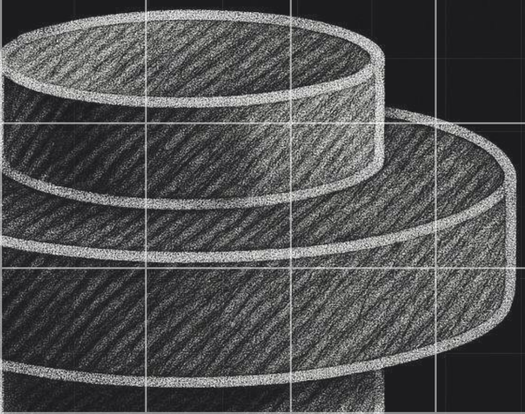
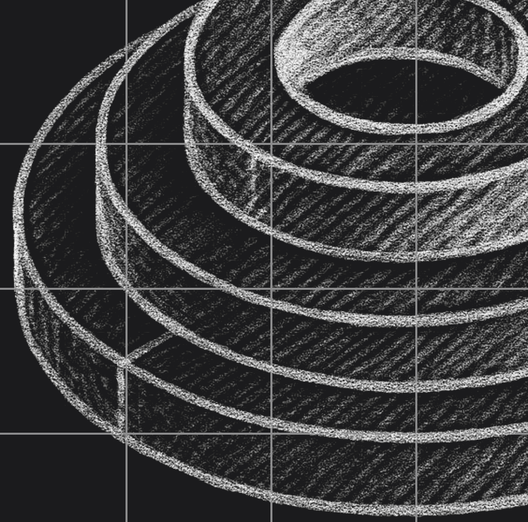
EBOOK

Buyer's Guide to Speech-to-Text APIs

Gladia





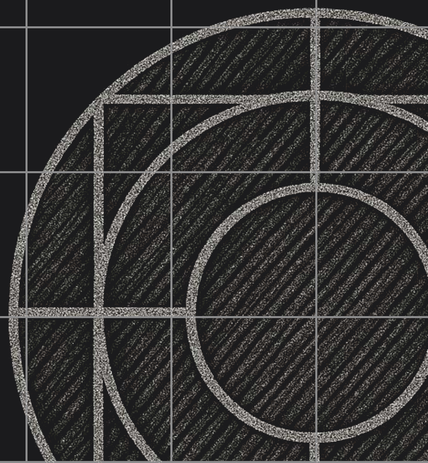


Table of Contents

At Gladia, we've helped hundreds of companies— from fast-moving startups like Attention to enterprises like Aircall —build and scale AI-powered audio features. Along the way, we've seen firsthand how different businesses have vastly different criteria when **choosing a speech-to-text (STT) provider**. Some prioritize real-time transcription with ultra-low latency, while others focus on compliance and multilingual accuracy.

That's why we created this buyer's guide—to **help you navigate the market** with key **evaluation criteria**, **vendor questions**, and **industry insights**. Whether you're integrating STT for customer support agent assist, video transcription, call analytics, or another use case, this guide will help you find a solution that aligns with your goals.

Start Here: A Practical Framework for Scoping STT Projects by Jean-Louis Q.

As someone who's worked with a wide range of companies to integrate speech-to-text—from early-stage startups to global platforms—one of the most common mistakes I see is jumping straight into a vendor search without first clarifying internal objectives. The question isn't just can you build it in-house—but should you, based on your team's focus, infrastructure, and product goals.

That's why I always recommend teams start with this simple framework to define their requirements. It's been used by founders, CTOs, and product leaders across industries to bring clarity to their STT search.

"I want Who **to do** What **in** Environment **using** Device **"**

Jean-Louis Queguiner
Co-founder and CEO at Gladia

Who — Understanding your target users

When thinking about who is using or benefiting from your STT-powered product, it's easy to focus on broad factors like spoken language, geography, dialect, and accent. But inherited attributes (gender, age) and acquired attributes (education, expertise) are equally important. These factors influence STT bias, meaning the provider you choose needs to demonstrate strong performance across your very specific user base.

