Groq Automatic Speech Recognition (ASR) API

The Automatic Speech Recognition (ASR) API from Groq provides developers and enterprises access to powerful ASR models (also known as Speech-to-Text), including Whisper Large v3, Whisper Large v3 Turbo, and Distil-Whisper. Running on Groq® LPU™ AI inference technology, the Groq ASR API provides ultra-low latency audio transcription and translation. It is available at console.groq.com.

Whisper Large v3, Whisper Large v3 Turbo, and Distil-Whisper are available on-demand or can be deployed as private, dedicated instances in GroqCloud™. They can be fine-tuned for specific languages or industries.

All three models provide high-quality speech but they differ in terms of speed, accuracy, and cost. In this document, we'll guide you through the key features and performance of each model, helping you make an informed decision for your specific use case.

Model Overview

Whisper Large v3

Whisper Large v3 is a multilingual ASR model that supports transcription and translation in multiple languages. It offers a high level of accuracy and speed, with a Word Error Rate (WER) of 10.3% and a real-time speed factor of up to 300x. It is ideal for applications that require fast and accurate transcription and translation in various languages.

Real-world examples:

- Global customer support: A multinational company uses Whisper Large v3 to provide customer support in English, Spanish, French, and Mandarin. The model transcribes and translates customer calls, enabling the company to respond promptly and effectively to customer inquiries.
- International business meetings: A company uses Whisper Large v3 to transcribe and translate business meetings, enabling participants to focus on the discussion rather than taking notes.
- Multilingual subtitling: A video streaming platform uses Whisper Large v3 to generate subtitles in multiple languages for TV shows and movies, making them more accessible to a global audience.

Whisper Large v3 Turbo

Whisper Large v3 Turbo is a pruned and fine-tuned version of Whisper Large v3 that supports transcription in multiple languages. It provides an incredible balance of the benefits of Whisper Large v3 and Distil-Whisper in terms of speed, quality, and capabilities with a WER of 12% and a real-time seed factor of up to 247x. This model is ideal for applications that require fast and accurate multilingual transcription.

Real-world examples:

- Voice-controlled interfaces: A multinational smart home company uses Whisper Large v3 Turbo to power their voice-controlled interface and provide rapid multilingual speech recognition.
- Real-time closed captioning : A non-profit organization uses Whisper Large v3 Turbo to provide real-time transcription for individuals with hearing impairments, helping them to follow conversations and lectures in real-time.
- Customer service call transcriptions: A multinational retail company uses Whisper Large v3 Turbo to transcribe customer interactions, making it easier to analyze and respond to customer feedback and concerns.

