Análise Estratégica e Blueprint de Implementação para um Serviço de Transcrição e Tradução Potencializado pela API Gemini

Resumo Executivo

Este relatório apresenta uma análise aprofundada da viabilidade e das vantagens estratégicas da migração de um serviço de legendagem, transcrição e tradução de uma stack auto-hospedada, como a baseada no Whisper da OpenAI, para a API Gemini do Google. A análise valida de forma conclusiva a hipótese de que tal migração é uma decisão estrategicamente sólida, oferecendo uma combinação superior de eficiência de custos, desempenho e funcionalidades avançadas.

A principal conclusão financeira é que a API Gemini, em particular o modelo gemini-2.5-flash-lite, proporciona um custo por hora de processamento de áudio excecionalmente baixo, criando um potencial de margem significativo para um produto SaaS. A estrutura de preços da API, que separa os custos de entrada (ingestão de áudio) e saída (geração de texto), oferece uma granularidade que permite modelos de negócio mais flexíveis e justos em comparação com a precificação tradicional por minuto.

Com base nesta análise, o relatório propõe uma estratégia de preços de três níveis (Gratuito, Pro, Enterprise), alinhada diretamente com a hierarquia de modelos da Gemini. Uma recomendação central é a estruturação do nível gratuito em torno dos limites de taxa da API (Requisições Por Minuto/Tokens Por Minuto) em vez de uma simples quota mensal, garantindo a sustentabilidade do modelo freemium e incentivando a conversão para planos pagos.

Finalmente, o documento fornece um conjunto de recomendações acionáveis para a implementação técnica e o lançamento do produto. A orientação é focar na família de modelos Flash para o lançamento inicial, adotar uma arquitetura serverless para escalabilidade e custo-benefício, e planear a integração futura de capacidades avançadas,

como a Live API, para garantir uma vantagem competitiva a longo prazo.

Secção 1: O Ecossistema de Áudio Gemini: Uma Análise de Capacidades

Esta secção estabelece a base técnica, demonstrando que a API Gemini não é apenas um substituto para tecnologias existentes, mas uma atualização significativa que pode expandir a proposta de valor do serviço.

1.1. Além da Transcrição: Desbloqueando Funcionalidades Avançadas

A migração para a API Gemini representa uma mudança fundamental de um "serviço de transcrição" para uma "plataforma de inteligência de áudio". As capacidades nativas do modelo vão muito além da simples conversão de áudio em texto, permitindo a criação de um produto com múltiplas camadas de valor.

- **Funcionalidade Principal**: A API Gemini oferece transcrição e tradução robustas, acionadas por prompts de texto, para uma vasta gama de formatos de áudio, incluindo MP3, WAV, FLAC e AAC.[1] A requisição central é elegantemente simples: fornecer o ficheiro de áudio e um prompt de texto como "Gere uma transcrição da fala".[1]
- **Diarização de Oradores (Speaker Diarization)**: Uma funcionalidade crítica para inúmeros casos de uso, como a transcrição de reuniões, entrevistas ou depoimentos legais. A Gemini pode identificar e diferenciar nativamente múltiplos oradores no mesmo ficheiro de áudio.[3] Esta capacidade é ativada através de uma engenharia de prompts eficaz. Um prompt como "Pode transcrever esta entrevista, no formato de timecode, orador, legenda. Use Orador A, Orador B, etc. para identificar os oradores" instrui o modelo a gerar uma transcrição estruturada e de fácil leitura.[4] Esta funcionalidade, por si só, eleva o serviço para além da transcrição básica.
- **Sumarização e Análise de Conteúdo**: Sendo um modelo generativo multimodal, a Gemini pode executar tarefas de raciocínio complexo sobre o conteúdo do áudio na mesma chamada de API. É possível solicitar resumos, fazer perguntas específicas sobre o conteúdo ("Qual foi a decisão final tomada na reunião?") ou analisar segmentos específicos usando timestamps.[1] Isto permite a criação de funcionalidades de alto valor, como "resumos automáticos de reuniões", "extração de tópicos chave" ou "identificação de itens de ação".
- **Análise de Timestamps e Segmentos**: A API permite a análise precisa de segmentos de

áudio, referenciando timestamps no formato $MM:SS$ dentro do prompt.[1] Esta funcionalidade é crucial para a criação de transcrições interativas, onde os utilizadores podem clicar numa parte do texto para ouvir o áudio correspondente, ou para focar a análise do modelo em partes específicas de uma longa gravação.

