
Retrieval-Augmented Generation (RAG)

Going Beyond the LLM

Thaís de Araújo de Medeiros
thais.araujo.707@ufrn.edu.br

WHO AM I

THAÍS MEDEIROS

Member of the **Conect2ai**
Research Group | Projects:
CNPq and **Rota2030**



**Technician in
Textiles**
IFRN



**Master's Degree and Doctoral
Candidate**
Graduate Program in Electrical and
Computer Engineering
PPgEEC | UFRN



Computer Engineer
UFRN



**Bachelor of Science
and Technology**
UFRN

AGENDA



Introduction to RAG

Architecture & Key
Components

System Building

Hands-On

Which Brazilian actress won the Golden Globe in 2025?



I don't know which Brazilian actress won the Golden Globe in 2025 because my training data goes only up to **June 2024**. Any events that happened after that date, including the 2025 ceremony, are outside of my knowledge. Without searching the web, I cannot provide a confirmed answer.



Which Brazilian actress won the Golden Globe in 2025?

Context:

Brazil woke up to a historic achievement: Fernanda Torres became the first Brazilian actress to win the Golden Globe in 2025. She was awarded for her performance in *I'm Still Here (Ainda Estou Aqui)*, Globoplay's first original film, in the category of Best Actress in a Motion Picture – Drama.

At the ceremony in Hollywood, Walter Salles celebrated the recognition and highlighted the film's impact on Brazilian audiences. Based on the book by Marcelo Rubens Paiva, the feature film portrays the struggle of Eunice Paiva to find her husband, Rubens Paiva, who disappeared during the military dictatorship.



Fernanda Torres won the Golden Globe in 2025, becoming the first Brazilian actress to receive the award. She earned it for her performance in *I'm Still Here (Ainda Estou Aqui)*, Globoplay's first original film, in the category of Best Actress in a Motion Picture – Drama.



Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis^{†‡}, Ethan Perez^{*},

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†],

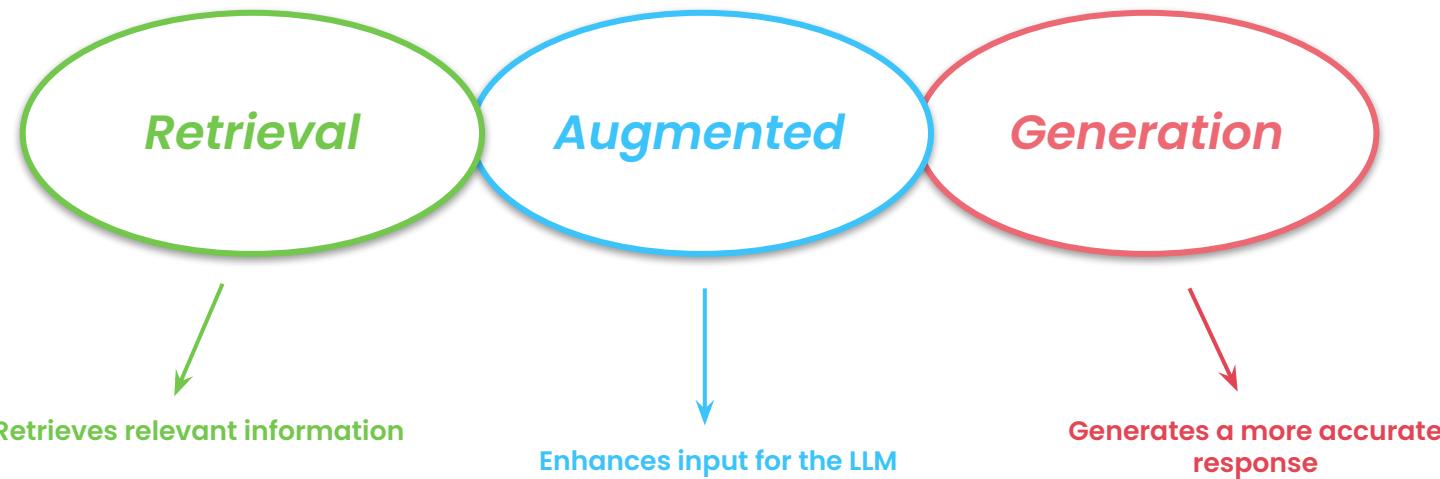
Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel^{†‡}, Sebastian Riedel^{†‡}, Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; ^{*}New York University;
plewis@fb.com

