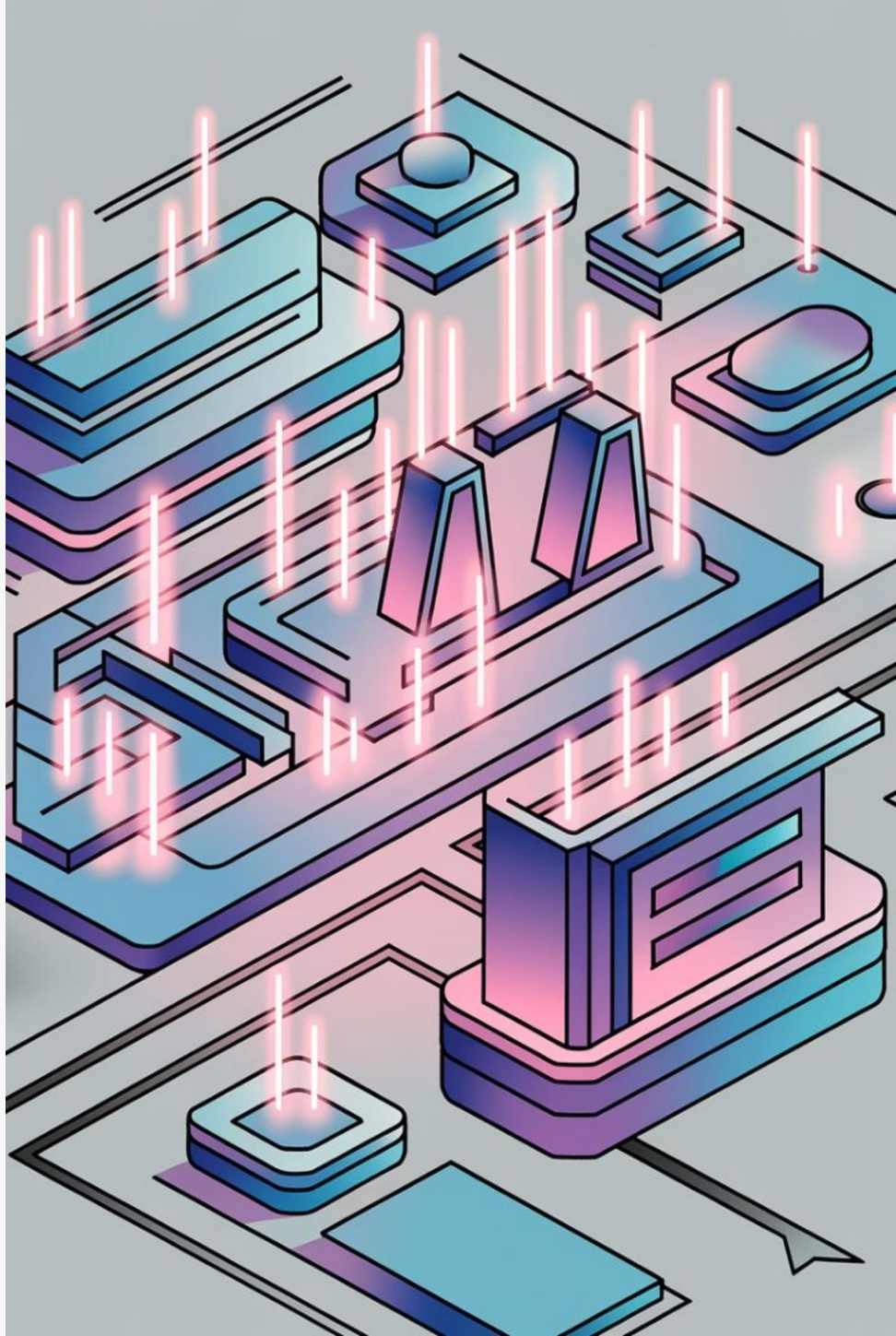
Arquitetura de Software para IA Generativa

A IA Generativa vai muito além do chat. Ela expande sistemas com funcionalidades antes impossíveis ou caras.