Distil-Whisper

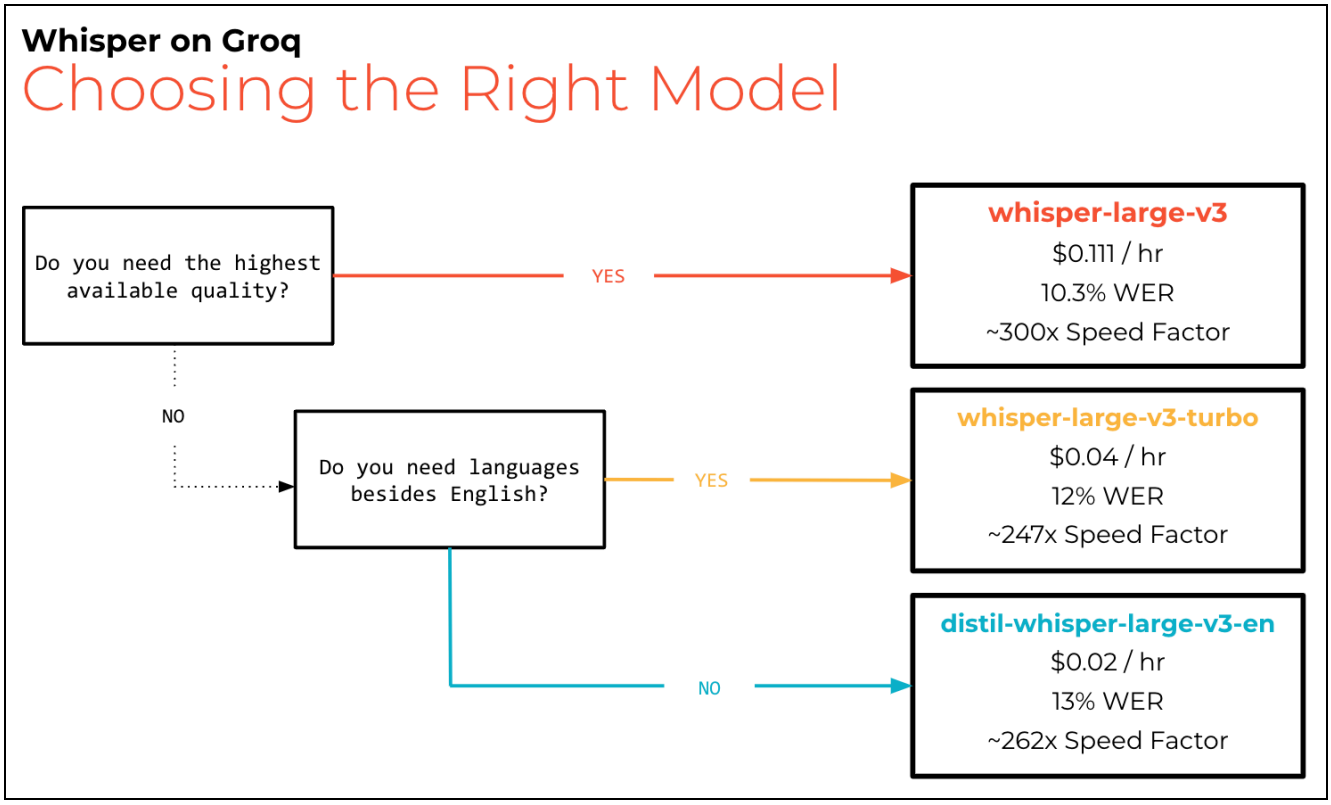
Distil-Whisper is a compressed version of Whisper Large v3, fine-tuned specifically for English speech recognition. It offers incredible speed while maintaining accuracy, with a WER of 13% and a real-time speed factor up to 262x. This model is ideal for applications that require fast and accurate English transcription, such as real-time customer service chatbots, automated speech-to-text systems, and voice-controlled interfaces.

Real-world examples:

- Customer service chatbots: A company uses Distil-Whisper to power its English customer service chatbot, enabling customers to interact with the chatbot in real-time and receive accurate and helpful responses.
- Transcribing audio and video recordings: A media company uses Distil-Whisper to transcribe audio and video recordings in English, enabling journalists to focus on editing and analysis rather than transcription.
- Automated speech-to-text systems: A healthcare organization uses Distil-Whisper to transcribe medical dictations in English, enabling doctors to focus on patient care rather than paperwork.

Whisper Model Comparisons on Groq

Not sure which model is best for your usage? See below for a decision tree and a quick comparison of the three models to help guide you.



Model Specifications

Item	Whisper Large v3	Whisper Large v3 Turbo	Distil-Whisper
Transcription Speed[1]	~300x real-time speed factor	~247x real-time speed factor	~262x real-time speed factor
Accuracy[1]	~10.3% WER	~12% WER	~13% WER
Language Support	Multilingual	Multilingual	English only
Use Cases	Translation and transcription	Transcription only	Transcription only
Pricing[2]	$0.111 per hour	$0.04 per hour	$0.02 per hour
FIle Based	Support for file-based transcription only - no streaming support		
API Compatibility	Compatible with OpenAI API specification		
Supported Audio Formats	mp4, mpeg, m4a, wav, mp3, ogg, webm, opus		
Max Audio File Size	Default up to 25 MB (~30 minute audio files); Up to 100 MB on Whisper Large v3 for paying customers Note: Groq's Whisper implementation is fastest at audio files above 4 minutes		
SDKs Available	Python SDK, JavaScript SDK		
Audio Upload	Upload audio files directly, no audio URL required		
Rate Limits	Visit console.groq.com for the latest rate limits		

1. Speed and accuracy as of 1/20/2025 measured by ArtificialAnalysis.ai
2. Pricing as of 1/20/2025 see groq.com/pricing for the latest pricing. Minimum charge is 10 sec per request.

Whisper models require a minimum of 30 seconds of audio. Audio files shorter than 30 seconds are padded with silence. We recommend audio files longer than 30 seconds to optimize throughput.

Interested in Learning More?

For more information visit groq.com or contact us at sales@groq.com.

Legal Statements.

Groq Inc. HQ
301 Castro St. Suite 200
Mountain View, CA 94041

Mailing Address
PO Box 1778
Mountain View, CA 94041

www.groq.com
console.groq.com