepgram models anywhere. We offer our own hosted cloud service, but we can package a model for use in virtual private clouds or on-premise applications. This is great news if you need to process sensitive data which may be subject to regulatory or security requirements.
- **A developer-friendly web UI** gives your team more than just an API key and a user manual. The Deepgram Console offers account management and self-serve onboarding tools, which makes getting up to speed with Deepgram a breeze.
- **APIs and SDKs** deliver an excellent developer experience. Interact with Deepgram in your choice of programming languages, and use power tools to build and get context-rich JSON outputs from Deepgram's API. Deepgram has official SDKs for Node.js, Python, and .NET. Our awesome community has developed SDKs for Deno and Go, too!

Detailed Performance Comparison: OpenAI Whisper & Deepgram

In each of the following sections, we will provide data and commentary on OpenAI Whisper as a leading "build" option and Deepgram as a leading "buy" option.

Accuracy

The goal of all speech-to-text solutions is to produce highly accurate transcripts in an easily usable format. To evaluate which speech solution is the best fit for you, we recommend side-by-side accuracy testing using files that are representative of the audio that you will be processing in production.

The industry best practice for measuring transcription quality is Word Error Rate (WER).

Think of WER in terms of the following formula:

WER + Accuracy Rate = 100%

Accordingly, a transcript that is 80% accurate has a WER of 20%.

WER as an industry standard focuses on error rate rather than on accuracy because the error rate can be further broken down into distinct error categories. These categories provide useful insights into the types of errors present in a transcript. WER can thus also be defined using the formula:

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

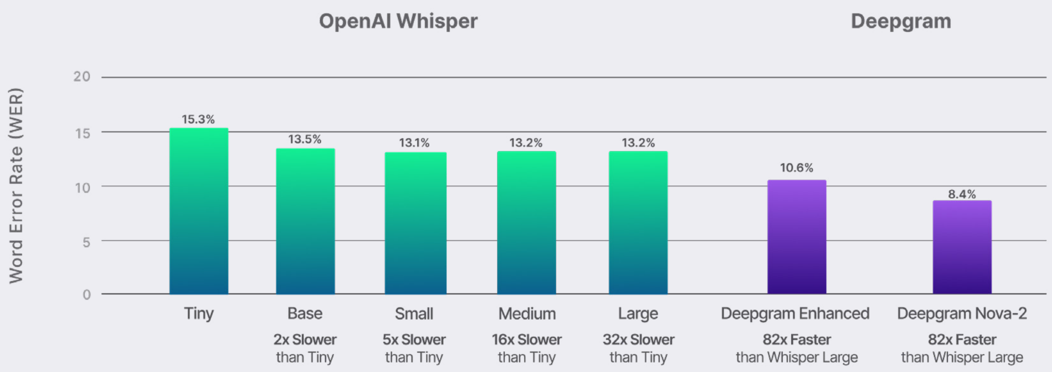
We encourage readers to adopt some healthy skepticism toward any vendor claims around accuracy. This applies both to the qualitative claim that OpenAI's model "approaches human level robustness on accuracy in English" as well as to the WER statistics published in Whisper's documentation.

One of the drawbacks of WER as a benchmarking tool is that it is highly sensitive to the difficulty of the audio data it measures. For example: In testing our own product, it's possible to observe significant variance in WER from a single model by submitting two files—one containing "easy" audio (i.e., audio containing simple, slowly-spoken vocabulary and good diction, recorded with high quality equipment in a quiet environment) and the other containing challenging real-world audio (i.e., audio of a fast-spoken conversation full of industry jargon, where the speakers are far away from the microphone in a noisy environment and frequently talk over each other). **The self-reported WER figures from Whisper and other vendors represent easy audio.**

To run the following WER analysis, we submitted 254 test files to Whisper and the Deepgram Enhanced model. The audio in these files represents real-world audio data from phone calls and meetings. The files contain a range of audio durations and a range of topics, which is important when benchmarking ASR capabilities of a general model. Bear in mind that human-level accuracy can be benchmarked in the 3-5% WER range.