Vamos explorar como integrar IA de forma segura e eficiente nos seus produtos.

por Marcelo Pinheiro



Sobre Mim

Marcelo Pinheiro

Formação

Análise de Sistemas com pós-graduação em Arquitetura de Soluções.

Profissional

Gerente de IA e Arquiteto de Soluções na Levva.

Pessoal

Pai da Elena (2 anos), tutor do Luci (gato preto) e casado com Carol.

Além da Telinha de Chat

Filtragem de Currículos

IA que seleciona candidatos baseada em critérios específicos.

Padronização de Dados

Transformação de diversos formatos de planilhas em padrão único.

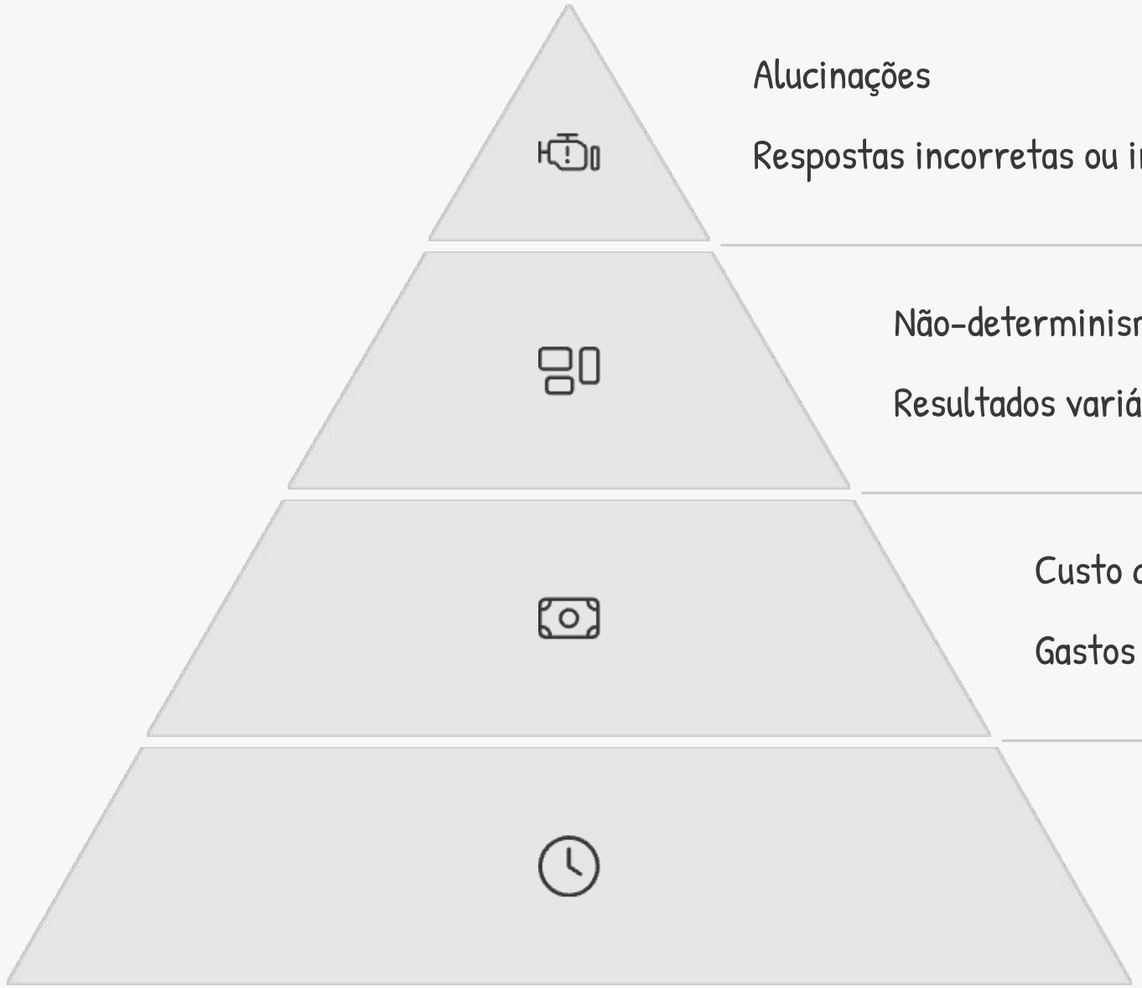
Qualificação de Leads

Análise automática de potenciais clientes no CRM.

Sugestão de Ações

Recomendação da próxima melhor ação para vendedores.

Desafios da IA Generativa



Alucinações

