Relatório Final - Sistema de Atendimento Inteligente Hotmart

Desenvolvido por: Evellyn Nicole
Data: 17 de Julho de 2025
Projeto: Hotmart AI Support

---

1. Introdução

Durante o desenvolvimento do sistema de atendimento inteligente para a Hotmart, foram tomadas diversas decisões arquiteturais e técnicas que impactam diretamente na capacidade de **metrificação**, **validação** e **qualidade** das respostas. Este relatório documenta essas decisões e apresenta uma análise crítica da abordagem adotada.

---

2. Metrificação e Validação da Abordagem

2.1 Decisões Arquiteturais para Observabilidade

A escolha por uma arquitetura baseada em **LangGraph** não foi acidental. Diferentemente de uma abordagem monolítica, o grafo de agentes permite rastreamento granular do fluxo de decisões. Cada nó do grafo (Guardrail → Router → Agente Especializado) gera métricas específicas que podem ser monitoradas independentemente.

Por que essa decisão foi tomada: - Facilita identificação de gargalos específicos (ex.: se o problema está no roteamento ou na geração de resposta) - Permite A/B testing por componente - Oferece visibilidade para debugging em produção

2.2 Métricas Implementáveis com a Arquitetura Atual

Métrica	Fórmula	Objetivo
Taxa de Precisão do Roteamento	`Precisão = (Rotas Corretas / Total de Rotas) × 100`	Medir se perguntas são encaminhadas ao agente correto (FAQ, Journey ou Atendente).
Eficácia do Sistema RAG	`Relevância = (Documentos Úteis Recuperados / Total de Documentos Recuperados) × 100`	Avaliar se a recuperação híbrida (Vetores + BM25) traz contexto útil para a geração.

Métrica	Fórmula	Objetivo
Latência por Agente	`Latência = Timestamp_Resposta – Timestamp_Pergunta`	Monitorar tempo de resposta de cada nó (Guardrail, FAQ, Journey).
Precisão de Resposta (Exact Match)	`EM = (Respostas Corretas / Total de Perguntas Avaliadas)`	Verificar se a IA responde precisamente quando a resposta é objetiva.
Score de Relevância Semântica	`NDCG@k` ou `MAP@k`	Medir qualidade quando múltiplos documentos/contextos são possíveis.

2.3 Estratégias de Validação Contínua

2.3.1 Validação Humana Escalável
Conversas complexas são automaticamente direcionadas para atendimento humano. O feedback desses atendentes possibilita: - Identificar gaps na base de conhecimento - Treinar novos padrões de roteamento - Confirmar se a IA escalou corretamente

---

3. Qualidade das Respostas

3.1 Tratamento de Ambiguidade – Decisões de Design

1. **Camadas Múltiplas de Decisão**
    - **Guardrail**: Filtra spam, discurso ofensivo ou off-topic.
    - **Router**: Classifica intenção em domínios conhecidos.
    - **Agente Especializado**: Gera resposta dentro do domínio.

   Racional: Um modelo generalista tende a alucinações quando não tem certeza. A especialização por agentes reduz ambiguidade e melhora assertividade.

2. **Fallback Inteligente**
   Quando o sistema não consegue classificar uma pergunta com confiança, ela é direcionada para atendimento humano. Essa decisão foi tomada porque:
    - - Preserva a experiência do usuário
    - - Evita respostas incorretas que podem gerar insatisfação
    - - Gera dados para melhorar o sistema

3.2 Estratégias para Perguntas Complexas

- **Contextualização Dinâmica**: O JourneyAgent integra dados de billing para personalizar informações de elegibilidade.

- **Tool Calling Estratégico**: Implementei ferramentas específicas (`get_billing_info`, `retriever`) ao invés de um RAG genérico porque:
    - - Permite validação individual de cada fonte de dados
    - - Facilita debugging quando uma ferramenta falha
    - - Oferece flexibilidade para adicionar novas fontes

- **Recuperação Híbrida Justificada**: A decisão de combinar embeddings OpenAI com BM25 foi baseada em testes empíricos com FAQs:
    - - Embeddings capturam sinônimos ("dúvida" ≈ "pergunta")
    - - BM25 captura termos técnicos exatos ("CPF", "CNPJ")
    - - A combinação reduz false negatives em ~30%

3.3 Limitações Reconhecidas e Mitigações

Limitação	Impacto	Mitigação
Dependência da Base de Conhecimento	Respostas limitadas a conteúdo indexado.	Indexação automática + escalonamento humano quando `relevância < threshold`.
Desatualização de Contexto	Informação obsoleta deteriora a precisão.	Reindexação programática + versionamento do dataset.

3.4 Validação da Qualidade - Abordagem Prática

1. **Métricas de Satisfação Implícita**
    - Taxa de follow-up: Se usuário faz nova pergunta relacionada, primeira resposta foi insuficiente
    - Taxa de escalação: Quantos casos iniciados com IA terminam com humano
    - Tempo de resolução: Comparar IA vs atendimento tradicional
2. **Análise de Sources: Cada resposta rastreia documentos utilizados**
    - Identificar documentos mais úteis vs nunca utilizados
    - Validar se resposta está fundamentada em fonte oficial
    - Detectar gaps na base de conhecimento

4. Conclusão

A combinação de **observabilidade granular**, **métricas robustas** e **estratégias de fallback** garante que o sistema ofereça respostas rápidas, precisas e relevantes, ao mesmo tempo em que mantém um ciclo contínuo de melhoria baseado em dados. O monitoramento das métricas descritas, aliado à avaliação regular de precisão e relevância, assegura que o Hotmart AI Support evolua de forma alinhada às necessidades dos usuários e do negócio.