1.2. Vias Arquitetónicas: API Padrão para Processamento em Lote vs. Live API para Aplicações em Tempo Real

A plataforma Gemini oferece duas arquiteturas distintas para o processamento de áudio, cada uma adequada a diferentes casos de uso e roteiros de produtos.

- **Processamento em Lote (API Padrão)**: Esta é a arquitetura primária para o serviço inicial. Os utilizadores carregam um ficheiro de áudio ou vídeo completo [1], e o sistema processa-o de forma assíncrona. Este modelo é ideal para transcrever conteúdos gravados, como reuniões, podcasts, palestras e entrevistas. É a base para a maioria dos serviços de transcrição no mercado.
- **Processamento em Tempo Real (Live API)**: A Live API é uma funcionalidade em preview que permite interações de voz bidirecionais de baixa latência e em tempo real.[8] Processa fluxos contínuos de áudio para fornecer respostas imediatas, criando uma experiência de conversação natural. Embora a sua implementação seja mais complexa, abre caminhos futuros para produtos de alto valor, como legendagem ao vivo para eventos, tradução em tempo real para chamadas ou assistentes de voz com IA. A Live API suporta diferentes arquiteturas de modelo: "Áudio nativo" para a fala mais natural e realista, e "Meia-cascata" para maior fiabilidade em ambientes de produção.[8] A consideração desta API desde o início pode criar uma vantagem competitiva significativa a longo prazo.

1.3. Otimização da Ingestão de Dados: Lidando com Ficheiros de Áudio e Limites da API

Uma arquitetura robusta e escalável depende da gestão eficiente da ingestão de dados.

- **Estratégia de Gestão de Ficheiros**: A API suporta dois métodos para fornecer dados de áudio: passá-los "inline" (codificados em base64) para ficheiros pequenos, ou carregá-los primeiro através da API de Ficheiros para ficheiros maiores. O limiar crítico é um tamanho total de requisição de 20 MB.[1] Para um serviço escalável, a prática recomendada é utilizar sempre uma solução de armazenamento na nuvem (como o

Google Cloud Storage) e fornecer o URI do ficheiro na requisição. Esta abordagem evita limitações de memória e erros de tamanho de requisição, especialmente em ambientes serverless.[3]

- **Formatos e Duração Suportados**: A API suporta uma lista abrangente de formatos de áudio.[1] De forma crítica, modelos como o gemini-2.5-flash podem processar até aproximadamente 8.4 horas de áudio num único prompt, o que é mais do que suficiente para a maioria dos casos de uso.[2] Esta grande janela de contexto é uma vantagem significativa sobre sistemas mais antigos, que muitas vezes exigem que ficheiros longos sejam divididos em pedaços menores antes do processamento.[3]

Secção 2: Modelação Financeira: Uma Análise de Custos Definitiva da API Gemini

Esta secção fornece os dados financeiros centrais solicitados, detalhando os custos com precisão e comparando-os com a concorrência para informar a estratégia de precificação do serviço.

2.1. A Economia dos Tokens: Desconstruindo Como Áudio e Texto se Traduzem em Custos de API

Compreender a mecânica de tokenização é fundamental para calcular os custos com precisão.