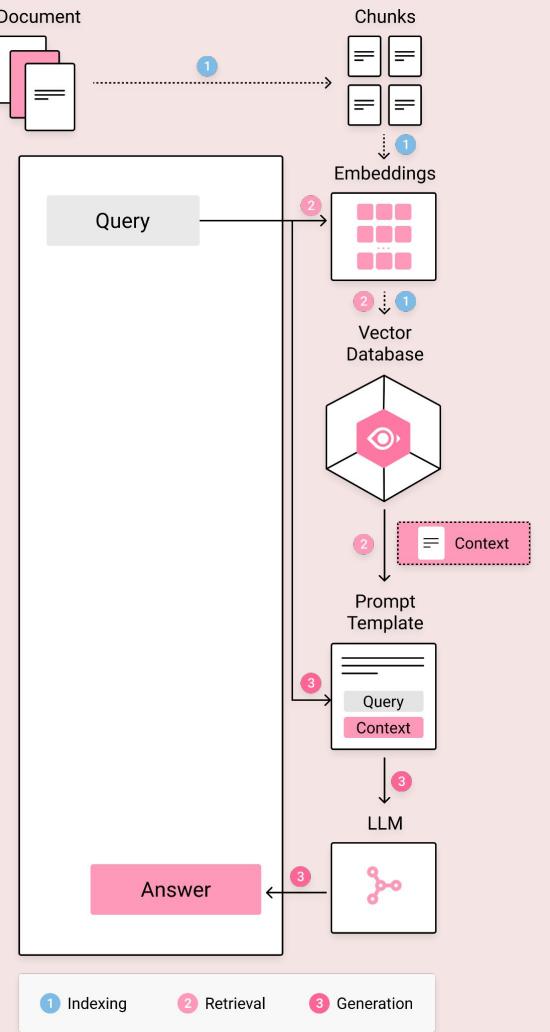
Abstract

Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-art results when fine-tuned on downstream NLP tasks. However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, their performance lags behind task-specific architectures. Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-trained models with a differentiable access mechanism to explicit non-parametric memory have so far been only investigated for extractive downstream tasks. We explore a general-purpose fine-tuning recipe for retrieval-augmented generation (RAG) — models which combine pre-trained parametric and non-parametric memory for language generation. We introduce RAG models where the parametric memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, accessed with a pre-trained neural retriever. We compare two RAG formulations, one which conditions on the same retrieved passages across the whole generated sequence, and another which can use different passages per token. We fine-tune and evaluate our models on a wide range of knowledge-intensive NLP tasks and set the state of the art on three open domain QA tasks, outperforming parametric seq2seq models and task-specific retrieve-and-extract architectures. For language generation tasks, we find that RAG models generate more specific, diverse and factual language than a state-of-the-art parametric-only seq2seq baseline.