Respostas incorretas ou inventadas

Não-determinismo

Resultados variáveis para mesmos inputs

Custo de tokens

Gastos com processamento de texto

Performance

Tempo de resposta mais lento

Consumo de IAs Generativas

APIs Comerciais

OpenAI, Anthropic, Cohere, etc.

Provedores Cloud

Azure, AWS, GCP com serviços robustos.

Frameworks

LangChain, LlamaIndex, Haystack facilitam integração.

Tópicos da Apresentação

Camada de Orquestração

Como e onde chamar LLMs com contexto adequado usando frameworks como LangChain e Agents.

Gerenciamento de Contexto

Importância de vetores, estratégias eficientes de memória para IA generativa.

Performance e Economia

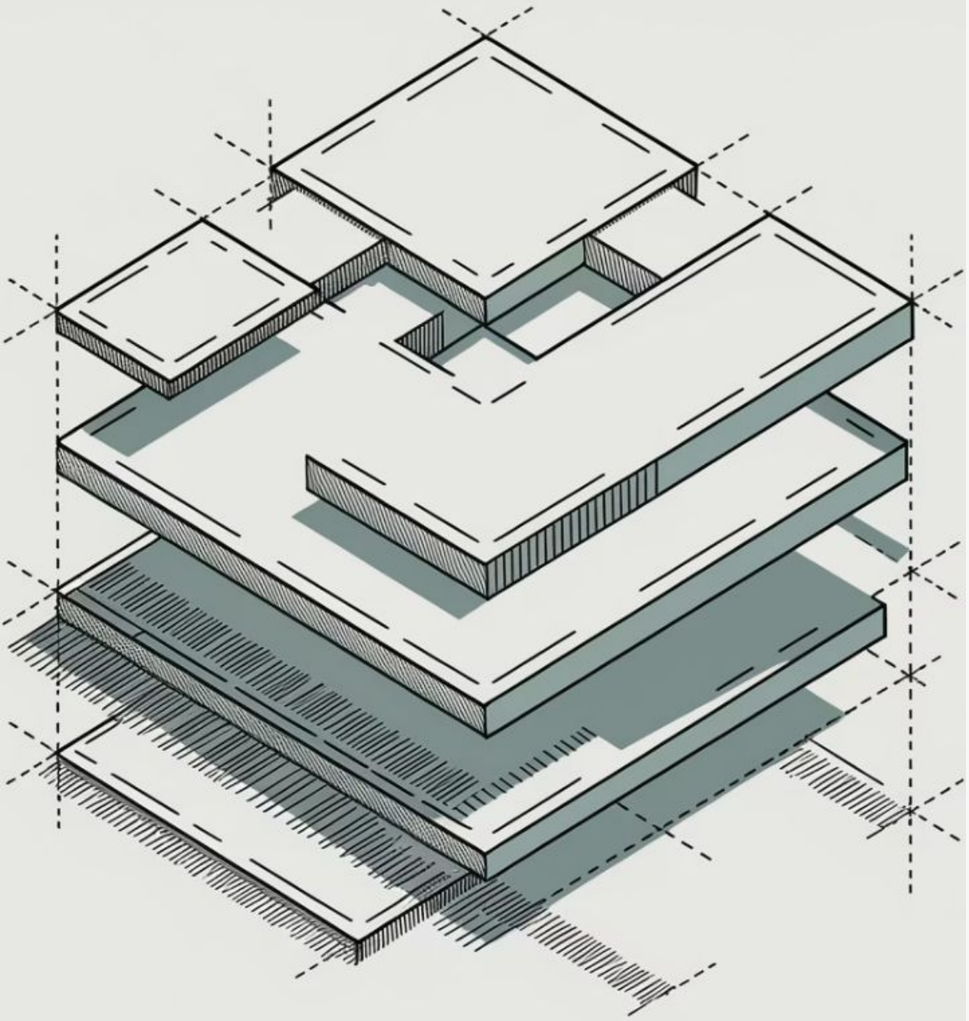
Técnicas de caching e assincronicidade para reduzir gastos de tokens e latência.