What — Defining your use case

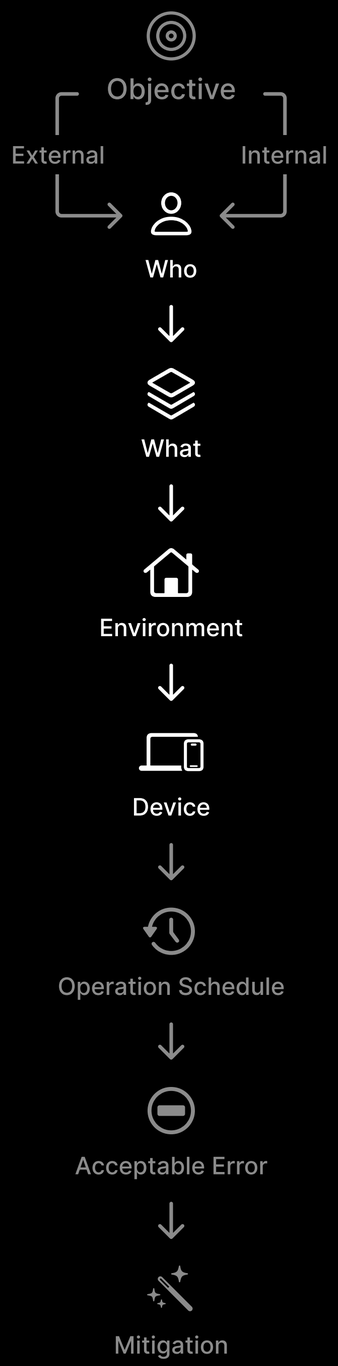
Your use case determines the core STT capabilities you need. Are you transcribing customer service calls in real time? Capturing medical information with high accuracy? Processing podcast audio asynchronously? Mapping this out will help you clarify whether you need real-time or async transcription, as well as establish key quality benchmarks for accuracy, reliability, and processing speed to test against different providers.

Environment — Where your STT will be used

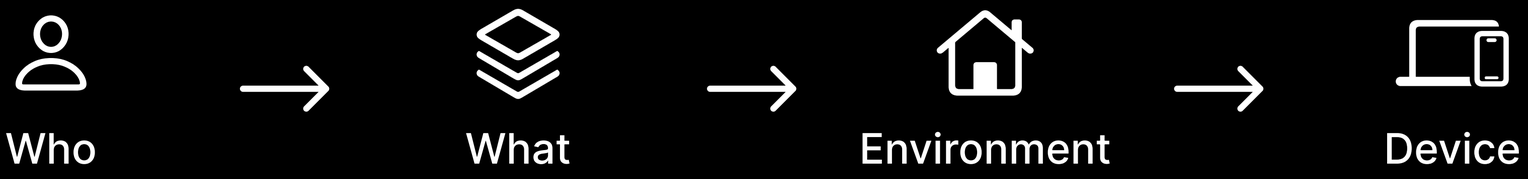
AI performance is heavily impacted by the quality and format of the audio or video it processes. A customer support call in a controlled, headset-based environment is vastly different from a live-streamed conference with multiple speakers and background noise. The environment matters because it determines how well your STT provider needs to filter noise and handle interruptions to deliver accurate transcriptions.

Device — How your AI system will be deployed

The device you deploy STT on impacts latency, streaming capabilities, and error management. Will your AI be running on a cloud-based contact center platform? Used in low-bandwidth environments? Different deployment models introduce different failure modes, and choosing the wrong setup can lead to poor performance, data loss, and user frustration. Your vendor should align with your hosting, connectivity, and real-time processing needs.



Take the time to do this right.



A clear understanding of **what you’re building, who you’re building for, and how it’ll be used** is the foundation for building something that actually works—something your customers will use, love, and come to rely on.

Accuracy

Accuracy is **one of the most important factors** when evaluating a STT provider, especially for applications where transcription quality directly impacts automation or compliance.

If your product depends on **extracting insights from speech**—whether that's in call center analytics, AI-powered agent assist, or media captioning—poor accuracy can result in lost meaning, faulty automation, and frustrated customers.

In other use cases, like media production for example, minor transcription errors generally aren't catastrophic, but **can still be frustrating** for end-users. This is all to say it's really important you consider who you're building your product for and how errors will affect user experience.

How is accuracy measured?

Accuracy is most commonly **measured using Word Error Rate (WER)** which calculates the percentage of words in a transcription that are incorrect compared to a reference transcript.