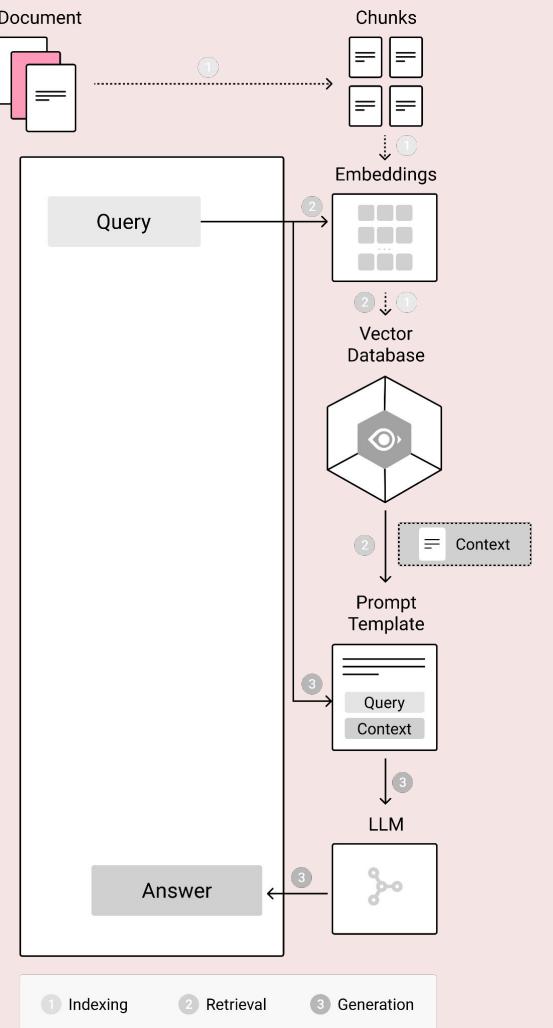
CONCEPT



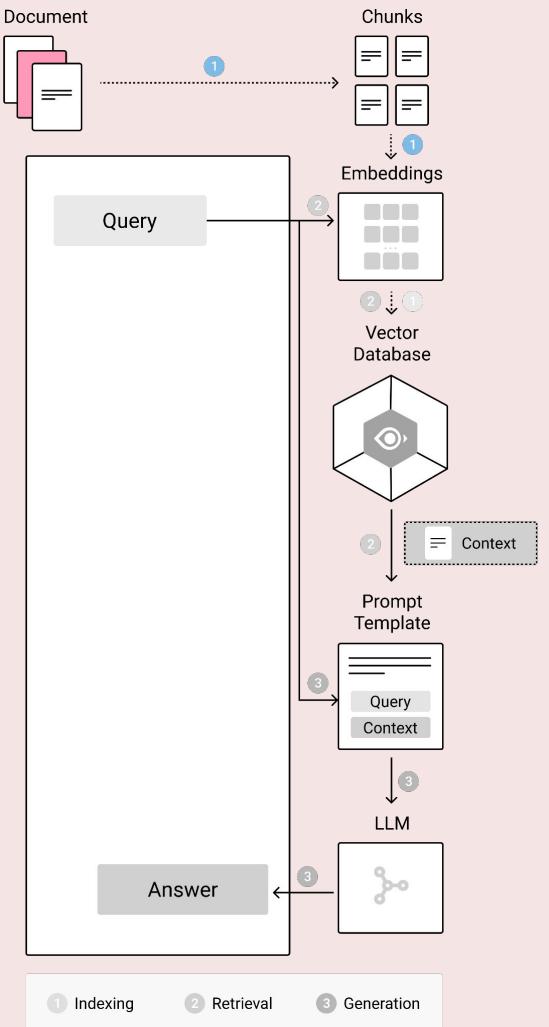
A TYPICAL RAG SETUP



TYPES OF DOCUMENTS



CHUNKING



The air filter is a component responsible for preventing impurities from entering the engine.

chunk 1

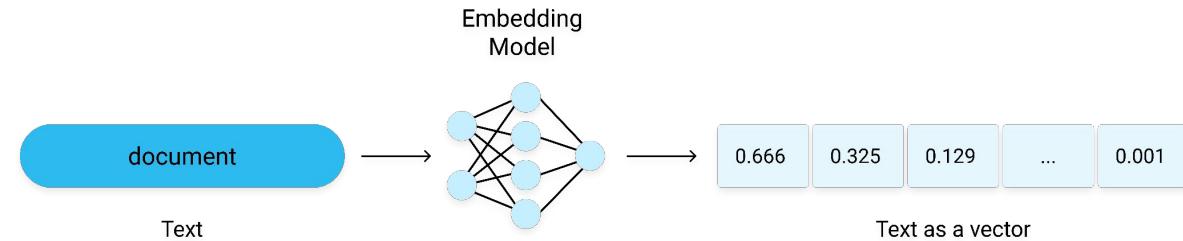
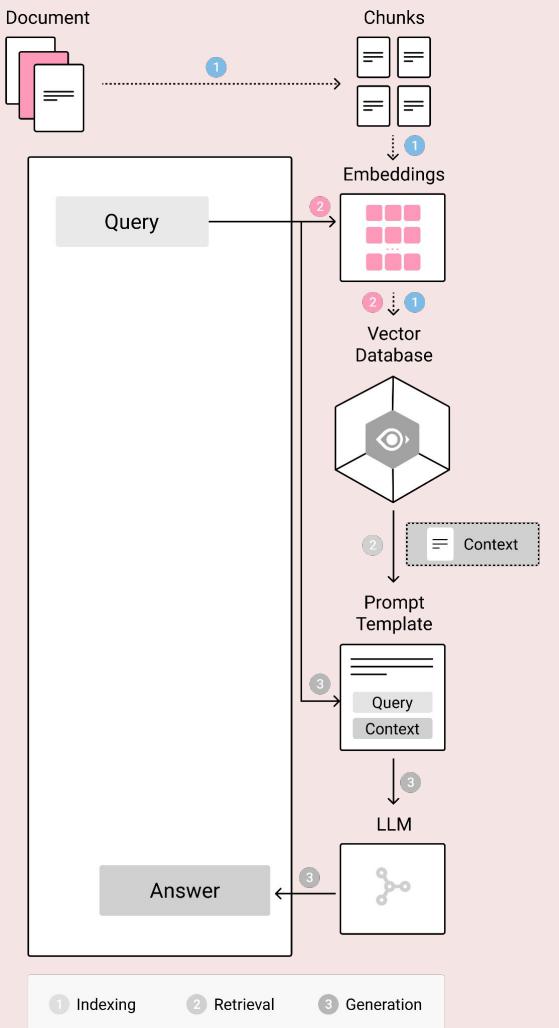
overlap