Segurança e Compliance

Proteção contra vazamento de dados sensíveis e informações pessoais identificáveis.

Observabilidade e Testes

Registro de prompts, avaliação de resultados e manejo do não-determinismo.



Camada de Orquestração

Isolamento

Separe a implementação da LLM do restante do código. Crie um serviço dedicado para LLMs.

Abstração

Nenhuma outra camada precisa saber o que é um prompt. Use interfaces amigáveis e semânticas

Desacoplamento

Se não estiver usando um framework, evite acoplamento com providers específicos. Modelos melhores surgem semanalmente.

Prompts mais estruturados

Usar adequadamente prompts de sistema e prompts de usuário pode ajudar na eficiência do consumo de tokens da sua aplicação. Templates também.

Gerenciamento de Contexto

Dados Vetoriais

Armazene dados em bases como Pinecone, Faiss ou Weaviate.

Pesquisa Semântica

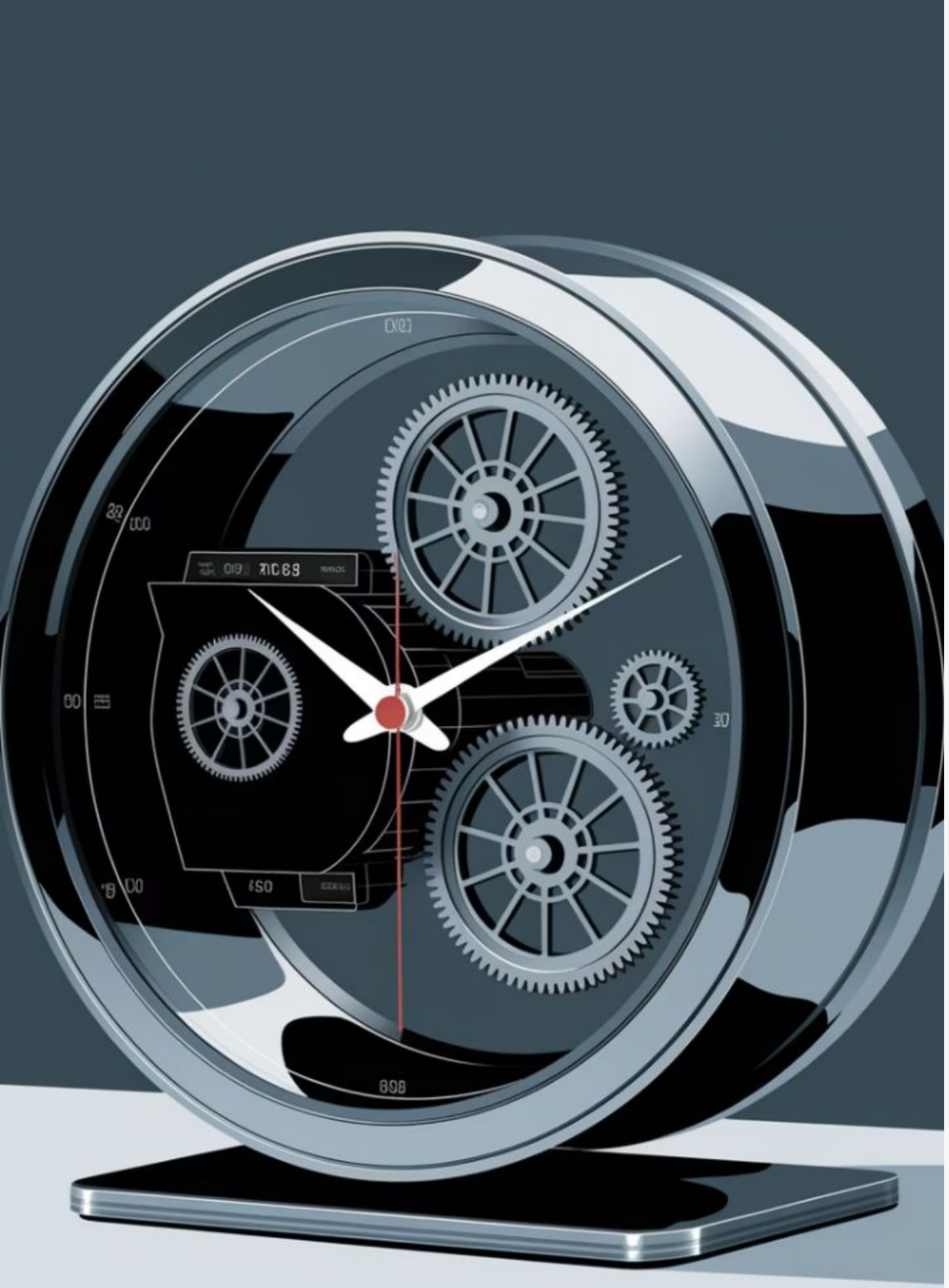
Busque contexto relevante usando significado, não palavras-chave.

Otimização

Equilibre tamanho do prompt, resposta e contexto para economizar tokens.

Montagem de Prompt

Combine contexto com instruções claras para a LLM.



Performance e Economia

1

Caching Semântico

Salve interações similares em banco de dados vetorial. Busque respostas existentes antes de chamar LLMs caras.

Economize dinheiro e melhore performance com consultas similares.

2

Processamento Assíncrono

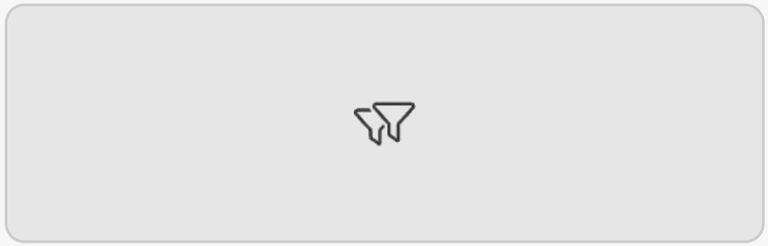
Use APIs em batelada para processos não urgentes. Economize até 50% no custo de tokens.

Ideal para tarefas que podem esperar até 24 horas por resposta.

