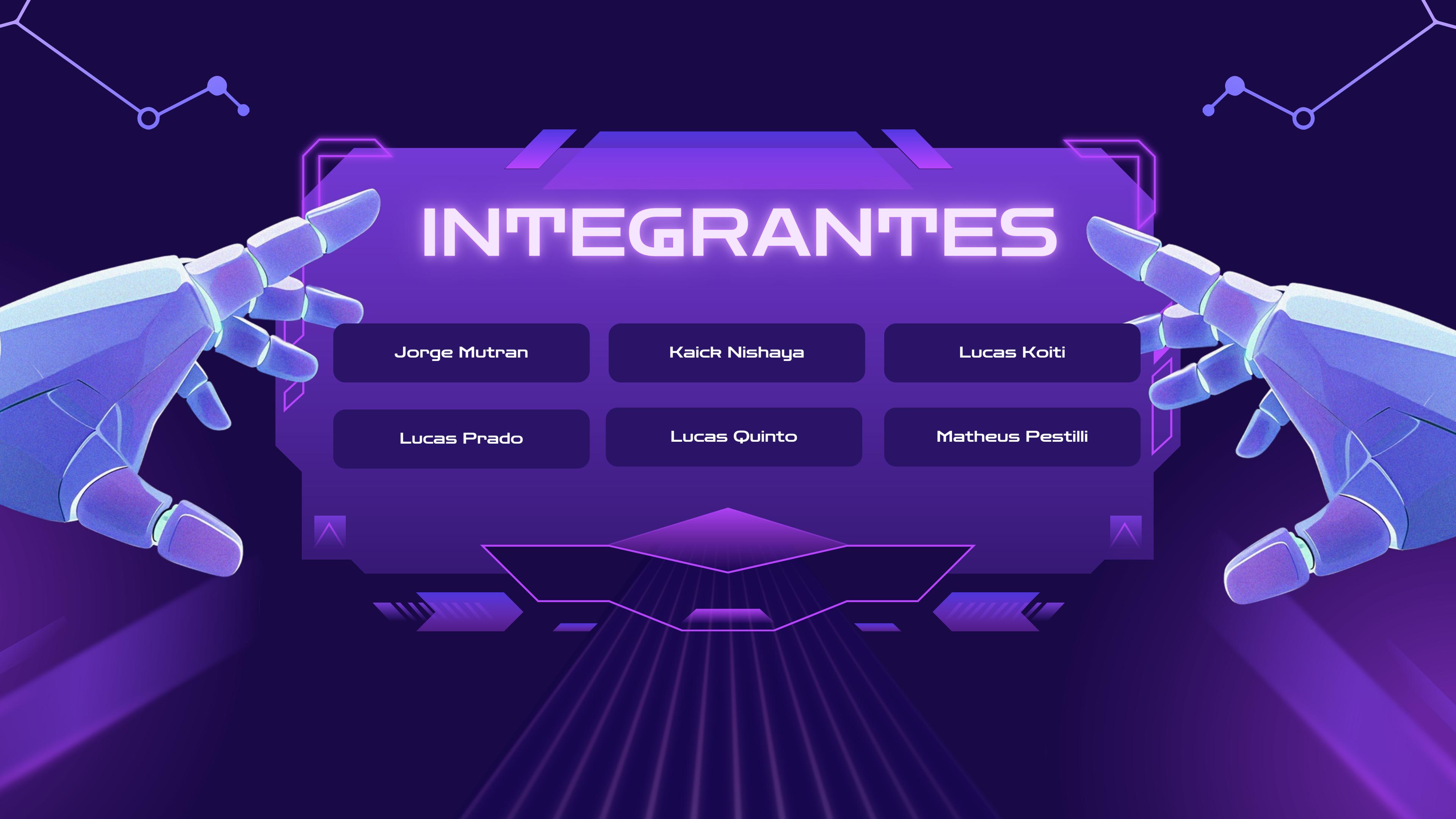


# SISTEMA DE Q&A COM RAG E HUGGING FACE





# INTEGRANTES

Jorge Mutran

Kaick Nishaya

Lucas Koiti

Lucas Prado

Lucas Quinto

Matheus Pestilli



