INFOX

Tecnologia que transforma

Usamos o melhor da tecnologia
para transformar negócios

www.infoxnet.com.br

INFOX
Tecnologia que transforma

Construindo uma API em Express.js para Extração Estruturada de Dados de PDFs com LangChain, Gemini e RAG

Tiago V. de Arruda

M.Sc. em Ciência da Computação
INFOX

INFOX
Tecnologia que transforma

ExpressJs

Framework WEB minimalista para NodeJs

https://expressjs.com/

INFOX
Tecnologia que transforma

LLMs – Grandes modelos de Linguagem

Modelos de inteligência artificial treinados com grandes quantidades de texto para entender, gerar e responder em linguagem natural.

Eles conseguem conversar, resumir textos, responder perguntas e criar conteúdos de forma parecida com a humana.

INFOX
Tecnologia que transforma

Langchain

Framework de orquestração de código aberto que simplifica a criação de aplicativos com grandes modelos de linguagem (LLMs). Ele fornece ferramentas e componentes para conectar LLMs a várias fontes de dados, permitindo a criação de fluxos de trabalho complexos e de várias etapas.

https://www.langchain.com/
https://js.langchain.com/

INFOX
Tecnologia que transforma

Gemini

Mais recente e avançada família de modelos de inteligência artificial generativa da Google.

Alguns modelos:

- Gemini: geração de texto
- Imagen: geração de imagem
- Veo: geração de vídeo
- Lyria: geração de música

https://ai.google.dev/gemini-api/docs

INFOX
Tecnologia que transforma

RAG - Geração Aumentada por Recuperação

Retrieval-Augmented Generation

Técnica que melhora as respostas de modelos de linguagem ao buscar informações externas relevantes antes de gerar o texto.
Em vez de depender só do que “aprendeu” no treinamento, o LLM consulta uma base de conhecimento e usa esses dados atualizados no momento da resposta, utilizando embeddings e vector store.

INFOX
Tecnologia que transforma

Embedding - Vetorização semântica

É uma forma de transformar textos (ou outros dados) em números que representam seus significados.
No contexto de IA generativa, isso serve para que o modelo entenda e compare o sentido das palavras ou frases, e não apenas as letras.

Por exemplo: as palavras "carro" e "automóvel" terão embeddings muito parecidos, porque significam coisas semelhantes.

INFOX
Tecnologia que transforma

Token

Unidade básica de processamento de texto em modelos de IA.
O modelo não processa palavras inteiras, mas fragmentos menores.

Palavra Simples	**Palavra Complexa**	**Regra Geral**
"gato" = 1 token	"extraordinário" = 2-3 tokens	~4 caracteres = 1 token

Exemplo Prático: A frase "O cachorro correu" pode ser dividida em 4-5 tokens: ["O", " cachorro", " cor", "reu"] (a divisão exata varia por modelo)

- Custos de API são calculados por tokens
- Limites de contexto são medidos em tokens
- Performance do modelo é afetada pelo total de tokens processados

INFOX
Tecnologia que transforma

Janela de Contexto

Quantidade máxima de tokens que o modelo consegue processar simultaneamente. É a "memória de trabalho" do modelo.

GPT-3.5	4.096 tokens (~3.000 palavras)
GPT-4	8.192 - 32.768 tokens
Claude Somnet	200.000 tokens (~150.000 palavras)
Gemini 1.5	Até 1.000.000 tokens

Composta por:

- Prompt do usuário
- Histórico da conversa
- Instruções do sistema
- Resposta sendo gerada

INFOX
Tecnologia que transforma

Engenharia de Prompt

Arte e ciência de formular instruções eficazes para obter melhores resultados dos modelos de IA.

Prompt Ruim

"Fale sobre python"

vago, sem contexto = resultado genérico

Prompt Bom

"Explique os conceitos de decorators em Python para um desenvolvedor Java, com 3 exemplos práticos de uso"

Específico, contextualizado = resultado focado e útil

Técnicas:

- Seja específico e detalhado no pedido
- Forneça contexto e exemplos
- Especifique formato desejado da resposta

INFOX
Tecnologia que transforma

Temperatura e Criatividade

Parâmetro que controla a aleatoriedade e criatividade das respostas do modelo (valores de 0 a 2).

Temperatura Baixa (0 a 0.3)

- Previsível
- Determinístico
- Focado
- Consistente

Temperatura Média (0.5 a 0.7)

- Balanceado
- Natural
- Versátil
- Padrão

Temperatura Alta (0.8 a 2.0)

- Criativo
- Variado
- Experimental
- Imprevisível

Técnicas:

- Temperatura baixa para: respostas factuais e traduções precisas
- Temperatura alta para: escrita criativa, geração de ideias, conteúdo artístico

INFOX
Tecnologia que transforma

Hands-On

<u>Dependências</u>

API
- npm i express nodemon dotenv

Lanchaing + Gemini
- npm i langchain @langchain/core @langchain/google-genai

Upload/leitura de PDF
- npm i pdf-parse
- npm i multer

MUITO OBRIGADO!

Tiago V. Arruda

Analista de Sistemas

tiago.arruda@infoxnet.com.br

www.infoxnet.com.br