chunk 2

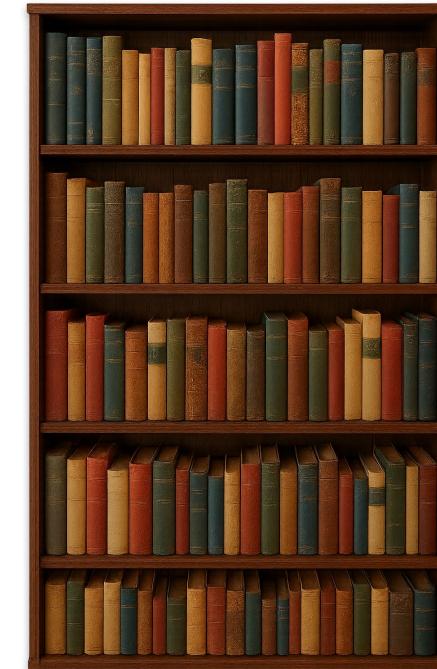
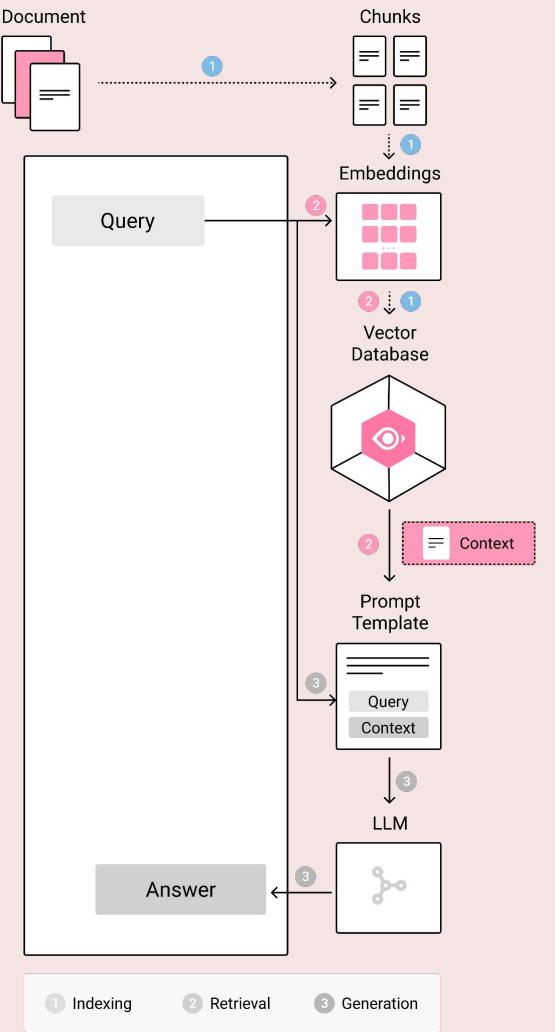
overlap