- **A Taxa de Conversão Fundamental**: O dado mais crítico para toda esta análise financeira é a forma como o áudio é precificado. Ao contrário do texto, que é tokenizado com base em caracteres ou palavras, o áudio é tokenizado com base na sua duração. A API Gemini utiliza uma taxa fixa: **32 tokens por segundo de áudio de entrada**.[11] Este valor é o pilar de todos os cálculos de custo subsequentes.
- **Tokenização de Texto**: Para a saída (o texto transcrito e traduzido), aplicam-se as regras de tokenização de texto padrão. Um token é aproximadamente equivalente a 4 caracteres ou cerca de 0.75 palavras em inglês.[11] Esta métrica é usada para estimar o custo da geração do texto final.
- **Componentes de Custo**: A faturação baseia-se no número de tokens de entrada (provenientes do áudio) e no número de tokens de saída (provenientes do texto

gerado).[14] A API
countTokens pode ser utilizada *antes* de enviar a requisição principal para obter uma contagem precisa dos tokens de entrada, o que é essencial para a estimativa de custos e para a gestão dos limites da janela de contexto.[15] A chamada a esta API de contagem é gratuita.[11] Esta separação entre os custos de ingestão de áudio e de geração de texto é uma vantagem estratégica. Significa que funcionalidades com saídas de texto mais curtas (como resumos) serão inerentemente mais baratas de fornecer do que transcrições completas, permitindo uma precificação de funcionalidades mais granular e justa.

2.2. Cálculo Central: Benchmark de Custo para Uma Hora de Processamento de Áudio (Transcrição e Tradução)

Esta subsecção detalha o cálculo passo a passo para determinar o custo de processar uma hora de conteúdo de áudio, que servirá de base para a precificação do serviço.

- **Passo 1: Calcular Tokens de Entrada para 1 Hora de Áudio.**
  - 1 hora = 3.600 segundos.
  - Tokens de Entrada = 3.600 segundos×32 tokens/segundo=115.200 tokens.
- **Passo 2: Estimar Tokens de Saída para Transcrição.**
  - Assumindo uma taxa de fala média de 150 palavras por minuto.
  - Total de palavras em 1 hora = 150 palavras/minuto×60 minutos=9.000 palavras.
  - Tokens de Saída de Transcrição Estimados =
    9.000 palavras/0.75 palavras/token≈12.000 tokens.
- **Passo 3: Estimar Tokens de Saída para Tradução.**
  - O texto traduzido terá uma contagem de palavras semelhante.
  - Tokens de Saída de Tradução Estimados ≈12.000 tokens.
  - Total de Tokens de Saída =
    12.000(transcric¸a˜o)+12.000(traduc¸a˜o)=24.000 tokens.
- **Passo 4: Aplicar a Precificação Específica do Modelo.**
  - Utilizando as tabelas de preços [17], o custo total é calculado para cada modelo relevante. Por exemplo, para o
    gemini-2.5-flash-lite:
    - Custo de Entrada = (115.200/1.000.000)×$0.30(prec¸o de entrada de a´udio por 1M de tokens)=$0.03456.
    - Custo de Saída = (24.000/1.000.000)×$0.40(prec¸o de sa ı´da de texto por 1M de tokens)=$0.0096.
    - Custo Total = Custo de Entrada + Custo de Saída = **$0.04416**.

2.3. Tabela Essencial 1: Análise Comparativa de Custos por Hora de Áudio

Esta tabela consolida a análise financeira, fornecendo uma comparação clara e direta que é fundamental para a tomada de decisões de negócio.

Modelo / Serviço	Custo de Entrada (1h de Áudio)	Custo de Saída (Transcrição + Tradução)	Custo Total por Hora	Notas / Diferenciador Chave
Gemini 2.5 Flash-Lite	$0.035	$0.010	**~$0.045**	O mais económico para tarefas de alto volume.
Gemini 2.0 Flash	$0.081	$0.010	**~$0.091**	Modelo anterior, mas ainda muito competitivo em preço.
Gemini 2.5 Flash	$0.115	$0.060	**~$0.175**	Equilíbrio ideal entre velocidade, custo e capacidades.
Lemonfox (Whisper v3)	N/A	N/A	**~$0.17**	Alternativa Whisper de baixo custo.[19]
OpenAI Whisper API	N/A	N/A	**~$0.36**	Benchmark de mercado; Preço de $0.006/minuto

