

warning "Atualização de backend (outubro de 2026)" A inferência do GitHub Models foi descontinuada em 30 de julho de 2026. As referências abaixo são mantidas apenas como contexto histórico enquanto este lab é migrado. Use o Lab 013 ? Backends de LLM gratuitos para configurar o Foundry Local ou o Ollama antes de continuar.

## Lab 038: Otimização de Custos de IA

Nível: L300 Trilha: ? Pro Code Tempo: ~45 min ? Custo: Azure Free Tier ? A maioria das estratégias são gratuitas para implementar

O que Você Vai Aprender

Entender os fatores de custo de APIs de IA (tokens, escolha de modelo, frequência de requisições) Aplicar 5 estratégias-chave para reduzir custos: cache, batching, roteamento de modelos, compressão de prompts e streaming Implementar cache semântico para evitar chamadas LLM duplicadas Construir um roteador de modelos que usa modelos baratos para consultas simples e modelos caros para as complexas Monitorar e alertar sobre custos com Azure Monitor

### Introdução

Um agente de IA em produção pode facilmente custar \$50?500/mês mesmo com uso modesto, se não otimizado. A boa notícia: a maioria das estratégias de redução de custos também melhora a latência e a experiência do usuário.

As principais alavancas de custo:

Fator | Impacto | Otimização | Escolha do modelo | Diferença de custo de 10?50× | Rotear para modelo mais barato quando possível | Contagem de tokens | Custo linear | Comprimir prompts, cortar histórico | Acertos de cache | Elimina o custo inteiramente | Cache semântico para consultas repetidas | Volume de requisições | Custo linear | Batch, debounce, deduplicar | Tamanho da saída | Custo linear | Usar max\_tokens, saída estruturada |

### Parte 1: Meça Sua Linha de Base

Antes de otimizar, meça o que você está gastando:

[diagram or code sample]

### Parte 2: Estratégia 1 ? Cache Semântico

Um cache semântico armazena respostas de LLM e as reutiliza para consultas semanticamente similares, mesmo quando a formulação difere:

[diagram or code sample]

### Parte 3: Estratégia 2 ? Roteador de Modelos

Use modelos baratos para consultas simples, modelos caros para as complexas:

[diagram or code sample]

### Parte 4: Estratégia 3 ? Compressão de Prompts

Corte o histórico de conversas para manter apenas as mais recentes + resumo:

[diagram or code sample]

### Parte 5: Monitoramento de Custos com Azure Monitor

Em produção, acompanhe os custos com Azure Monitor:

[diagram or code sample]

No Portal do Azure, crie um alerta:

[diagram or code sample]

### ? Verificação de Conhecimento

question "1. O que é cache semântico e como ele difere do cache de correspondência exata?"

Cache de correspondência exata retorna resultados em cache apenas quando a entrada é idêntica byte a byte. Cache semântico usa embeddings para encontrar respostas em cache para entradas semanticamente similares ? "Mostre-me tendas" e "Que tendas vocês têm?" obtêm a

mesma resposta em cache mesmo com formulações diferentes. O cache semântico é muito mais eficaz para IA conversacional.

question "2. Quando você NÃO deve usar um modelo barato em um roteador de modelos?" Use um modelo premium para: raciocínio em múltiplos passos (cadeia de pensamento), comparações entre opções complexas, recomendações personalizadas que requerem pesar muitos fatores, tarefas onde erros são custosos (médico, financeiro, jurídico). Use modelos baratos para: consultas, listagens, formatação, perguntas e respostas simples onde a tolerância a erros é maior.

question "3. Quais são os riscos de comprimir agressivamente o histórico de prompts?" O agente pode perder contexto que precisa: um usuário mencionou uma alergia 5 turnos atrás e o resumo comprimido perdeu isso. Um histórico mal resumido pode causar respostas contraditórias em conversas longas. Sempre mantenha os N turnos mais recentes na íntegra; apenas resuma turnos mais antigos. Teste com cenários que exigem contexto de longo alcance.  
Resumo

Estratégia | Economia Potencial | Complexidade | Cache semântico | 40?80% das chamadas de API | Média | Roteamento de modelos | Redução de custo de 10?40× | Baixa | Compressão de prompts | Redução de 20?60% dos tokens | Baixa | Batching de requisições | 10?30% (trade-off de latência) | Média | Limitação de saída (max\_tokens) | 5?20% | Muito Baixa |  
Próximos Passos

Observe o que você otimizou: ? Lab 033 ? Observabilidade de Agentes Avalie a qualidade após otimização: ? Lab 035 ? Avaliação de Agentes Deploy completo em produção: ? Lab 028 ?  
Deploy MCP no Azure

Kit de apoio

O build do site gera downloads estáveis em PDF e ZIP. O ZIP contém o guia do lab e todos os arquivos de apoio declarados para este lab.

Baixar PDF gerado Baixar ZIP gerado

Supporting file: CHEATSHEET.md

Arquivos de apoio: Lab 038: Otimização de Custos de IA

Comece pelo caminho local ou gratuito documentado. Registre premissas e evidências de validação. Mantenha serviços hospedados opcionais, salvo quando o lab os exigir.

Supporting file: README.md

Lab 038: Otimização de Custos de IA

Use estes arquivos com o guia do lab e o kit gerado.