Segurança e Compliance

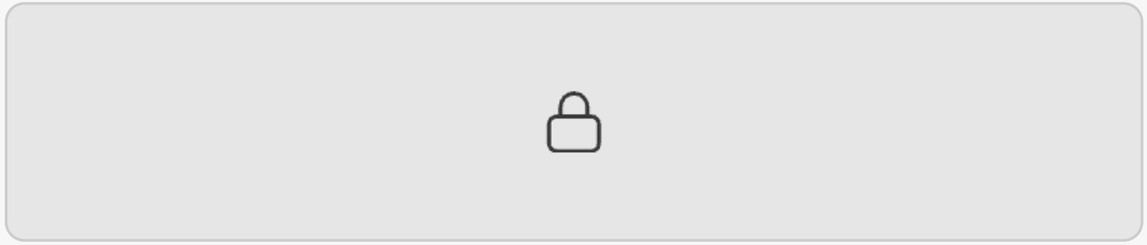
Guardrails

LLMs que validam respostas de outras LLMs



Sanitização

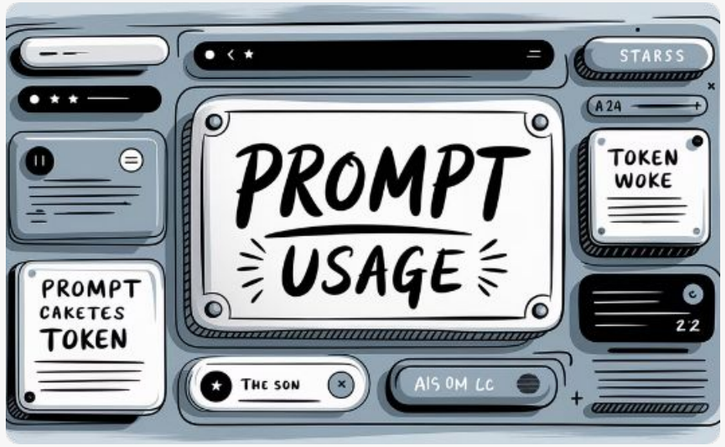
Limpe inputs do usuário antes de enviar à LLM