WORD ERROR RATE FORMULA

$$WER = \frac{(\text{SUBSTITUTIONS} + \text{DELETIONS} + \text{INSERTIONS})}{\text{WORDS IN REFERENCE TEXT}}$$

A **lower WER means higher accuracy**, but it's not a perfect metric. It varies widely depending on the use case, audio complexity, and speaker characteristics.

A STT model that performs well on clean, near-perfect audio might **struggle in real-world conditions** with background noise, accents, or domain-specific terminology. It also doesn't consider the fact that in different contexts, the accuracy of some words is more important than others.

For example, in healthcare, an **incorrect medication name or dosage could have serious consequences**. Even in customer support, failing to capture a name or issue description correctly can lead to poor automation and frustrated users.

Questions to ask yourself

- [ ] For my use case, do minor errors affect usability?
- [ ] Do I need support for industry-specific vocabulary or jargon?
- [ ] How does accuracy impact automation or downstream processing?
- [ ] Does my use case involve multiple speakers, accents, or noisy environments?

Questions to ask vendors

- [ ] What is your WER for general speech vs. domain-specific use cases?
- [ ] Do you support custom vocabulary and fine-tuning?
- [ ] How does accuracy vary across accents, dialects, and languages?
- [ ] How does the STT system perform in noisy environments?
- [ ] What accuracy optimizations are available for streaming vs. batch processing?

What makes Gladia stand out?

Gladia’s next-gen STT model, delivers market-leading accuracy—beating Deepgram and AssemblyAI across languages and audio conditions. Unlike models built for clean data, Gladia is tailored for real-world enterprise use, like call centers with noise, interruptions, and overlapping speech.

Stress-tested across diverse datasets

We benchmark on datasets like Mozilla Common Voice & Google FLEURS for accent and audio diversity. Unlike vendors tied to one benchmark, we test across versions—including real-world data—for true robustness.

Key Data Extraction (KDE)

We go beyond transcription by extracting structured data like names, orgs, and locations. Powered by proprietary ASR systems, this enables fast, accurate capture for CRMs, automation, and insights from any audio.

Adaptability to specific use cases

While you get high accuracy out of the box, you can easily fine-tune Gladia to your terminology, audio environments & user speech patterns. This is especially useful for improving recognition of repeated or unique phrases, without the overhead of building custom models from scratch.

Benchmark-beating accuracy

Gladia reaches a 94% Word Accuracy Rate (WAR) across English, French, Spanish, and other major languages—beating Deepgram’s English-only Nova-3 and AssemblyAI’s latest model.

Voice Activity Detection (VAD) control

Gladia’s speech_threshold (0–1) fine-tunes noise filtering. A 0.999 setting keeps only direct mic input—ideal for noisy call centers. Lower values allow more ambient speech. VAD control adapts to your environment.

Latency

Latency is critical for any use case that involves real-time or near-real-time interaction. If your product relies on instant transcription, live analytics, or conversational AI, latency **can make or break the user experience.**

Traditionally, teams faced a trade-off between speed and accuracy. Real-time meant compromising quality. But that's changing.

Advances in model design and optimization have made real-time speech-to-text truly production-ready, solving the long-standing trade-off between latency and accuracy. With accuracy and latency in balance, **real-time is quickly becoming the future of voice-powered products**.

That said, some use cases like media transcription, quality assurance or call analytics don't depend on instant outputs.

For these asynchronous examples, latency might be less critical, but the speed of transcription still impacts scalability and time-to-value.

For real-time use cases, latency isn't just about speed per stream. It's also about handling many streams at once. If you're transcribing multiple live calls or audio inputs simultaneously, **support for high concurrency** is essential to maintain performance and avoid bottlenecks.

How is latency measured?

Two key metrics matter most when evaluating latency in STT systems: latency on interrupt and latency on final.

Latency on Interrupt—also known as Time to First Byte (TTFB)—measures how quickly the system begins responding after a speaker starts talking. This determines how natural and responsive your speech AI feels.

Latency on Final captures how quickly a complete transcript is delivered after the user finishes speaking. This is crucial for triggering downstream actions, such as passing transcripts to an LLM.

Questions to ask yourself

- [ ] Does my use case require real-time or async transcription?
- [ ] What is the maximum latency my users can tolerate?
- [ ] How will latency affect the user experience and product performance?