				.[20]

Nota: Os custos da Gemini são calculados com base nos preços por 1M de tokens para entrada de áudio e saída de texto, conforme as fontes.[17] O modelo Gemini 2.5 Pro não está incluído devido à ausência de um preço explícito para entrada de áudio, mas o seu uso seria para um nível premium com um custo significativamente mais elevado, focado na qualidade e em capacidades de raciocínio avançado.

A tabela demonstra inequivocamente a vantagem de custo da família de modelos Gemini Flash, especialmente o gemini-2.5-flash-lite, que é aproximadamente 8 vezes mais barato que a API Whisper da OpenAI para a mesma tarefa. Esta diferença de custo cria uma oportunidade substancial para oferecer preços competitivos aos utilizadores finais, mantendo margens de lucro saudáveis.

Secção 3: Estratégia de Go-to-Market: Arquitetando os Seus Níveis de Preços SaaS

Esta secção traduz a análise financeira numa estratégia de negócio acionável, ajudando a estruturar as ofertas de produtos de uma forma que seja sustentável e alinhada com o valor percebido pelo cliente.

3.1. Alavancando a Hierarquia de Modelos Gemini para Diferenciação de Produtos

A análise da Secção 2 revela uma clara hierarquia de custo-desempenho dentro da família de modelos Gemini: Flash-Lite -> Flash -> Pro. Esta estrutura interna mapeia-se perfeitamente para um modelo de preços SaaS em camadas (tiered), permitindo uma diferenciação clara entre as ofertas.[22] Em vez de tratar todos os modelos como iguais, eles podem ser usados para alimentar diferentes planos, cada um com a sua própria proposta de valor.

- **Nível Básico/Standard**: Potencializado pelo Gemini 2.5 Flash-Lite ou 2.0 Flash. Este nível oferece o serviço principal (transcrição/tradução) com o menor custo possível. É ideal para atrair utilizadores com grandes volumes de conteúdo, estudantes ou startups que são sensíveis ao preço. A sua proposta de valor é a acessibilidade e a eficiência.
- **Nível Pro/Business**: Potencializado pelo Gemini 2.5 Flash. Este nível pode ser comercializado como mais rápido, mais fiável e com limites de taxa mais elevados. É o

"cavalo de batalha" para a maioria dos utilizadores profissionais e pequenas equipas que necessitam de um desempenho consistente para as suas operações diárias.
- **Nível Enterprise/Premium**: Potencializado pelo Gemini 2.5 Pro. Este é o nível para clientes que exigem a máxima precisão, lidam com linguagem técnica ou de domínio específico (médica, jurídica), ou necessitam de funcionalidades avançadas como resumos complexos e análise profunda de conteúdo. O custo mais elevado da API é justificado por um preço premium, direcionado a casos de uso de alto valor.[22]

3.2. Estruturando uma Oferta Freemium Sustentável

A oferta de um plano gratuito é uma ferramenta poderosa para a aquisição de utilizadores, mas deve ser estruturada para evitar canibalizar as receitas.

