ADR - Decisão sobre LLM para Plataforma de Atendimento

Título

Escolha de LLM para plataforma de atendimento ao cliente automatizado

Status

Aprovado

Contexto

A empresa está desenvolvendo uma plataforma de atendimento ao cliente automatizado utilizando IA, com planos de expansão para SDR e vendas. É necessário decidir entre utilizar APIs de LLMs comerciais (como Claude, GPT) ou implementar LLMs open-source hospedados internamente. A decisão impacta custos, escalabilidade, privacidade de dados e tempo de desenvolvimento.

Opções Consideradas

1. APIs de LLMs Comerciais

Vantagens:

- Rápido para implementar
- Sem necessidade de manutenção de infraestrutura de IA
- Modelos geralmente mais poderosos e atualizados
- Escalabilidade facilitada
- Free-tiers disponíveis para iniciar com custo reduzido

Desvantagens:

- Custos contínuos que aumentam com o uso
- Menor controle sobre o modelo
- Possíveis preocupações com privacidade de dados

2. LLMs Open-Source Hospedados Internamente

Vantagens:

- Maior controle sobre o modelo e dados
- Custos fixos de infraestrutura (potencialmente mais baratos em grande escala)
- Possibilidade de personalização mais profunda
- Dados não saem da infraestrutura da empresa

Desvantagens:

- Maior complexidade técnica
- Necessidade de recursos computacionais significativos
- Necessidade de equipe com conhecimento especializado
- Tempo de implementação maior

3. Abordagem Híbrida

Vantagens:

- Flexibilidade para diferentes casos de uso
- Otimização de custos para casos frequentes
- Balanceamento entre tempo para mercado e controle
- Transição gradual possível

Desvantagens:

- Maior complexidade de implementação e manutenção
- Necessidade de orquestração entre sistemas
- Potencial inconsistência nas respostas

Decisão

Escolhida: Abordagem híbrida com foco inicial em APIs comerciais

Implementação em fases:

Fase 1 (MVP e início das operações):

- Implementar integração com APIs de LLMs comerciais (Claude/GPT)
- Utilizar free-tiers e planos iniciais
- Desenvolver sistema de gerenciamento de créditos
- Preparar a infraestrutura para eventual migração parcial

Fase 2 (com crescimento da base de clientes):

- Implementar LLMs open-source para casos de uso mais comuns e repetitivos
- Manter APIs comerciais para casos complexos ou de alta precisão
- Desenvolver sistema de roteamento inteligente entre LLMs

Justificativa

- **Tempo para mercado**: APIs comerciais permitem implementação mais rápida e foco em diferenciais de produto
- **Recursos iniciais limitados**: Free-tiers das APIs comerciais são suficientes para MVP e primeiros clientes
- **Modelo de créditos**: Sistema de créditos pré-pagos permite transferir custos para clientes de forma transparente
- **Conhecimento da equipe em IA**: A equipe possui conhecimento técnico para eventual implementação de modelos próprios
- **Foco no idioma português**: Alguns modelos open-source podem ter desempenho inferior em português comparado às APIs comerciais
- **Flexibilidade**: Abordagem híbrida permite adaptação conforme crescimento e necessidades específicas

Consequências

- Dependência inicial de fornecedores externos (risco mitigado pela possibilidade de migração parcial futura)
- Necessidade de desenvolver sistema de controle de créditos por tipo de operação
- Implementação gradual de infraestrutura para LLMs próprios
- Potencial necessidade de ajustes no sistema RAG para diferentes tipos de LLM
- Desenvolvimento de lógica de roteamento entre diferentes LLMs

Métricas para Revisão da Decisão

Esta decisão será revisada quando:

- Base de clientes atingir 50 empresas ativas
- Volume mensal de conversas ultrapassar 100.000
- Custo mensal com APIs comerciais ultrapassar 30% da receita
- Surgimento de modelos open-source com performance comparável em português

Monitoramento

Serão monitorados os seguintes aspectos para informar futuras decisões:

- Custos por conversa em cada abordagem
- Qualidade das respostas (avaliação humana e métricas automáticas)
- Tempo de resposta
- Utilização de recursos computacionais

Data: 20/04/2025

Participantes: [Equipe do Projeto]

Próxima revisão agendada: 20/07/2025