chunk 3

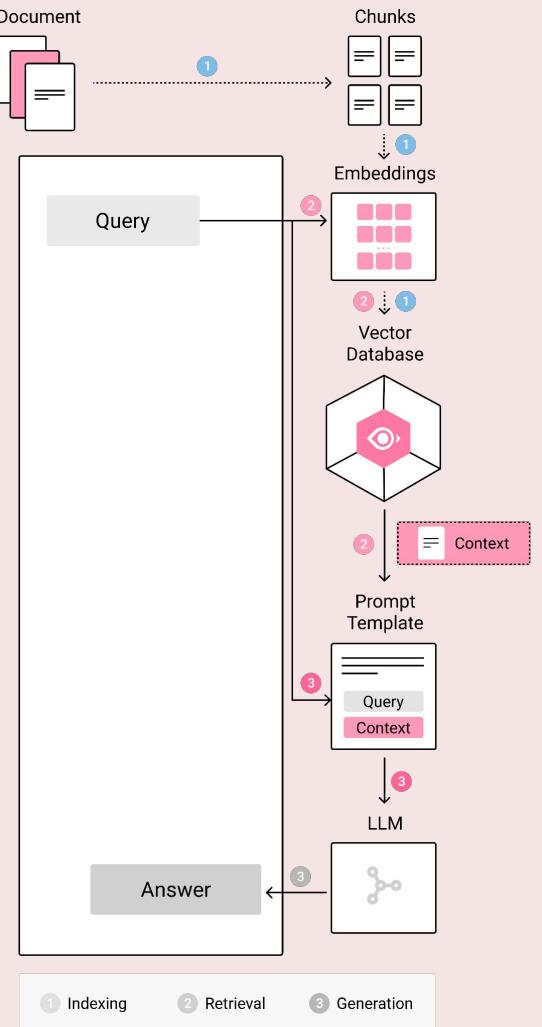
EMBEDDING



VECTOR DATABASE



PROMPT TEMPLATE



You're a friendly customer support. You know this:

<EXTERNAL KNOWLEDGE>



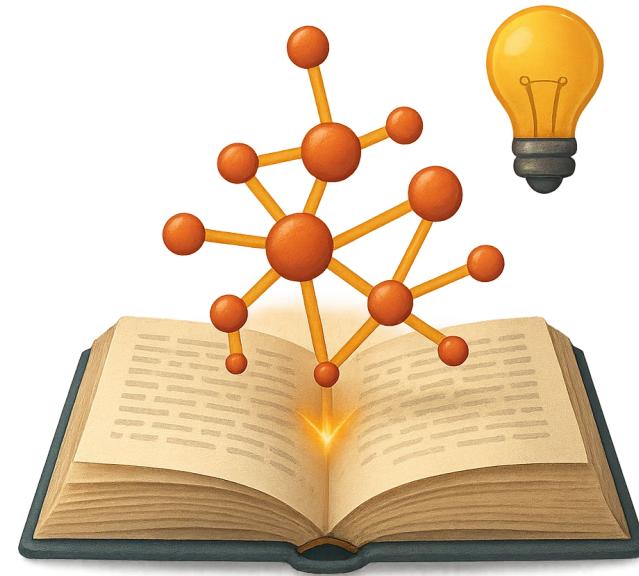
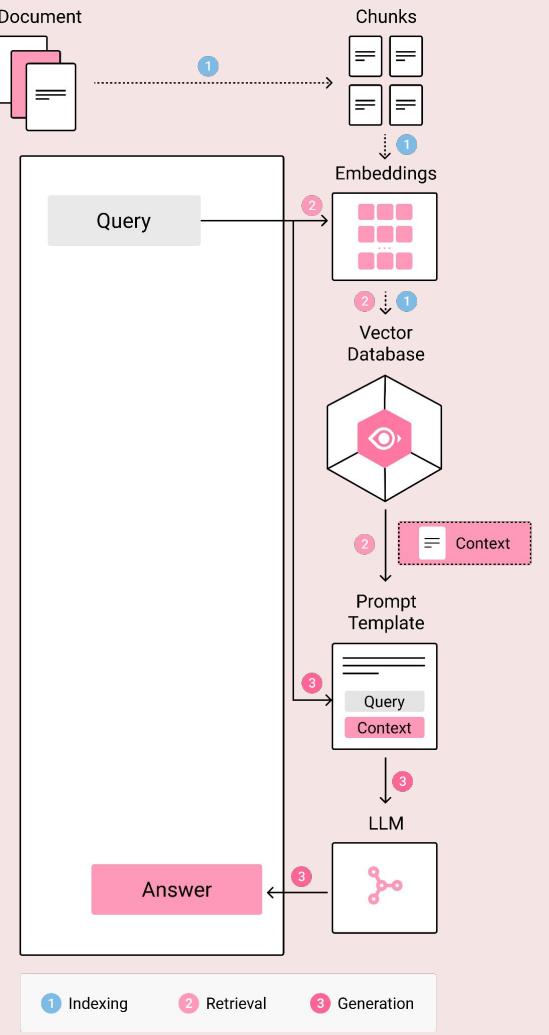
A customer asked this:

<USER QUERY>

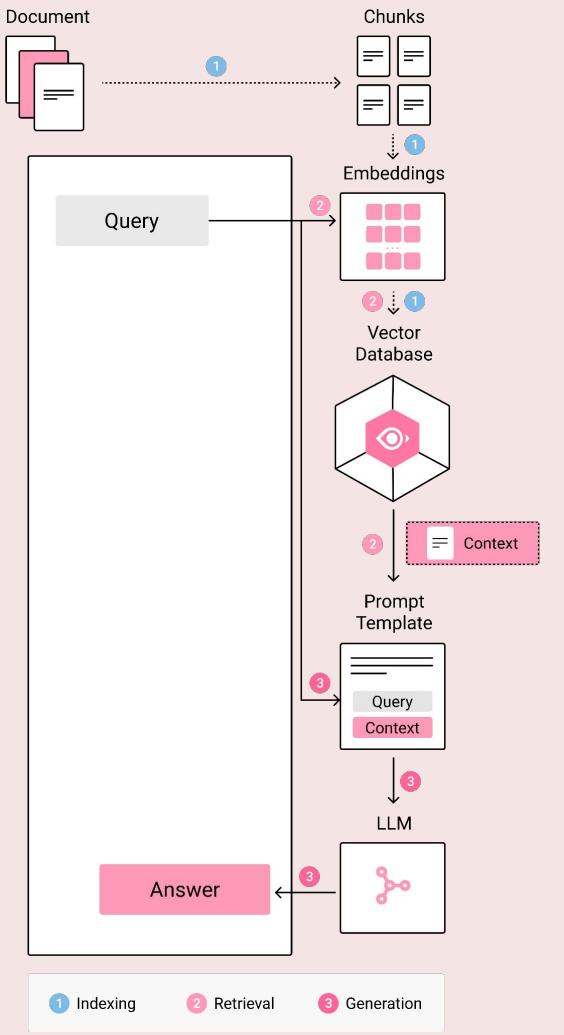


Please generate an answer for the customer.

LARGE LANGUAGE MODEL



LARGE LANGUAGE MODEL



🤗 Hugging Face

🦙 Ollama

OpenAI
Gemini

HOW TO BUILD?

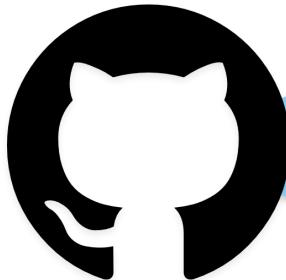


LlamaIndex



LangChain

REPOSITORY



<https://github.com/conect2ai/2025-UNIBS-Courses/>

TUTORIAL

<https://ai.google.dev/>

The screenshot shows the Google AI for Developers homepage. At the top, there's a navigation bar with links for Google AI for Developers, Models, Solutions, Code assistance, Showcase, and Community. A search bar and language selection (Português) are also at the top. A banner at the top of the main content area says "NEW Veo 3 is now available in the Gemini API! >". The main feature is a large white text "AI for every developer" on a dark background. Below it is a subtext: "Unlock AI models to build innovative apps and transform development workflows with tools across platforms." To the right, there's a screenshot of the Google AI Studio interface showing an audio recording segment with a play button and a progress bar (0:00 / 1:49). The interface includes sections for "USER" (instructions to split audio into segments and include transcriptions), "MODELS" (listing Gemini 2.5 Pro, Gemini 2.5 Flash, Gemini 2.0 Flash, Veo 2, and Imagen 3, with Gemini 2.5 Pro selected), and a "Stop" button. A red arrow points from the "Explore models in Google AI Studio" button to the "Stop" button in the screenshot. At the bottom, the text "Start building" is displayed.

Google AI for Developers

Models Solutions Code assistance Showcase Community

Search

Português

NEW Veo 3 is now available in the Gemini API! >

AI for every developer

Unlock AI models to build innovative apps and transform development workflows with tools across platforms.