Build vs. Buy: Deepgram ASR Speed and Quality Benchmarks Compared to Open AI Whisper

Word error rate (WER) data displayed below is based on a test of 254 audio titles processed by Deepgram's Enhanced and Nova-2 models and each size of OpenAI's Whisper ASR models.



Deepgram offers superior speed, accuracy, and cost.

Aside from out-of-the-box accuracy, it is important to consider important questions like:

	Whisper (Build)OpenAI	**Deepgram Enhanced (Buy)**
Can I adjust the model accuracy to fit my use case?	**No.** OpenAI's Whisper model can not be fine-tuned or customized as the training code is not released to the public. There is only one model type offered and it can not be altered (e.g. by use case).	**Yes.** Deepgram offers a handful of models trained on data from various use cases, including phone call data, meetings data, earnings calls, and more. Deepgram also offers its customers the option to train a model on the data that matters to them. This allows for higher accuracy on the accents, keywords, and acoustic environments unique to the customer's use case.
Can I expect the model to improve over time?	**No.** OpenAI's Whisper model has been trained on over 680,000 hours of data, but it won't benefit from any updates from the Open AI team. Any further improvements or changes would be made in-house by your engineering or research teams.	**Yes.** Deepgram makes regular improvements to its use case and language models. Customers who opt to have a model trained on their unique data can work with their Deepgram contact to establish the optimal update frequency.

Costs

Since it's open source, it's tempting to think that Whisper is free to use. There may not be a licensing fee, but when it comes to implementing Whisper at virtually any production scale, the costs stack up, and fast.

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

People hoping to deploy Whisper to a public cloud face significant costs as well. We ran benchmark tests on AWS EC2 P4—high-performance compute instances built around NVIDIA A100 Tensor Core GPUs, designed for ML training—and processing 1 hour of audio using Whisper's Large model cost approximately $0.559. Scaling up to a project with 10,000 hours of audio, you'd be looking at nearly $5,600 in cloud computing costs if using Whisper. By comparison, **Deepgram costs approximately 40% less, and delivers higher overall quality with much faster turnaround for a comparable amount of audio.**

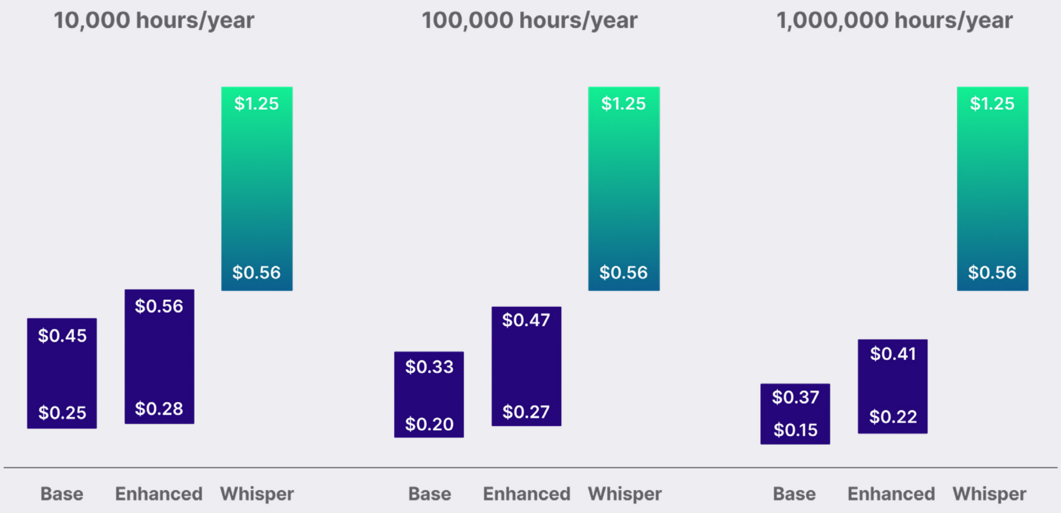
The cost of buying or renting high-end hardware to run Whisper as an in-house speech recognition solution is, if anything, its smallest cost factor. Building on top of open source software requires a nontrivial amount of dedicated engineering and research time to integrate Whisper into existing systems, and to continuously optimize system performance to improve accuracy and efficiency as inference requirements scale up. ML engineers and researchers are fairly well-compensated for the value they can provide. In other words, this really is a case where the “free” option can be the most expensive.