Questions to ask vendors

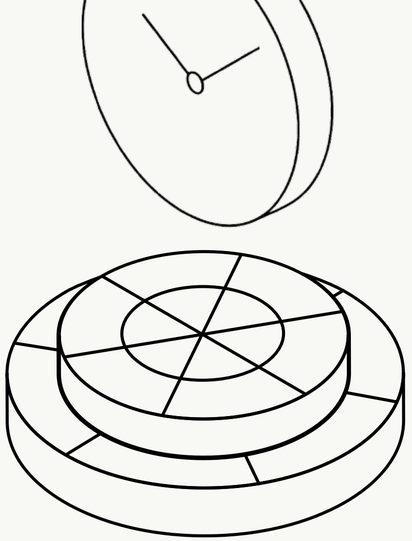
- [ ] What is your average end-to-end latency for real-time and async processing?
- [ ] Do you offer streaming transcription, and how quickly does the first token appear?
- [ ] How does latency change with different languages, accents, or noisy environments?
- [ ] How many concurrent requests does the vendor support?
- [ ] What happens if network issues or connection drops occur?
- [ ] How does latency scale with volume?
- [ ] Can model parameters be adjusted based on specific latency needs?
- [ ] How does your system balance partial vs. final transcript speed?

In real-time use cases like agent assist, live captions, and voice AI, low E2E latency ensures a seamless experience—anything above 300ms can introduce noticeable delays.

For asynchronous (batch) processing, ≈1 min per hour of audio is considered best-in-class, but remember that async use cases tend to prioritize accuracy and cost over speed.

What makes Gladia stand out?

When speed matters, Gladia delivers. It is engineered for real-time use cases where responsiveness is critical—without sacrificing accuracy or functionality.



Industry-leading latency

Gladia achieves an ultra-low 270ms time-to-first-byte (TTFB) and 698ms average latency for final transcription, outperforming competitors by over 500ms in side-by-side testing.

Real-time accuracy without trade-offs

Many providers cut corners on quality to reduce delay—Gladia balances speed and accuracy, even in noisy, multi-speaker environments typical of call centers, live meetings, or AI agent interactions.

Fast async processing too

For non-real-time use cases like media transcription or call analytics, Gladia transcribes 3-minute files in ~30 seconds, and 60-minute files in under a minute, making it ideal for high-volume async workloads.

Language

Language coverage is critical for any STT application that serves multilingual users, global audiences, or industries requiring precise transcription across different dialects and accents.

But it's not just about the number of languages a STT provider supports—it's about how **well the model performs across these languages**, including with regional **accents, dialect variations, and mixed-language speech** (code-switching).

For companies that serve call centers and products that support global enterprise communication, poor language support can lead to inaccurate transcriptions, misinterpretations, or missing data.

Evaluating language coverage

Remember that raw numbers don't tell the full story. A provider might claim to support many languages. But if performance degrades for certain dialects, accents, or mixed-language speech, that makes the API virtually unusable.

"Supported" languages **may actually be unusable for real-world applications**.

To help understand whether a STT system meets the needs of your audience, consider the following:

- Baseline transcription and translation accuracy per language
- Accent and dialect robustness
- Code-switching support
- Multilingual speaker diarization

All of this is influenced by the data the model was trained on. Most STT providers optimize heavily for high-resource languages like English but may struggle with less commonly spoken languages or regional accents.

Questions to ask yourself

- [ ] Do I need high accuracy in one or multiple languages?
- [ ] How important is support for regional accents and dialects in my user base?
- [ ] Will I need speaker diarization (aka separation) for multilingual conversations?
- [ ] Am I targeting a specific country/region where certain languages/dialects dominate?

Questions to ask vendors

- [ ] How many languages are actively optimized?
- [ ] How frequently do you update language models? What improvements do you prioritize?
- [ ] How does accuracy vary by language, dialect & accent? Can you provide benchmarks?
- [ ] Can your STT model recognize & process code-switching within a single conversation?
- [ ] How does your system handle multilingual speaker diarization?
- [ ] Do you offer customization or fine-tuning for specific languages or regional accents?
- [ ] How does language accuracy compare in real-time streaming vs. async transcription?

What makes Gladia stand out?

Gladia offers the broadest and most accurate language support of any STT provider on the market—powering real-time transcription, code-switching, and translation across 100 languages. Whether you're enabling a global workforce or supporting multilingual customers, Gladia makes it possible to serve users fluently, no matter where they are.

