1. Introduction

Virtual assistants like Ale rely on real-time transcription to process and respond to user queries promptly. Key challenges include minimizing latency while maintaining high transcription accuracy across diverse audio conditions such as background noise, accents, and varying speech tempos.

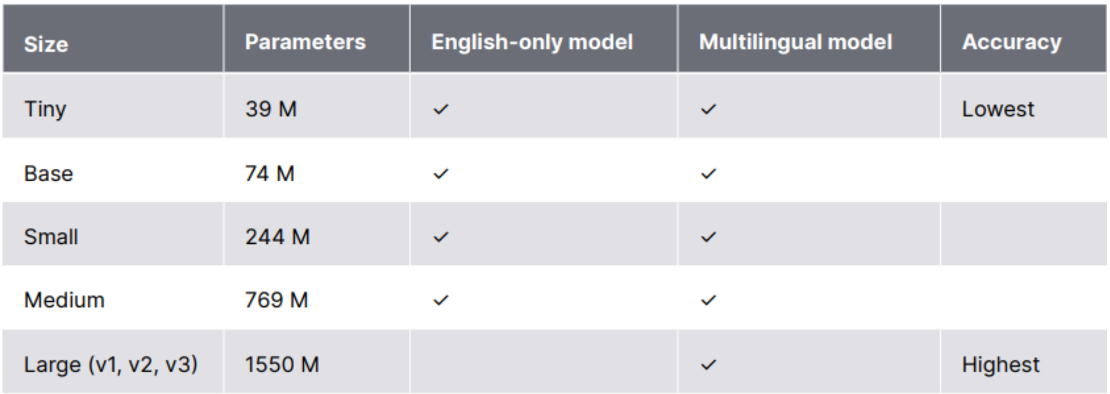
Speech-to-Text (STT) technology enables the automatic transcription of spoken language into text, serving as a critical component in voice-based applications such as virtual assistants, transcription services, and accessibility tools. This paper compares commercial, cloud-based and optimized for real-time transcription STT Deepgram to one of the most recent open source projects to hit the Natural Language Understanding scene: Whisper from OpenAI – A transformer-based, open-source model renowned for its multilingual transcription capabilities and robustness in noisy environments.

2. Model Overview

2.1. Whisper by OpenAI

Architecture:

- Whisper is a family of encoder/decoder ASR models trained in a supervised fashion, on a large corpus of crawled, multilingual speech data.
- Its architecture is "deceptively simple" and comprises a stack of 2D CNNs followed by a symmetric transformer encoder/decoder stack.
- The model ingests 80-dimensional log-mel filterbank features derived from audio transcoded to 16kHz. These are relatively "standard" features.
- Whisper models are available in several sizes, representing a range of model capacities (see table below). There is substantial variation in speed and accuracy across the capacity range, with the largest models generally producing the most accurate predictions but running up to ~30x slower than the smaller ones.



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

Capabilities:

- The Whisper source code takes care of audio pre-processing and can natively handle long-form audio provided directly as input.
- Whisper was trained in a supervised fashion on a very large corpus comprising 680k hours of crawled, multilingual speech data.
- Whisper performs multiple tasks (language detection, voice activity detection, ASR, and translation) despite the decoder only having a single output head.
- In ASR and translation modes, Whisper naturally adds punctuation and capitalization to its output. This is important for end users as it improves the readability of the transcripts and enhances downstream processing with NLP tools.
- Whisper predicts "segment-level" timestamps as part of its output.

- Whisper employs a unique inference procedure that is generative in nature. The default behavior is to infer sequentially on 30-second windows of audio. The audio window is embedded with the encoder and then mapped to a predicted text sequence auto-regressively by the decoder, which uses the encoder output as a context vector.
- Since it's a generative encoder/decoder model, Whisper is prone to some particular failure modes like pathologically repeating the same word or n-gram.

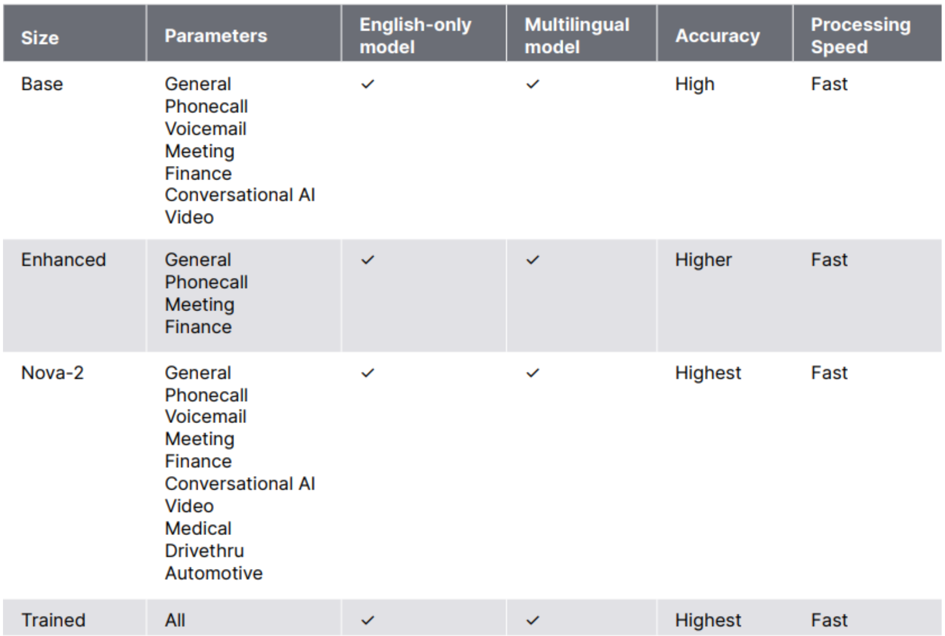
Deployment:

- Whisper requires on-premise deployment and significant computational resources.
- The real-world speed may vary significantly depending on many factors including the language, the speaking speed, and the available hardware. The *.en* models for English-only applications tend to perform better, especially for the *tiny.en* and *base.en* models [1, 5, 7].

2.2. Deepgram

Architecture:

- 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 (see overview table below) [2, 3].
- Deepgram is the first software provider to perform speech training and inference on GPUs. Similar deep neural network (DNN) approaches are typically reserved for image recognition.
- Deepgram serves hundreds of models simultaneously, rather than just the one or two permitted in traditional speech pipelines.
- Custom speech models can be trained on request.



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

Capabilities:

- Deepgram's speech-to-text technology stands out for its precision and speed. By utilizing end-to-end deep learning models, Deepgram achieves higher accuracy rates than traditional transcription methods. This technology can handle diverse accents, dialects, and noisy environments, ensuring reliable performance in real-world scenarios.
- Real-time transcription is one of Deepgram's key features, enabling instant conversion of speech to text. This is particularly advantageous for applications such as live captioning, real-time customer support, and interactive voice response (IVR) systems. The ability to transcribe speech in real-time enhances user experience and operational efficiency.

- Deepgram's audio intelligence features allow for advanced analysis of audio content. By detecting sentiment, intent, and topics within conversations, businesses can gain valuable insights into customer behavior and preferences..

Deployment:

- Deepgram is an enterprise-grade, cloud-developer service.
- Programmable API allows developers without deep data science expertise to run speech recognition models at scale.
- Deepgram is Kubernetes-ready with Docker images, and has pre-built VM images to enable rapid deployment to most cloud providers.
- Deepgram has been built from the bottom up with performance as its highest priority. Existing frameworks and toolkits (e.g., PyTorch, TensorFlow) are not designed for this level of transcription speed, and are too slow to perform at enterprise requirements.
- Deepgram's patented infrastructure outperforms other deep learning frameworks by over 30%.

3. Comparison Criteria

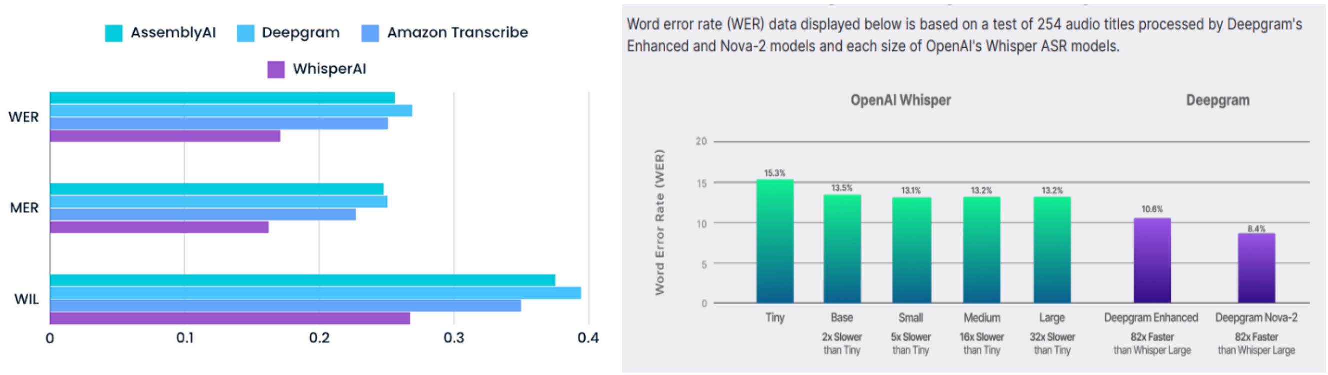
3.1. Accuracy:

- **Word Error Rate (WER)**:

Some contradictive results have been reported for WER values of Whisper and Deepgram models. While independent researchers report better performance of Whisper, Deepgram in its official report claim lower WER rate in its turn.

