

THIAGO MEMELLI

Engenheiro de IA | Python | Sistemas LLM | Engenharia de Backend

Vitória, ES - Brasil | tmemelli@gmail.com | thiagomemelli.com.br
linkedin.com/in/thiagomemelli | github.com/tmemelli

PERFIL PROFISSIONAL

Engenheiro de software com mais de 13 anos de experiência em desenvolvimento, sustentação e evolução de sistemas corporativos, incluindo ambientes de missão crítica e operação 24/7.

Atualmente focado em Engenharia de IA, desenvolvendo aplicações e infraestrutura para sistemas baseados em LLMs com Python, FastAPI, processamento assíncrono e integração com múltiplos provedores de inteligência artificial.

Experiência prática recente na construção de APIs de inferência resilientes, concorrência assíncrona, retries, fallbacks entre modelos e provedores, rate limiting, tratamento de falhas, contratos tipados, testes automatizados e publicação de aplicações em ambiente real.

Interesse especial em Engenharia de LLM, RAG, embeddings, busca vetorial e construção de sistemas de IA preparados para produção.

PROJETO EM DESTAQUE

LLM Inference Service

Engenharia de IA / Infraestrutura de LLM

Demo: app.thiagomemelli.com.br | API: api.thiagomemelli.com.br/docs

Serviço assíncrono de inferência desenvolvido em Python e FastAPI para abstrair e tornar resiliente o acesso a múltiplos provedores de LLM.

- Integração com múltiplos provedores: Groq e Gemini
- Controle de concorrência por provedor
- Backoff exponencial com jitter
- Fallback de modelo
- Proteção contra ciclos de fallback
- Rate limiting por cliente
- Métricas de tokens, provedor, modelo e latência
- API pública documentada com OpenAPI / Swagger
- 44 testes automatizados
- Processamento assíncrono com Asyncio
- Retry para falhas transitórias
- Timeout individual por provedor
- Fallback de provedor
- Classificação e normalização de erros dos SDKs
- Rastreamento por request ID
- Validação de dados e configurações com Pydantic
- Frontend público integrado à API
- Validação estática e qualidade com mypy e Ruff

Tecnologias: Python • FastAPI • Asyncio • Pydantic • Groq API • Gemini API • Pytest • mypy • Ruff • APIs REST • OpenAPI

OUTROS PROJETOS

CashFlow API

API financeira desenvolvida com Python e FastAPI para gerenciamento de receitas, despesas, categorias e relatórios. Inclui assistente de IA integrado à OpenAI para responder perguntas utilizando os dados financeiros do usuário.

- API REST com FastAPI
- Integração com OpenAI API
- Autenticação JWT
- SQLAlchemy ORM
- Validação com Pydantic
- Docker
- 67 testes automatizados com Pytest
- Documentação OpenAPI / Swagger

Tecnologias: Python • FastAPI • OpenAI API • SQLAlchemy • Pydantic • Pytest • Docker • JWT

Sales Insights Pipeline

Pipeline automatizado de processamento e análise de dados de vendas desenvolvido em Python. Realiza validação e limpeza de arquivos CSV, cálculo de indicadores financeiros, identificação de outliers, geração automática de insights e criação de relatórios e visualizações. Benchmark documentado de aproximadamente 8,5 segundos para processamento de 100 mil registros.

Tecnologias: Python • Pandas • NumPy • Matplotlib • Seaborn • POO • Processamento de Dados

COMPETÊNCIAS TÉCNICAS

Engenharia de IA: APIs de LLM • Sistemas LLM com múltiplos provedores • Inferência de LLM • Engenharia de Prompts • Estratégias de Retry e Fallback • Tratamento de Erros • Rate Limiting

Python e Backend: Python • FastAPI • Asyncio • Pydantic • SQLAlchemy • APIs REST • OpenAPI • JWT

Arquitetura e Qualidade: POO • Type Hints • Protocols • Dataclasses • Pytest • mypy • Ruff • Concorrência Assíncrona

Dados: SQL • SQLite • SQL Server • Modelagem de Dados

Ferramentas e Ambiente: Linux • Git • GitHub • Docker • VS Code

Em aprofundamento: RAG • Embeddings • Busca Vetorial • Bancos de Dados Vetoriais • Redis

EXPERIÊNCIA PROFISSIONAL

Engenharia de IA / Backend Python - Projetos Independentes

Agosto 2025 - Presente

Desenvolvimento de projetos voltados à modernização da stack técnica e especialização em Engenharia de IA.

- Desenvolvimento e publicação de serviço de inferência para LLMs com múltiplos provedores
- Construção de APIs assíncronas com Python e FastAPI
- Implementação de mecanismos de retry, fallback, timeout, concorrência e rate limiting
- Integração com APIs de modelos de linguagem
- Desenvolvimento orientado a contratos tipados e separação de responsabilidades
- Criação de testes automatizados e validação estática de código
- Desenvolvimento e publicação de aplicações acessíveis pela internet
- Construção de projetos práticos voltados a Backend, IA e processamento de dados

Experiência Internacional e Hospitalidade

Junho 2018 - Julho 2025

Mudei-me para a Europa e trabalhei no setor de hospitalidade, atuando em atendimento de restaurantes e funções de cozinha. Trabalhei como garçom e cozinheiro, desenvolvendo forte orientação ao cliente, trabalho em equipe, adaptabilidade, comunicação em ambientes multiculturais e capacidade de atuar sob pressão em operações de ritmo acelerado.

Em 2025, retornei à engenharia de software com foco na transição para Engenharia de IA, IA Generativa e sistemas backend modernos em Python.

Analista de Sistemas Sênior

SPG Negócios de TI e Telecom | Maio 2006 - Junho 2018 | Vitória, ES

Atuação em sistemas de alta criticidade e operação 24/7 para grandes empresas do setor de telecomunicações, incluindo Oi e Brasil Telecom.

- Sustentação e evolução de sistemas corporativos críticos
- Análise e resolução de incidentes em produção
- Desenvolvimento e manutenção de sistemas
- Integração entre plataformas e sistemas
- Investigação de problemas de performance
- Atuação com bancos de dados e processamento de grandes volumes de informações
- Participação em análise técnica e definição de soluções

Ambiente: Natural/Adabas • SQL Server • SOA • Sistemas Corporativos

Desenvolvedor de Sistemas

InfoSis Consultoria | Março 2005 - Março 2006 | Vitória, ES

- Desenvolvimento de aplicações
- Análise de sistemas
- Desenvolvimento com Delphi e SQL Server
- Modelagem de dados
- Levantamento de requisitos junto a clientes

FORMAÇÃO

MBA em Gerenciamento de Projetos - Fundação Getulio Vargas (FGV), 2009 - 2011

Bacharelado em Ciência da Computação - FAESA, 2002 - 2005

IDIOMAS

Português: Nativo

Inglês: Intermediário - Boa leitura de documentação técnica e experiência trabalhando com ferramentas, APIs e documentação em inglês.

ÁREAS DE INTERESSE

Engenharia de IA • Engenharia de LLM • Sistemas RAG • IA Generativa • Backend Python • Infraestrutura de IA

Modalidade: Remoto • Presencial • Híbrido