True global language support

Gladia supports 100 languages, including 42 underserved languages that aren’t available from other leading STT providers—covering high-population markets like Bangladesh, India, and the Philippines with native-level accuracy.

Fine-tuning by language(s)

For even greater performance, developers can pre-set one or multiple expected languages in a call or conversation, reducing misclassification and speeding up transcription—ideal for contact centers and multilingual products.

Robust code-switching capabilities

Gladia can handle real-time language mixing within a single conversation—essential for international teams, global support centers, and multicultural media content.

Accuracy across dialects & accents

Unlike models that falter outside of English or clean audio, Gladia was trained and evaluated to perform reliably across regional variations and accent-heavy speech.

Features

Many STT providers offer optional features that enhance transcription capabilities beyond basic STT:

Speaker diarization

Differentiates between multiple speakers in a conversation, useful for meetings, interviews, customer service calls, and legal transcriptions.

Abc Custom vocabulary

Allows businesses to improve accuracy for industry-specific jargon, product names, or brand terms.

Sentiment analysis & emotion detection

Helps businesses extract emotional tone and intent from conversations, particularly valuable in customer support and call center analytics.

Timestamps

Automatically aligns text with specific points in the audio or video file, enabling easy searchability, subtitle generation, and content editing.

Are add-ons compulsory?

Add-ons can be essential, optional, or unnecessary depending on your use case. If you're in a high-stakes industry like healthcare, finance, or legal, features like custom model training and speaker diarization might be **non-negotiable**.

Likewise, for customer support and sales analytics, sentiment analysis and emotion detection have become table stakes.

But it's important to note that **enabling add-ons like speaker diarization or emotion detection can impact latency** and increase processing times and therefore may be a trade-off to consider.

This varies by vendor so be sure to evaluate how each vendor **balances feature richness with processing efficiency**.

Questions to ask yourself

- [ ] Do I need speaker diarization, or is single-speaker transcription sufficient?
- [ ] How critical is accurate industry terminology, and will I need custom vocabulary?
- [ ] Would sentiment analysis or emotion detection provide meaningful insights for my business?
- [ ] Is it better to build these features in-house or rely on an external provider?

Questions to ask vendors

- [ ] Is speaker diarization included by default, or does it require additional fees?
- [ ] How does sentiment analysis and emotion detection work—what signals does your model use, and how accurate is it across different speaker tones?
- [ ] What impact do these add-ons have on latency, processing time, and cost?
- [ ] Are there API endpoints or integrations for pulling sentiment and emotion data into our analytics stack?
- [ ] What's on your product roadmap for these features—are there planned improvements, and how often are they updated?

What makes Gladia stand out?

Gladia’s STT platform goes beyond transcription with modular add-ons that help teams extract more value from voice data—whether you’re powering real-time agents, enriching meeting transcripts, or automating workflows. These features are designed to be flexible, developer-friendly, and production-ready out of the box.

Speaker diarization

Accurately differentiate between multiple speakers—even in noisy or overlapping conversations—using a hybrid approach powered by open source models and pyannoteAI. Ideal for interviews, support calls, and multi-party recordings.

Word-level timestamps

Attach precise timing to each word in the transcript—essential for editing tools, searchable content, and automated summaries.

Prosody control

Choose whether to include or remove filler words and verbal tics ("um," "uh," etc.) based on your use case.

Custom vocabulary & spelling

Improve accuracy for brand names, acronyms, and technical terms with custom vocabulary and custom spelling, and ensure proper formatting with casing and punctuation support. Available in both real-time and async modes—unlike many providers that offer it only post-processing.

Sentiment & emotion detection

Analyze tone, intent, and emotional cues async or in real time—especially valuable in contact center analytics and customer experience tools.

Named Entity Recognition (NER)

Automatically identify and label key information such as names, organizations, and locations—critical for structuring data and triggering actions in downstream systems.