Explore models in Google AI Studio

User

Split audio recording into segments and include transcriptions

0:00 / 1:49 generative-ai_audio_c...

MODELS

Gemini 2.5 Pro

Gemini 2.5 Flash

Gemini 2.0 Flash

Veo 2

Imagen 3

Running... Stop

Start building

<https://ai.google.dev/>

The screenshot shows the Google AI Studio interface. On the left, there's a sidebar with options like Chat, Stream, Generate media, Build, History, and Enable saving. The main area has a title 'Google AI Studio' and a search bar containing 'Generate Python code for a simple calculator app →'. Below the search bar, there's a 'What's new' section with four cards: 'URL context tool', 'Native speech generation', 'Live audio-to-audio dialog', and 'Native image generation'. A large red arrow points from the top right towards the 'Get API key' button in the top right corner of the main workspace. The top right also includes 'Studio', 'Dashboard', 'Documentation', and a gear icon. The right side of the screen contains various configuration settings for the AI model, such as 'Run settings' (Gemini 2.5 Pro), 'Token count' (0 / 1.048.576), 'Temperature' (set to 1), 'Media resolution' (Default), 'Thinking' (Thinking mode off, Set thinking budget off), 'Tools' (Structured output, Code execution, Function calling, Grounding with Google Search, URL context), and 'Advanced settings'.

Google AI Studio

Get API key

Studio Dashboard Documentation

Run settings

Gemini 2.5 Pro

Token count 0 / 1.048.576

Temperature 1

Media resolution Default

Thinking

Thinking mode

Set thinking budget

Tools

Structured output

Code execution

Function calling

Grounding with Google Search

URL context

Advanced settings

What's new

- URL context tool
- Native speech generation
- Live audio-to-audio dialog
- Native image generation

Google AI models may make mistakes, so double-check outputs.

<https://ai.google.dev/>

The screenshot shows the Google AI Studio interface, specifically the 'Chaves de API' (API Keys) section. The top navigation bar includes 'Get API key', 'Studio', 'Dashboard', 'Documentation', and a gear icon. On the left, there's a sidebar with 'API keys', 'Usage & Billing', and 'Changelog'. The main content area has a heading 'Chaves de API' with a red arrow pointing to a 'Criar chave de API' button. Below it is a 'Testar rapidamente a API Gemini' section with a 'Guia de início rápido da API' and a code editor containing a curl command. A message below the editor says 'Suas chaves de API estão listadas abaixo. Também é possível visualizar e gerenciar seu projeto e as chaves de API no Google Cloud.' A 'Look up API Key for project' button is also present. At the bottom, there's a table header for projects and a note about API key security.

Google AI Studio

Get API key Studio Dashboard Documentation ☰

API keys Usage & Billing Changelog

Chaves de API

+ Criar chave de API

Testar rapidamente a API Gemini

Guia de início rápido da API

```
curl "https://generativelanguage.googleapis.com/v1beta/models/gemini-2.0-flash:generateContent" \
-H 'Content-Type: application/json' \
-H 'X-goog-api-key: GEMINI_API_KEY' \
-X POST \
-d '{
  "contents": [
    {
      "parts": [
        {
          "text": "Explain how AI works in a few words"
        }
      ]
    }
}'
```

Look up API Key for project

Número do projeto	Nome do projeto	Chave de API	Criação	Plano
Crie uma chave de API para ver seus projetos				

View status

Use as chaves de API com segurança. Não as compartilhe nem as incorpore em códigos públicos. O uso da API Gemini em um projeto com faturamento ativado está sujeito aos [preços do modelo de pagamento por uso](#).

Google AI Studio

Get API key Studio Dashboard Documentation ☰

Chaves de API

+ Criar chave de API

Testar rapidamente a API Gemini

Guia de Início rápido da API

Code

```
curl "https://generativelanguage.googleapis.com/v1beta/models/gemini-2.0-f1ash:generateContent" \
-H "Content-Type: application/json" \
-H "X-gcp-api-key: GEMINI_API_KEY" \
-X POST \
-d "{\n  \"contents\": {\n    \"parts\": [\n      {\n        \"text\": \"Explain how AI works\"}\n    ]\n  }\n}"
```

Criar chave de API

Selecione um dos seus projetos atuais do Google Cloud

Search Google Cloud projects

Q

↳ Criar uma chave de API em um projeto atual

Suas chaves de API estão listadas abaixo. Também é possível visualizar e gerenciar seu projeto e as chaves de API no Google Cloud.