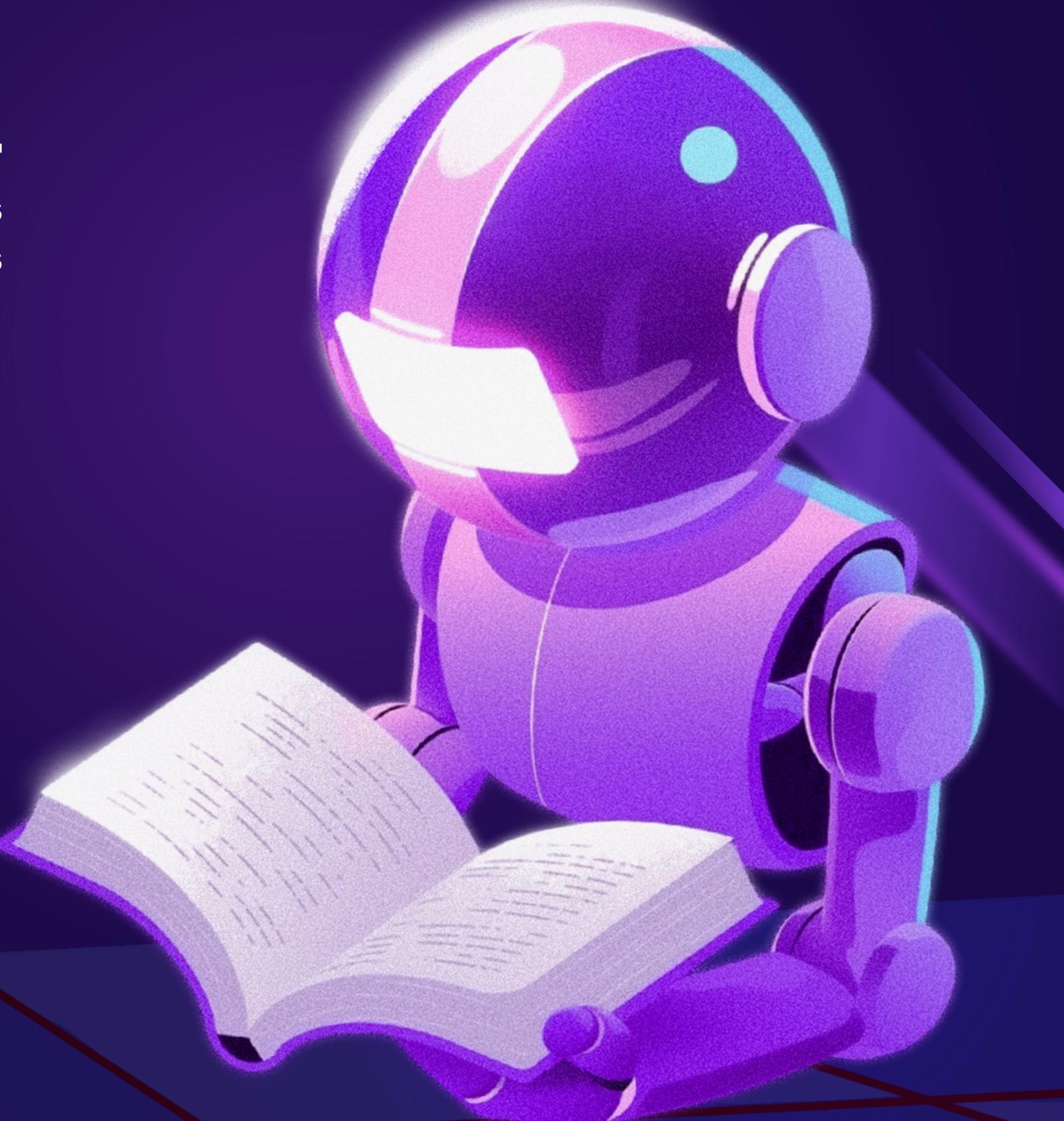
## Θ Problema

LLMs (como o GPT) podem "alucinar" (inventar fatos) e não têm acesso a dados privados ou muito recentes (ex: dados internos de uma empresa).