Pricing

Most STT providers offer a **pay-as-you-go** model, charging **per minute** or **per hour** of transcribed audio. This is ideal for early-stage startups, developers, or companies with variable usage, as it offers **flexibility without upfront commitments.**

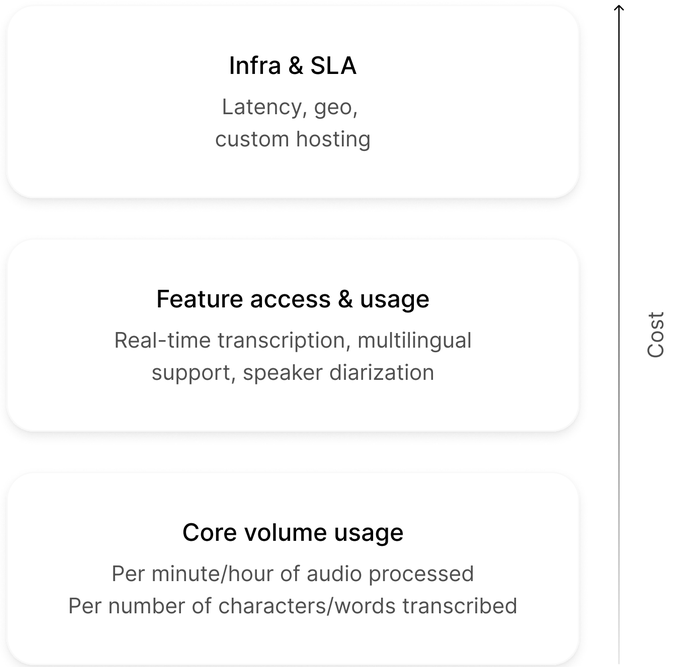
For businesses with **predictable, high-volume needs**, some providers offer **tiered subscription plans** or enterprise pricing with volume discounts, dedicated support, and custom infrastructure options.

Enterprise plans may also include **on-premise deployment or model fine-tuning**, which aren't available in lower tiers.

Feature add-ons, like speaker diarization, may be priced separately due to overheads. Same goes for real-time (streaming) and batch (async) processing, with **real-time STT often carrying a higher per-minute cost** due to the additional infrastructure required.

What are you paying for?

When negotiating a package, understand your vendor's pricing structure and work together to tailor the offer to your project specs and expected ROI.



Questions to ask yourself

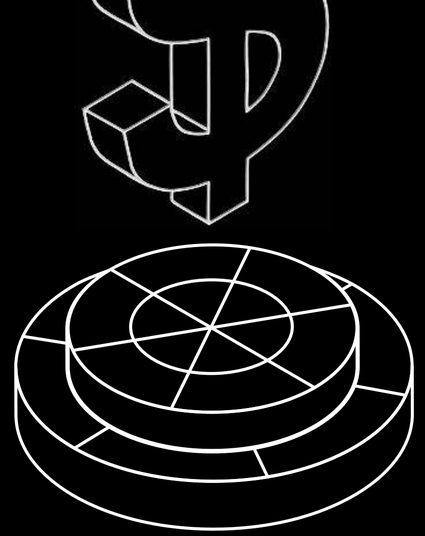
- [ ] How much audio do I expect to transcribe per month, and how might that change as I scale?
- [ ] Would a predictable subscription model or a usage-based approach better align with my business model?

Questions to ask vendors

- [ ] What is the cost per minute/hour for batch vs. real-time transcription?
- [ ] Do you offer volume discounts, and at what thresholds?
- [ ] Are features like speaker diarization, emotion detection, or custom vocabulary included or billed separately?
- [ ] What happens if I exceed my usage limits?
- [ ] Do you offer a free tier or trial credits for testing?

What makes Gladia stand out?

Gladia offers clear, predictable pricing designed to scale with your business so you’re never caught off guard by surprise fees or usage spikes. Unlike many vendors that meter add-ons separately, Gladia keeps your costs transparent and easy to forecast.



Bundled features without surprise charges

Advanced capabilities like speaker diarization, emotion detection, and named entity recognition can be included in your plan. That means you won’t need to track multiple metered line items as you scale.

Predictable costs at scale

Whether you're transcribing hundreds or millions of minutes per month, Gladia’s pricing model helps you plan spend with confidence, ideal for usage-heavy industries like contact centers, media platforms, and virtual meeting tools.

Flexible evaluation & onboarding

You can start in self-serve mode with free trial credits, then scale with technical support and enterprise features when you’re ready.

Integration

Ease of integration and developer experience can **significantly impact time-to-value**. A well-documented API, robust SDKs, and a web-based playground for quick testing allow engineering teams to move faster and reduce development effort.

If integration is **too complex or poorly supported**, even the best STT technology can fail to deliver ROI.