By contrast, Deepgram offers transparent pricing that scales with your needs: from three self-serve plans to enterprise plans priced to serve virtually any scale of voice processing workload.

Every enterprise customer is unique, so Deepgram doesn’t have a single per-hour rate for annual service agreements. Unlike Whisper from OpenAI, Deepgram offers both asynchronous batch processing and can work with streaming audio sources. The chart below shows the price range of processing one hour of audio using Deepgram’s Basic and Enhanced ASR models, depending on the size of the anticipated annual workload. For illustrative purposes, we show price ranges across two orders of magnitude of workload.

Deepgram: Average Per-hour Price Ranges for Batch Processing Enterprise-Scale ASR, vs. Whisper Large

Estimated cost of running Whisper Large is based on benchmark tests conducted by Deepgram. The $0.56 rate is based on pricing for AWS Ec2P4 instances. Whisper costs may exend beyond $1.25 per hour of audio processed if a cloud computing instnace isn't adequately provisioned.

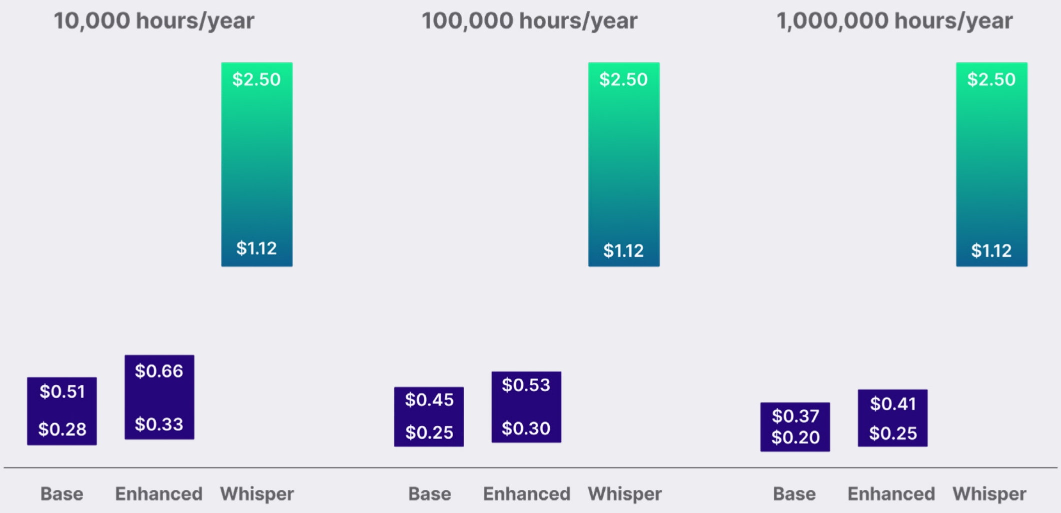


As previously mentioned, Whisper does not work with streaming audio right out of the box. Deepgram does.

Here is a range of average prices similar to the chart above, but this time based on processing real-time audio streams instead of asynchronous batch processing.

Deepgram: Average Per-hour Price Ranges for Real-Time Streaming Enterprise-Scale ASR, vs. Whisper Large

Estimated cost of running Whisper Large is based on benchmark tests conducted by Deepgram. Whisper does not readliy support connecting and processing sources of real-time streaming audio. Cloud deployments which attempt to package and analyze snippets of audio with Whisper cost 2x, or more, versus batch processing.