Look up API Key for project

Número do projeto	Nome do projeto	Chave de API	criado	Promo

Crie uma chave de API para ver seus projetos

View status

Use as chaves de API com segurança. Não as compartilhe nem as incorpore em códigos públicos. O uso da API Gemini em um projeto com faturamento ativado está sujeito aos [preços do modelo de pagamento por uso](#).

<https://ai.google.dev/>

Google AI Studio

Get API key Studio Dashboard Documentation ☰

Chaves de API

+ Criar chave de API

Testar rapidamente a API Gemini

Guia de Início rápido da API

Code

```
curl "https://generativelanguage.googleapis.com/v1beta/models/gemini-2.8-flash:generateContent" \
-H "Content-Type: application/json" \
-H "X-goog-api-key: GEMINI_API_KEY" \
-X POST \
-d "{\n  \"contents\": {\n    \"parts\": [\n      {\n        \"text\": \"Explain how AI works\"\n      }\n    ]\n  }\n}"
```

Criar chave de API

Selecione um dos seus projetos atuais do Google Cloud

Search Google Cloud projects

Look up API Key for project

Número do projeto	Nome do projeto	Chave de API	criação	Promo
Crie uma chave de API para ver seus projetos				

View status

Use as chaves de API com segurança. Não as compartilhe nem as incorpore em códigos públicos. O uso da API Gemini em um projeto com faturamento ativado está sujeito aos [preços do modelo de pagamento por uso](#).

Google AI Studio

Get API key Studio Dashboard Documentation ☰

Chaves de API

+ Criar chave de API

Testar rapidamente a API Gemini

Guia de Início rápido da API

Code

```
curl "https://generativelanguage.googleapis.com/v1beta/models/gemini-2.8-flash:generateContent" \
-H "Content-Type: application/json" \
-H "X-goog-api-key: GEMINI_API_KEY" \
-X POST \
-d "{\n  \"contents\": {\n    \"parts\": [\n      {\n        \"text\": \"Explain how AI works\"\n      }\n    ]\n  }\n}"
```

Criar chave de API

Selecionar um dos seus projetos atuais do Google Cloud

Search Google Cloud projects

Gemini API (gen-lang-client-0581720549)

↳ Criar uma chave de API em um projeto atual

Suas chaves de API estão listadas abaixo. Também é possível visualizar e gerenciar seu projeto e as chaves de API no Google Cloud.

Look up API Key for project

Número do projeto	Nome do projeto	Chave de API	criação	Promo

Crie uma chave de API para ver seus projetos

View status

Use as chaves de API com segurança. Não as compartilhe nem as incorpore em códigos públicos. O uso da API Gemini em um projeto com faturamento ativado está sujeito aos [preços do modelo de pagamento por uso](#).

<https://ai.google.dev/>

Google AI Studio

Get API key Studio Dashboard Documentation ☰

Chaves de API

+ Criar chave de API

Testar rapidamente a API Gemini

Guia de Início Rápido da API

Code

```
curl "https://generativelanguage.googleapis.com/v1beta/models/gemini-2.0-flash:generateContent" \
-H "Content-Type: application/json" \
-H "X-goog-api-key: GEMINI_API_KEY" \
-X POST \
-d '{ \
  "contents": { \
    "parts": [ \
      { \
        "text": "Olá, mundo!" \
      } \
    ] \
  } \
}'
```

Chave de API gerada

Use suas chaves de API com segurança. Não compartilhe nem as incorpore em códigos que o público possa ver.

AlzaSyAMaONxG9FojqVGFnj9X3T9_pv-vzErmTE 

Suas chaves de API estão listadas abaixo. Também é possível visualizar e gerenciar seu projeto e as chaves de API no Google Cloud.

Look up API Key for project

Número do projeto	Nome do projeto	Chave de API	criação	Promoção
1456	Gemini API ☰	...mTE	25 de ago. de 2025	Configurar o faturamento Ver detalhes de uso ☰

[View status](#)

Use as chaves de API com segurança. Não as compartilhe nem as incorpore em códigos públicos. O uso da API Gemini em um projeto com faturamento ativado está sujeito aos [preços do modelo de pagamento por uso](#).

HANDS-ON