Fortunately, most STT providers offer detailed documentation to support teams through the process.

That means decision-makers should focus on **ensuring the right stakeholders** are involved early and that the integration meets **both technical and business requirements.**

What does “good” look like?

While the fundamentals of integration are similar across providers, the complexity varies based on use case, tech stack, and operational scale.

A CCaaS or CPaaS provider, for example, will have very different requirements than a media production platform or enterprise knowledge management tool.

That said, there are a few core things to look for in a STT API. It should:

- **Be well-documented**, easy to set up, and offer SDKs for common programming languages
- **Offer a web-based playground** to test transcription quality before integration
- **Adapt to different audio formats**, latency requirements, and use cases
- **Be compatible with existing workflows**, whether it’s processing live VoIP calls in a CCaaS platform or syncing transcripts with a CRM for post-call analysis
- Provide **flexible authentication methods**
- Include **robust logging and error-handling** mechanisms to help teams diagnose issues quickly

Questions to ask yourself

- [ ] What systems must this STT API connect with?
- [ ] Do we have experience with real-time APIs, streaming, or AI transcription?
- [ ] How much engineering bandwidth can we realistically allocate to this integration?
- [ ] How does the expected time-to-value align with our product roadmap?

Questions to ask vendors

- [ ] Do you offer a web-based playground or demo environment for testing?
- [ ] What SDKs, libraries & integration docs do you provide?
- [ ] How does authentication work?
- [ ] Are there dependencies or required configurations for different environments?
- [ ] What protocols do you support for audio & input?
- [ ] Does your API support bi-directional streaming?
- [ ] What is the expected integration time, and are there common challenges?
- [ ] Are there any upcoming changes to your API or product roadmap that we should be aware of?

What makes Gladia stand out?

Gladia makes it easy to go from prototype to production with clear documentation and hands-on support every step of the way. Our team doesn't just help you connect the API...we partner with you to build a product your customers will love.

Dedicated engineering support

As a startup, we're able to offer direct, reactive support from our technical team—including product guidance from our CEO and co-founder, who's often in the weeds with customers building for scale.

Clear, developer-friendly documentation

Gladia offers robust SDKs, client libraries, and integration guides, plus a web-based playground to test audio inputs and tweak model parameters before deployment

Fast time-to-value

Teams can spin up integrations quickly thanks to clean docs, smart defaults, and a flexible architecture that supports both real-time streaming and async batch processing.

Workflow compatibility

Whether you're building for CCaaS platforms, syncing transcripts to CRMs, or powering media editing tools, Gladia adapts to your workflow—not the other way around.

Hosting & Scalability

Choosing the right hosting and scalability model for your STT API is key to performance, reliability, & long-term success. Your needs—real-time transcription, batch processing, or strict compliance—will shape cost, flexibility, and usability.

Hosting Options	Overview	Best For...
Public Cloud (Fully Managed)	STT provider handles infra, scaling, and maintenance via AWS, GCP, or Azure — the fastest, easiest setup.	Startups, SaaS platforms, and businesses that need real-time transcription and seamless scalability.
Private Cloud (Customer-Managed)	A dedicated instance offering better data isolation and security, with more control over residency & compliance.	Regulated enterprises needing strong security and compliance, with cloud flexibility.
On-Premise Deployment	Runs on your servers for full control and low latency — but requires in-house deployment and maintenance.	For strict compliance needs (finance, healthcare, government) or full data control.
Hybrid Cloud	Merges on-prem control with cloud scaling — ideal for sensitive data needing managed environments.	Global orgs balancing security, compliance, and performance for large-scale transcription.

Questions to ask yourself

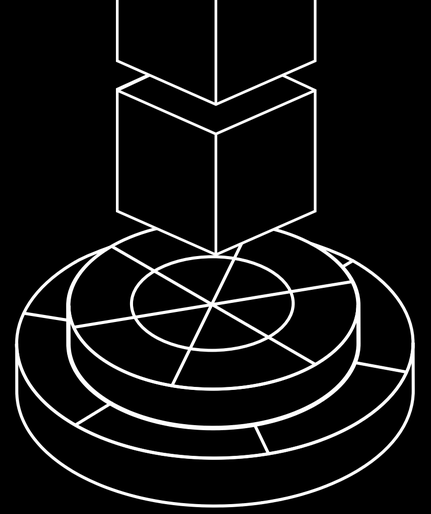
- [ ] How sensitive is the audio data we process?
- [ ] How much transcription volume do we anticipate over time?
- [ ] What internal infrastructure do we have (or lack)?
- [ ] Any regional or industry-specific data compliance requirements?