## A Solução

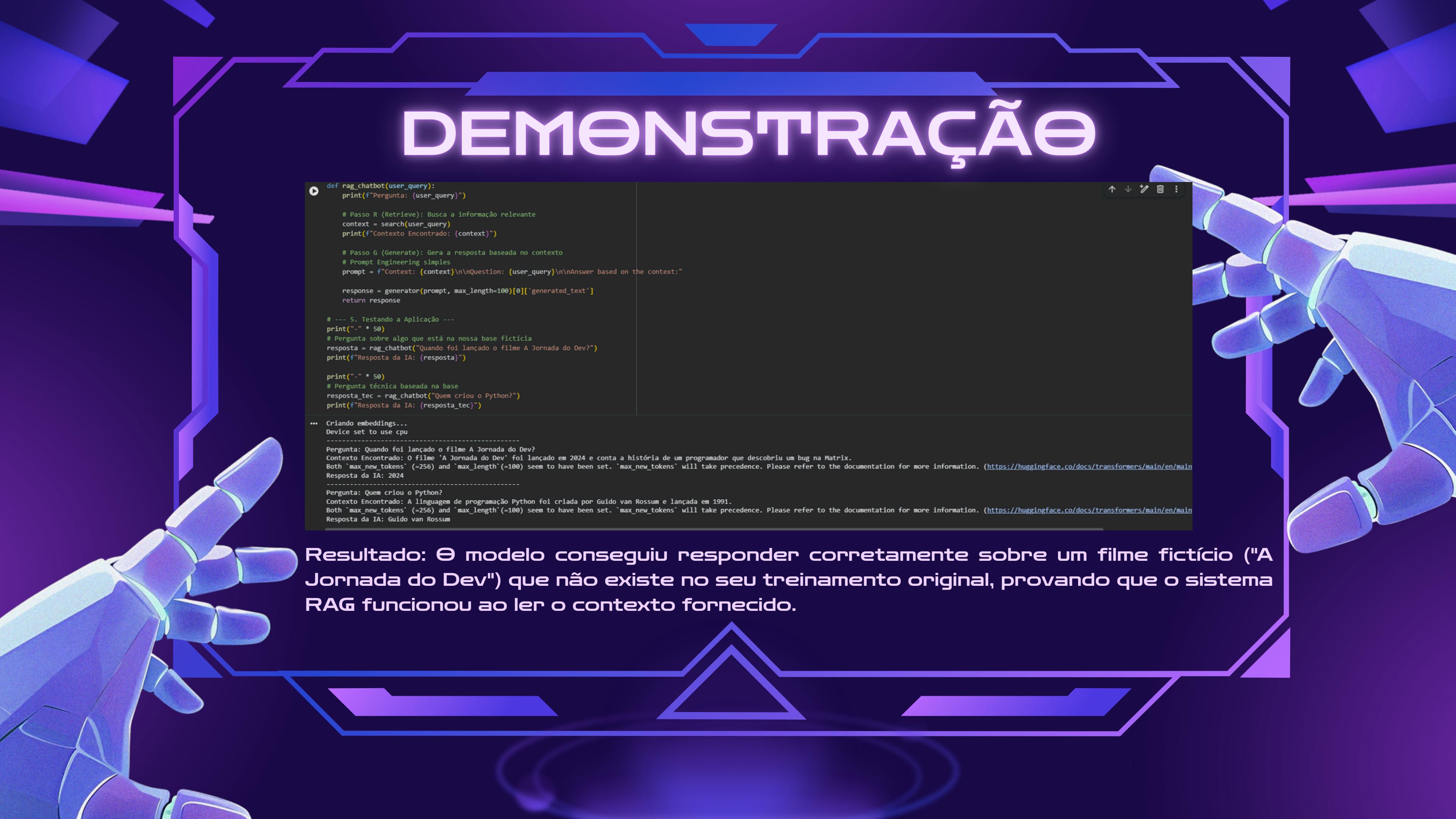
RAG (Retrieval Augmented Generation). Em vez de confiar apenas na memória do modelo, buscamos a informação correta em uma base de dados vetorial e pedimos para a IA formatar a resposta.



# Tecnologias Utilizadas

- 1 Python: Linguagem base.
- 2 Hugging Face sentence-transformers: Usado para criar embeddings (converter texto em vetores numéricos) para busca semântica.
- 3 FAISS (Facebook AI Similarity Search): Banco de dados vetorial rápido para encontrar o contexto mais similar à pergunta.
- 4 Modelo Generativo (google/flan-t5-base): LLM leve do Hugging Face que lê o contexto encontrado e responde à pergunta do usuário.

# DEMONSTRAÇÃO



```
def rag_chatbot(user_query):
    print(f"Pergunta: {user_query}")

    # Passo R (Retrieve): Busca a informação relevante
    context = search(user_query)
    print(f"Contexto Encontrado: {context}")

    # Passo G (Generate): Gera a resposta baseada no contexto
    # Prompt Engineering simples
    prompt = f"Context: {context}\n\nQuestion: {user_query}\n\nAnswer based on the context:"

    response = generator(prompt, max_length=100)[0]['generated_text']
    return response

# --- 5. Testando a Aplicação ---
print("-" * 50)
# Pergunta sobre algo que está na nossa base fictícia
resposta = rag_chatbot("Quando foi lançado o filme A Jornada do Dev?")
print(f"Resposta da IA: {resposta}")

print("-" * 50)
# Pergunta técnica baseada na base
resposta_tec = rag_chatbot("Quem criou o Python?")
print(f"Resposta da IA: {resposta_tec}")

...
Criando embeddings...
Device set to use cpu
-----
Pergunta: Quando foi lançado o filme A Jornada do Dev?
Contexto Encontrado: O filme 'A Jornada do Dev' foi lançado em 2024 e conta a história de um programador que descobriu um bug na Matrix.
Both `max_new_tokens` (=256) and `max_length` (=100) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/main)
Resposta da IA: 2024
-----
Pergunta: Quem criou o Python?
Contexto Encontrado: A linguagem de programação Python foi criada por Guido van Rossum e lançada em 1991.
Both `max_new_tokens` (=256) and `max_length` (=100) seem to have been set. `max_new_tokens` will take precedence. Please refer to the documentation for more information. (https://huggingface.co/docs/transformers/main/en/main)
Resposta da IA: Guido van Rossum
```

**Resultado:** O modelo conseguiu responder corretamente sobre um filme fictício ("A Jornada do Dev") que não existe no seu treinamento original, provando que o sistema RAG funcionou ao ler o contexto fornecido.