- **A Natureza do Nível Gratuito da API**: É crucial entender que o nível gratuito da API Gemini não é uma quota mensal de X horas de áudio. É um nível com **limites de taxa**.[18] Por exemplo, o nível gratuito para o
  Gemini 2.5 Pro está limitado a 5 Requisições Por Minuto (RPM) e 25 Requisições Por Dia (RPD).[29] Outros modelos têm limites mais generosos, como 15 RPM para o
  Gemini 2.0 Flash.[27]
- **Implicação para o Serviço**: Isto significa que o plano gratuito do serviço não pode oferecer o processamento ilimitado de ficheiros longos. Um único utilizador a submeter um ficheiro de uma hora ocuparia um slot de processamento por um tempo significativo, potencialmente bloqueando outros utilizadores. Portanto, o nível gratuito deve ser concebido com restrições que espelhem os limites da API:
  - **Limitar a duração máxima do ficheiro** (por exemplo, 5-10 minutos).
  - **Limitar o número de ficheiros** que um utilizador pode submeter por dia.
  - **Implementar um sistema de filas** para gerir requisições concorrentes e evitar exceder os limites de RPM do projeto.
- **Objetivo do Nível Gratuito**: O seu propósito é permitir que os utilizadores testem a qualidade e a velocidade do serviço, funcionando como uma ferramenta de geração de leads para fazer o upsell para um plano pago assim que atingirem os limites de uso.[28] O próprio Google AI Studio serve este propósito de "experimentar antes de comprar" e é totalmente gratuito.[14] A arquitetura da aplicação deve, desde o início, ser capaz de gerir graciosamente os erros de limite de taxa (HTTP 429) e impor um throttling a nível de utilizador que seja ligeiramente mais restritivo que os próprios limites da API, para evitar que o limite global do projeto seja esgotado por alguns utilizadores intensivos no plano gratuito.

3.3. Tabela Essencial 2: Proposta de Níveis de Preços SaaS e Mapeamento de Funcionalidades

Esta tabela fornece um blueprint concreto para a página de preços do produto, ligando cada funcionalidade e limite diretamente às capacidades e custos da API subjacente, garantindo que o modelo de negócio é tecnicamente viável e financeiramente sólido.

Funcionalidade / Limite	**Plano Gratuito**	**Plano Pro**	**Plano Enterprise**
Preço	$0	~$15 / mês	~$40 / mês
Utilizador Alvo	Estudantes, Hobbyistas	Profissionais, Pequenas Equipas	Empresas, Utilizadores Intensivos
Modelo Subjacente	Gemini 2.0 Flash	Gemini 2.5 Flash	Gemini 2.5 Pro
Duração Máx. do Ficheiro	10 minutos	2 horas	8+ horas
Processamento Mensal	1 hora no total	20 horas no total	100+ horas no total
Transcrição Principal	✔️	✔️	✔️ (Máxima Precisão)
Tradução Principal	✔️	✔️	✔️ (Máxima Precisão)
Velocidade de Processamento	Padrão	Rápida	A mais rápida
Diarização de Oradores	❌	✔️	✔️

Resumos com IA	❌	❌	✔️
Acesso à API	❌	✔️ (com limites de taxa)	✔️ (limites mais elevados)

Este modelo de preços não só cria um caminho de upgrade claro para os utilizadores, mas também se alinha com a forma como a própria Google escala o acesso à sua API. A plataforma Gemini oferece níveis de acesso pagos mais elevados (Tier 1, 2, 3) com base nos gastos cumulativos de um projeto no Google Cloud, cada um com limites de RPM/TPM significativamente maiores.[27] Isto cria um mecanismo de escala simbiótico: à medida que o serviço cresce e gasta mais na API, a sua própria capacidade aumenta automaticamente, evitando tetos de desempenho.

Secção 4: Blueprint Técnico e Recomendações de Implementação

Esta secção fornece conselhos acionáveis para construir um serviço robusto, escalável e económico sobre a API Gemini.

4.1. Um Framework de Decisão para a Seleção de Modelos

A escolha do modelo certo é uma decisão estratégica que afeta o custo, a velocidade e a qualidade do serviço.