It’s important to keep in mind that Deepgram handles annual ASR workloads many times larger than the 1 million-hour level noted in the above charts. At higher scale, Deepgram can offer an even lower marginal cost of automatic speech recognition and understanding capabilities.

Lastly, let’s talk about customization. Deepgram’s Base, Enhanced, and Nova-2 ASR model tiers are already at the front of the pack when it comes to accuracy and speed, outpacing Whisper by a mile. To extract even higher performance Deepgram offers optional model customization to suit even the most exacting set of end-user requirements. Depending on the amount of customization required, per-hour processing rates can increase, but the return on investment is worth it. Custom model training can boost accuracy further by 5% to 15%, taking word error rates below 5% in many cases.

Speed and Latency

The time it takes to generate a transcript can make or break your use case. If you’re building an IVR system where your user expects to hear a response in milliseconds, your transcription engine has to be snappy.

Deepgram offers batch processing for pre-recorded audio as well as real-time processing for streamed audio. OpenAI Whisper only offers batch processing for pre-recorded audio. If your use case relies on real-time speech processing, Whisper will not cover your use case unless you devote significant in-house engineering effort to making its model available in real time.

The table below compares processing times for OpenAI Whisper on AWS GPUs and Deepgram on Deepgram's GPUs:

	OpenAI Whisper	Deepgram
1 Hour of Pre-Recorded Speech (Batch Processing)	230s (large-v2)	30s
Latency (Live Streaming)	N/A - not currently available for live streaming audio	< 300-700ms

Features and Functionality

Most developers and companies that are in the market for speech processing solutions need more than just a raw transcript — they need rich features that help them build scalable products with voice.

One important functionality area relates to the experience of using the product's interface. OpenAI Whisper is provided as code. This code must be hosted and maintained on the user'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 GUI for its own models, but also makes Whisper available as a model option in its API and GUI.

Another functionality area includes transcription features that increase the usability of the raw transcript. In this category, you'll find:

- Input modes like pre-recorded audio vs real-time audio
- Formatting options that improve readability and make the transcript more useful for data science purposes, like punctuation, numeral formatting, paragraph and speaker labeling (also known as diarization), word-level timestamping, profanity filtering, and many more.
- Tools that improve the accuracy of the transcription output and make it more navigable, like keyword boosting, deep search, and find & replace.
- Support for languages beyond English.
- Use-case specific models that improve accuracy in common industry domains.

A third functionality area includes understanding features that analyze the data for context and insights that build a deeper understanding of the conversation. In this category, you'll find:

- Diarization, or the ability separate a transcript into sections based on speaker turn
- Sentiment analysis
- Redaction
- Topic detection
- Summarization
- Speaker identification
- Translation
- Language detection
- Entity detection

A fourth area relates to deployment options, including hosted, virtual private cloud, and on-prem options.

Finally, partner integrations allow you to use your speech provider seamlessly with other components of your stack.

Taken together, these functionalities related to interface, transcription, understanding, deployment, and integrations make the difference between a tool that is mainly designed for AI researchers vs a tool that is designed for software developers building scalable products. The table below summarizes the feature offerings of OpenAI Whisper and Deepgram.

Functionality Category	Capability	OpenAI Whisper	Deepgram
Software Type	Closed or open source	Open Source	Closed
User Interface	Application Programming Interface (API)		✓
	Graphical User Interface (GUI)		✓
	SDKs		✓
Transcription	Pre-Recorded	✓	✓
	Live Streaming		✓
	Language Support*	38	30
	Use Case Models	1-General	10
	Model Training		✓
	Word Level Timestamps		✓
	Punctuation	✓	✓
	Numeral Formatting	✓	✓