Overall, we can note the following objective statements:

- Whisper demonstrates low WER in noisy and multilingual environments;
- Deepgram performs well in clean environments but may require fine-tuning for diverse conditions.



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.

Fig 1. ASR models error rate evaluation. Source: [4, 6]

- **Handling of Accents and Dialects**:

- Whisper is known to be strong in diverse audio datasets due to its robust training.
- Deepgram is effective with pre-trained models but benefits from domain-specific tuning.

3.2. Latency:

- Whisper has got a higher latency due to computational demands of its transformer architecture.
- Deepgram: is optimized for low-latency, real-time transcription through its cloud infrastructure.

3.3. Ease of Integration:

- Whisper: As an open source software package, Whisper allows quickly develop a demo product which includes transcriptions of the ten languages that Whisper can currently process and conduct product or technical research on AI speech recognition or non-English to English language translation.
- Deepgram supports an easy integration using cloud-hosted APIs and SDKs for multiple programming languages: JavaScript, Python, .NET, and Go.

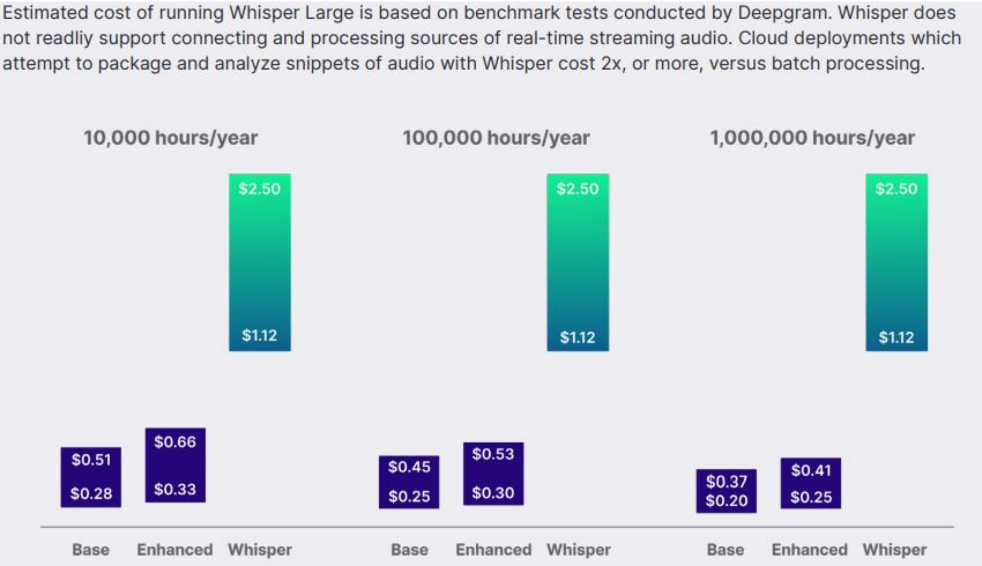
4. Resource Requirements.

Whisper involves high GPU/CPU requirements for local deployments and larger model size impacts scalability.

Deepgram offers minimal client-side resource usage but depends on stable internet connectivity.

5. Cost and Scalability:

As can be seen from Fig. 2, despite Whisper is free and open-source, it incurs infrastructure costs for deployment, which are expected to be higher than Deepgram's subscription-based pricing.



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.

Fig. 2. Average Per-hour Price Ranges for Real-Time Streaming Enterprise-Scale ASR, vs. Whisper Large. Source: [7]

4. Use Case in Ale

- **Whisper** is **s**uitable for scenarios requiring robust transcription across diverse datasets or offline use cases.
- **Deepgram** is ideal for real-time transcription with minimal setup and reliable cloud infrastructure.

Recommendation:

This white paper provides actionable insights to guide the selection of an STT model for Ale's virtual assistant use case, balancing accuracy, latency, and scalability.

✓ For Ale, Deepgram is preferred for real-time transcription due to its low latency and ease of integration. Whisper may be more appropriate for offline or complex multilingual scenarios where robustness outweighs latency concerns.

5. References and Benchmarks

1. Whisper GitHub. https://github.com/openai/whisper
2. Deepgram Documentation. https://developers.deepgram.com/docs/
3. *How to Make Your Application Voice-Ready*. Deepgram. https://deepgram.com/
4. Kiefer A. (Feb 5, 2024). *It Started With a Whisper.* Medium. https://medium.com/@askiefer/it-started-with-a-whisper-4090d26d95e4
5. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision. ArXiv, abs/2212.04356.*
6. Deepgram. (Dec 19, 2022) *Benchmarking Top Open Source Speech Recognition Models: Whisper, Facebook wav2vec2, and Kaldi*. https://deepgram.com/learn/benchmarking-top-open-source-speech-models
7. *Benchmark Report: OpenAI Whisper vs. Deepgram.* Deepgram. https://deepgram.com/