- **Compromisso Custo vs. Qualidade**: A decisão principal reside entre os modelos Flash e Pro. Para a maioria das tarefas de transcrição padrão, o Gemini 2.5 Flash oferece o melhor equilíbrio entre custo, velocidade e qualidade.[22] Relatos de utilizadores sugerem que o
  Flash é surpreendentemente capaz, mesmo para algumas tarefas de programação, enquanto o Pro se destaca em raciocínio profundo e instruções complexas onde a precisão é primordial.[26]
- **Recomendação**: Lançar com o Gemini 2.5 Flash como o modelo padrão para os níveis pagos e o Gemini 2.0 Flash para o nível gratuito. Oferecer o Gemini 2.5 Pro como uma opção premium de "Alta Precisão" no nível Enterprise. Os modelos mais recentes (série

2.5) são geralmente recomendados em detrimento das versões mais antigas, pois oferecem melhor desempenho e uma estrutura de preços simplificada.[25]

4.2. Melhores Práticas para uma Arquitetura Escalável e Robusta

Uma base técnica sólida é essencial para o sucesso a longo prazo.

- **Gestão Eficiente de Ficheiros**: Conforme estabelecido na Secção 1, não se devem passar ficheiros de áudio grandes "inline". A utilização de URIs do Google Cloud Storage é um requisito para um sistema de produção.[3]
- **Arquitetura Serverless**: Recomenda-se vivamente a utilização de funções serverless (como o Google Cloud Run ou Cloud Functions) acionadas pelo carregamento de ficheiros para um bucket do Cloud Storage. Esta arquitetura é económica (paga-se apenas pelo que se usa) e escala automaticamente para lidar com picos de procura.[3] É importante estar ciente dos tempos limite das funções para ficheiros de áudio longos e aumentá-los em conformidade (até 60 minutos para o Cloud Run).[3]
- **Engenharia de Prompts Avançada**: A qualidade da saída depende fortemente do prompt. A mudança de paradigma em relação a sistemas como o Whisper é que o prompt não é apenas um comando, mas uma ferramenta para controlar o formato da saída. Para tarefas complexas como a diarização com formatação específica, fornecer exemplos "few-shot" dentro do prompt pode guiar o modelo para o resultado desejado.[3] Isto permite a criação de funcionalidades voltadas para o utilizador que personalizam a saída, como "modelos" para notas de podcast, depoimentos legais ou ditados médicos, que são essencialmente prompts pré-configurados e otimizados.
- **Segurança da Chave de API**: As chaves de API nunca devem ser codificadas no código do lado do cliente. Devem ser armazenadas de forma segura como variáveis de ambiente no servidor backend.[6]

4.3. Navegando e Maximizando o Nível Gratuito

O nível gratuito é um ativo valioso tanto para o desenvolvimento como para a aquisição de utilizadores.

- **Desenvolvimento**: O nível gratuito deve ser utilizado extensivamente para desenvolvimento e testes. O Google AI Studio oferece um playground para experimentar com prompts e modelos sem escrever uma única linha de código, o que é inestimável para a prototipagem rápida.[28]

- **Testes de Utilizadores**: Conforme discutido na Secção 3, o plano gratuito voltado para o utilizador deve ser concebido para operar dentro dos limites de taxa da API. Isto implica a aplicação de limites no tamanho/duração dos ficheiros e no número de submissões diárias para garantir um uso justo e prevenir o abuso do sistema.[27]

4.4. Preparando o Serviço para o Futuro: Integrando a Live API

Mesmo que não seja implementada no lançamento, as implicações arquitetónicas da Live API devem ser consideradas.[8] Esta API utiliza WebSockets para comunicação, o que é diferente da API padrão baseada em REST. A implementação pode ser cliente-servidor (para menor latência) ou servidor-servidor.[8] Reconhecer este requisito futuro potencial durante o design inicial do backend tornará a adição de funcionalidades em tempo real muito mais fácil no futuro, protegendo o investimento técnico e posicionando o serviço para capturar casos de uso de maior valor.

Conclusão: Recomendações Estratégicas Finais

Este relatório fornece uma análise abrangente que valida a migração para a API Gemini como uma decisão estratégica fundamental. As conclusões indicam não apenas uma otimização de custos, mas uma oportunidade para transformar um serviço de transcrição numa plataforma de inteligência de áudio mais completa e competitiva.