Functionality Category	Capability	OpenAI Whisper	Deepgram
Transcription (con't)	Paragraphs		✓
	Utterances		✓
	Find & Replace		✓
	Profanity Filtering	✓	✓
	Deep Search		✓
	Keyword Boosting		✓
	Interim Results		✓
	Voice Activity Detection		✓
	Request Tagging		✓
Speech Understanding	Speaker Diarization		✓
	Language Detection	✓	✓
	Entity Detection		✓
	Redaction		✓
	Summarization		✓
	Topic Detection		✓
	Sentiment Analysis		✓
	Language Translation	✓	✓
	Speaker ID		✓
Deployment Options	Cloud Hosted		✓
	On-Prem or VPC Deployment		✓
Partner Integrations	UniMRCP		✓
	Twilio		✓

** Whisper has 38 language models available at WER of 30% or less, as assessed on “easy” audio files. We consider 30% WER to be the outer limit of transcript usability, and consider WER of 15% to be the threshold of a high quality model. We therefore exclude their language models for which OpenAI’s own studies indicate a WER of greater than 30%. Whisper’s documentation states it achieves “strong ASR results in ~10 languages.”*

Support

When putting a system into production, it is important to be able to handle any problems that arise and have the flexibility to address your users' needs. OpenAI offers no dedicated support options for Whisper. Whisper is maintained by a small number of individuals at OpenAI. Additionally, members of the OpenAI open source community can propose and modify the codebase — including introducing breaking changes. If you encounter problems with the code, dependencies, runtime, or model performance, you would have to file an issue in the OpenAI Whisper github repository.

Dedicated support is a central part of how we at Deepgram work with our users. We provide SDKs, sample code, and tutorials for a variety of user needs. Our team includes technical support engineers, sales engineers, solutions engineers, and customer success partners to help our users succeed with production-grade deployments. We work closely with customers to deeply understand their requirements and assist in ensuring they get the best accuracy and performance possible. We also offer community resources in addition to these dedicated support channels. Our goal is to give people the support options they need to succeed. On larger deployments, we may also include service level agreements to make sure that our customers' success is our success.

Scale

As a final point of comparison, let's talk about scalability in the context of speech data strategy.

By design, Deepgram is designed to scale with any applied ASR system, from bedroom hobbyist projects and skunkworks prototypes to some of the largest speech processing workloads on the planet. (And, since we count NASA as a flagship user, some workloads are quite literally out of this world.)

Can Whisper offer the same? In short: no. Whereas Deepgram seeks to deliver the best marginal cost of speech processing at any scale, the costs of running Whisper—at best—only scale linearly, and may compound over time. No matter how much speech data you need to process, with Whisper your cost structure is dictated by the raw processing power of the hardware you buy or rent to run the model.

Does Whisper make sense for certain projects? Of course. Whisper is an excellent entry point to learn about how ASR capabilities can inform or enhance a product or service, but beyond the R&D phase, it's slow and prohibitively costly to run at scale.

From there, Deepgram can handle the rest.

Conclusion

OpenAI has provided a great open source example of how to build accurate, end-to-end deep learning speech recognition. We are excited for the splash that Whisper's release made in the speech tech space, as OpenAI's results validate the end-to-end deep learning approach that Deepgram has been pursuing for nearly a decade.

For certain use cases, Whisper can be a great choice. OpenAI introduced Whisper as best suited for AI researchers interested in evaluating the performance of the Whisper model. It may also be a good tool for building a speech-enabled demo product, as long as the use case doesn't require streaming, advanced functionality, or large scale, and assuming that the user has GPU hosts available. Developers and companies that want to build scalable products with voice will likely gravitate toward a solution that is designed for that purpose. These differences in design become clear when you compare Whisper to Deepgram, as Deepgram offers higher accuracy, richer features, lower operating costs, faster processing speeds, convenient deployment options, model customization, dedicated support, and more.

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 the OpenAI Model and any of the Deepgram models mentioned in this report.

About Deepgram

Deepgram is a foundational AI company on a mission to understand human language. We give any developer access to the most advanced speech AI transcription and understanding with just an API call. Our models deliver the fastest, most accurate transcription alongside contextual features like summarization, sentiment analysis, and topic detection. Contact us to learn more at deepgram.com/contact-us.