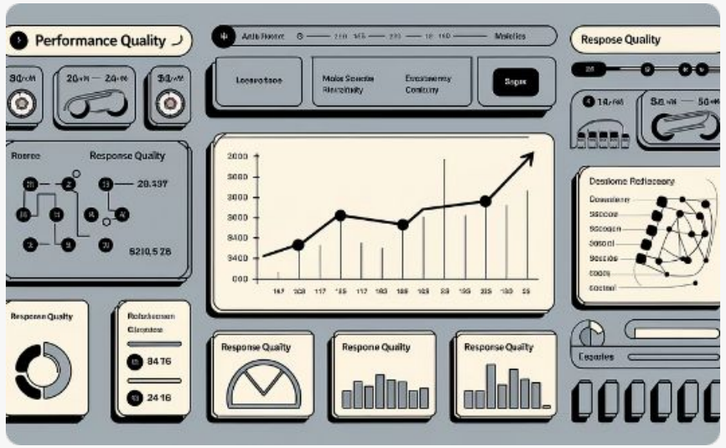


Proteção de Dados

Evite vazamento de informações sensíveis

Observabilidade





Logging Estruturado

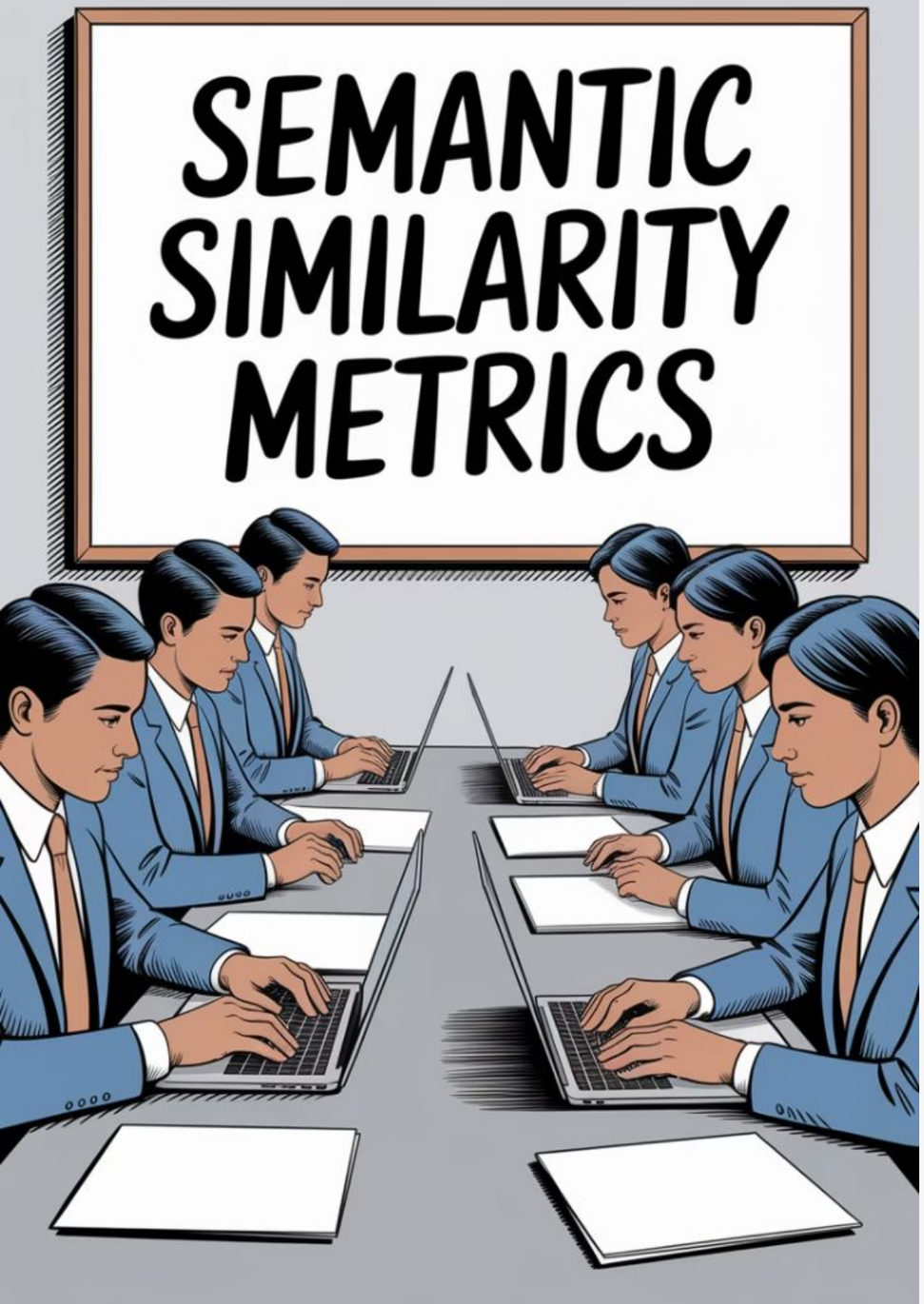
Registre prompts, contextos, respostas, tempo e consumo de tokens em cada interação.

Ferramentas Especializadas

Use Langsmith, Datadog ou Traceloop para monitorar interações com LLMs.

Otimização Contínua

Analise padrões para melhorar prompts e economizar tokens.



Testes e Garantia de Qualidade

Métricas Semânticas

Use similaridade semântica para comparar respostas com resultados ideais.

Avalie qualidade além de exatidão.

Testes Dinâmicos

Adapte testes para lidar com respostas não-determinísticas. Estabeleça faixas aceitáveis de variação.

Human in the Loop

Inclua revisão humana quando precisão absoluta for necessária.

Garanta respostas coerentes em casos críticos.

Obrigado pela Atenção!

Suas dúvidas e observações são valiosas para continuarmos esta conversa sobre arquiteturas para IA generativa.

Acompanhe a Levva nas redes sociais para mais conteúdos sobre tecnologias emergentes e IA.

LinkedIn levva

Siga-nos para artigos técnicos e novidades sobre IA generativa.

Instagram: @levva.io

Acompanhe eventos, webinars e bastidores de nossos projetos.

Eventos

Estamos sempre realizando eventos com e para a comunidade aqui no hub.