- **Reafirmação da Estratégia**: A migração para a API Gemini é fortemente recomendada. As poupanças de custos em comparação com a API Whisper são substanciais, e as oportunidades de expansão de funcionalidades são transformadoras para o negócio.
- **Plano de Ação Prioritizado**:
    1. **Estabelecer um Projeto Google Cloud e Chave de API**: Configurar imediatamente a faturação para aceder ao plano pago "Tier 1" e beneficiar de limites de taxa mais elevados para o desenvolvimento.[27]
    2. **Desenvolver uma Prova de Conceito com Gemini 2.5 Flash-Lite**: Focar primeiro no modelo mais económico para validar o fluxo de trabalho principal de transcrição/tradução e estabelecer um custo base.
    3. **Arquitetar para a Escala**: Implementar a arquitetura serverless utilizando o Google Cloud Storage e o Cloud Run desde o primeiro dia, seguindo as melhores práticas delineadas na Secção 4.
    4. **Finalizar o Modelo de Preços**: Utilizar os dados da "Tabela Essencial 1" e a

estrutura da "Tabela Essencial 2" para finalizar os preços de go-to-market.

5. **Lançar e Iterar**: Lançar com o modelo de três níveis. Monitorizar o uso e o feedback dos clientes para determinar se o mapeamento modelo-nível está otimizado ou requer ajustes. Manter a Live API no roteiro como uma funcionalidade chave para uma futura oferta "Enterprise Plus" ou "Tempo Real".

Referências citadas

1. Audio understanding | Gemini API | Google AI for Developers, acessado em agosto 30, 2025, https://ai.google.dev/gemini-api/docs/audio
2. Audio understanding (speech only) | Generative AI on Vertex AI - Google Cloud, acessado em agosto 30, 2025, https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/audio-understanding
3. How partners unlock scalable audio transcription with Gemini ..., acessado em agosto 30, 2025, https://cloud.google.com/blog/topics/partners/how-partners-unlock-scalable-audio-transcription-with-gemini
4. Transcript an audio file with Gemini 1.5 Pro | Generative AI on Vertex AI - Google Cloud, acessado em agosto 30, 2025, https://cloud.google.com/vertex-ai/generative-ai/docs/samples/generativeaionvertexai-gemini-audio-transcription
5. Analyze audio files using the Gemini API | Firebase AI Logic, acessado em agosto 30, 2025, https://firebase.google.com/docs/ai-logic/analyze-audio
6. Building Intelligent Audio Applications with Gemini AI and Google Cloud [Build With AI 2025], acessado em agosto 30, 2025, https://bensonarafat.medium.com/building-intelligent-audio-applications-with-gemini-ai-and-google-cloud-build-with-ai-2025-e919c2120a23
7. Gemini Transcribe, acessado em agosto 30, 2025, https://gemini-transcribe.fly.dev/
8. Get started with Live API | Gemini API | Google AI for Developers, acessado em agosto 30, 2025, https://ai.google.dev/gemini-api/docs/live
9. Live API capabilities guide | Gemini API | Google AI for Developers, acessado em agosto 30, 2025, https://ai.google.dev/gemini-api/docs/live-guide
10. Gemini 2.5 Flash | Generative AI on Vertex AI - Google Cloud, acessado em agosto 30, 2025, https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash
11. Count tokens for Gemini models | Firebase AI Logic - Google, acessado em agosto 30, 2025, https://firebase.google.com/docs/ai-logic/count-tokens
12. Calculating multimodal input tokens - Gemini by Example, acessado em agosto 30, 2025, https://geminibyexample.com/027-calculate-input-tokens/
13. Understand and count tokens | Gemini API | Google AI for Developers, acessado em agosto 30, 2025, https://ai.google.dev/gemini-api/docs/tokens
14. Billing | Gemini API | Google AI for Developers, acessado em agosto 30, 2025, https://ai.google.dev/gemini-api/docs/billing