Questions to ask vendors

- [ ] What hosting options do you provide?
- [ ] If using a cloud-based API, where is data stored and processed?
- [ ] How does the STT API scale? Any rate limits or performance drops?
- [ ] For on-premise deployments, what are the infrastructure requirements?
- [ ] What's the latency gap between cloud and on-premise?
- [ ] Can we migrate between hosting options as our needs change?
- [ ] Do you offer service-level agreements (SLAs) for uptime and scalability?
- [ ] Do you support inference optimization?

What makes Gladia stand out?

Gladia supports global enterprise customers with infrastructure designed for high throughput, low latency, and operational flexibility. Whether you're building real-time voice AI or processing thousands of hours of audio asynchronously, we make it easy to scale without compromising speed or reliability.



Global infrastructure footprint

Choose where your data lives—with hosting available in the U.S., Europe, or APAC, supporting worldwide customer bases and region-specific performance needs.

Enterprise-grade scalability

Gladia supports millions of minutes per month & was purpose-built to handle real-time workloads with bursty traffic, like those found in contact centers and virtual collaboration tools.

Deployment options for any environment

Choose from fully managed cloud, private cloud, or on-premise setups, depending on your team’s architecture and operational preferences.

Seamless transitions as you scale

Need to migrate from cloud to on-prem later? No problem. Gladia supports host model flexibility so your infrastructure can evolve with your business.

Security

Security and compliance are critical considerations when selecting a STT provider, particularly for industries handling **sensitive customer data**, regulated communications, or proprietary information.

Enterprises in **regulated industries**, CCaaS and CPaaS providers, and businesses using STT for internal documentation or knowledge management should be especially vigilant about **security risks and compliance requirements**.

But all organizations should assess **encryption, access control**, and **breach response protocols** to safeguard user data.

Must-have security features for voice data

Security certifications vary **depending on your industry, geography, and risk level**. Here are four key categories to help you evaluate whether a provider meets your organization's requirements

ISO 27001

Broadly valuable across all sectors as a global information security standard.

SOC 2 Type II

Often expected by enterprises to show a provider's operational security and trustworthiness.

GDPR / CCPA

Required if you're handling EU or California consumer data.

Specialized certs

HIPAA (healthcare), PCI DSS (payments), CJIS (law enforcement), FedRAMP (U.S. federal cloud).

Questions to ask yourself

- [ ] What regulatory frameworks do I have to comply with?
- [ ] What types of sensitive data will our STT system process?
- [ ] Who within our organization needs access to transcription data, and how should permissions be managed?

Questions to ask vendors

- [ ] How do you ensure secure API authentication and prevent unauthorized access?
- [ ] What security certifications and compliance frameworks do you support?
- [ ] How is audio data transmitted, stored, and encrypted?
- [ ] Can we control data residency for compliance purposes?
- [ ] What access controls and user permissions are available?
- [ ] What are your policies on data retention and deletion?

What makes Gladia stand out?

When it comes to audio data, security and compliance can’t be optional. That’s why enterprises across finance, healthcare, customer support, and legal tech trust Gladia to handle sensitive voice data with care—without sacrificing performance or control.

Certified compliance

Gladia is compliant with major regulatory frameworks including GDPR, HIPAA, and SOC 2, making it easier for your team to meet internal and external requirements.

Privacy by default

We never train on customer data, ensuring your voice data stays private and won’t be repurposed to improve our models.

End-to-end protection

All audio and transcripts are encrypted in transit and at rest, and we provide robust authentication, access controls, and logging features

Adaptable data policies

Support for region-specific data residency, custom retention, and on-premise deployments lets you tailor data handling to your organization’s needs.

Gladia

Audio infrastructure to transform your business.

Everything starts with reliable transcription.
Learn more at gladia.io.

Talk to an expert

Trusted by 500+ AI assistants and contact center platforms

musixmatch JELLYSMACK method recall VEED
alan TORTUS mojo Spoke Selectra
Brut. Circleback. aircall Attention momentum