15. Counting tokens | Gemini API | Google AI for Developers, acessado em agosto 30, 2025, https://ai.google.dev/api/tokens
16. CountTokens API | Generative AI on Vertex AI - Google Cloud, acessado em agosto 30, 2025, https://cloud.google.com/vertex-ai/generative-ai/docs/model-reference/count-tokens
17. Pricing - Google Gemini API, acessado em agosto 30, 2025, https://gemini-api.apidog.io/doc-965864
18. Gemini Developer API Pricing | Gemini API | Google AI for Developers, acessado em agosto 30, 2025, https://ai.google.dev/gemini-api/docs/pricing
19. Whisper API Pricing - Lemonfox.ai, acessado em agosto 30, 2025, https://www.lemonfox.ai/whisper-api
20. OpenAI Whisper Pricing Calculator - InvertedStone, acessado em agosto 30, 2025, https://invertedstone.com/calculators/whisper-pricing
21. Whisper API Pricing and Use Cases | by Ivan Campos | Sopmac Labs - Medium, acessado em agosto 30, 2025, https://medium.com/sopmac-labs/whisper-api-pricing-and-use-cases-6b05ef655015
22. Gemini Pro vs. Gemini Flash: Choosing the Right AI Model - Arsturn, acessado em agosto 30, 2025, https://www.arsturn.com/blog/gemini-pro-vs-gemini-flash-which-ai-model-to-use
23. Gemini 1.5 Flash vs Pro: Which Model Is Right for You? - PromptLayer Blog, acessado em agosto 30, 2025, https://blog.promptlayer.com/an-analysis-of-google-models-gemini-1-5-flash-vs-1-5-pro/
24. Gemini models | Gemini API | Google AI for Developers, acessado em agosto 30, 2025, https://ai.google.dev/gemini-api/docs/models
25. Gemini 2.5 Updates: Flash/Pro GA, SFT, Flash-Lite on Vertex AI ..., acessado em agosto 30, 2025, https://cloud.google.com/blog/products/ai-machine-learning/gemini-2-5-flash-lite-flash-pro-ga-vertex-ai
26. How do you feel Gemini 2.5 Flash differs from Gemini 2.5 Pro? : r/Bard - Reddit, acessado em agosto 30, 2025, https://www.reddit.com/r/Bard/comments/1kdgm7j/how_do_you_feel_gemini_25_flash_differs_from/
27. Rate limits | Gemini API | Google AI for Developers, acessado em agosto 30, 2025, https://ai.google.dev/gemini-api/docs/rate-limits
28. Gemini API: The Free Tier That Makes Developers Happy - DEV Community, acessado em agosto 30, 2025, https://dev.to/garciadiazjaime/gemini-api-the-free-tier-that-makes-developers-happy-28nk
29. Gemini 2.5 Pro Free API Limits: Complete Guide for Developers (2025) - Cursor IDE 博客, acessado em agosto 30, 2025, https://www.cursor-ide.com/blog/gemini-2-5-pro-free-api-limits-guide

30. Google Gemini Pricing Guide: What You Need to Know, acessado em agosto 30, 2025, https://www.cloudeagle.ai/blogs/blogs-google-gemini-pricing-guide
31. How to Use Google Gemini API Key for Free in 2025 - Apidog, acessado em agosto 30, 2025, https://apidog.com/blog/google-gemini-api-key-for-free/
32. Google AI Studio, acessado em agosto 30, 2025, https://aistudio.google.com/
33. Gemini 2.0: Flash, Flash-Lite and Pro - Google Developers Blog, acessado em agosto 30, 2025, https://developers.googleblog.com/en/gemini-2-family-expands/
34. Introduction to Gemini API: Scopes, Challenges and Best Practices | by Vedant Bhamare, acessado em agosto 30, 2025, https://medium.com/@vedantdbhamare/introduction-to-gemini-api-scopes-challenges-and-best-practices-026ebb9c4527
35. Mastering Gemini API For Developers: A Practical Guide - DhiWise, acessado em agosto 30, 2025, https://www.dhiwise.com/post/mastering-gemini-api-for-developers-a-practical-guide