O CÓDIGO INVISÍVEL DOS SUPERAGENTES DE IA

Protocolo de auditoria de fidelidade autoral

Material complementar do livro de Jefferson Lobo — Método DEL

jeffersonlobo.tech

APÊNDICE C

Protocolo de auditoria de fidelidade autoral

A auditoria garante que o agente continua soando como o autor ao longo do tempo. Três momentos: lançamento, periódico, por gatilho.

CHECKLIST 1 — AUDITORIA DE LANÇAMENTO

(antes de colocar o agente em produção)

PREPARAÇÃO

- ☐ Os três arquivos DEL foram gerados e revisados?
- ☐ Os arquivos e os textos de conteúdo foram carregados na plataforma?
- ☐ O prompt de identidade foi configurado com as cinco seções?
- ☐ O log de versão foi iniciado (data, versão 1.0)?

TESTE DE FIDELIDADE SINTÁTICA — 5 perguntas

Para cada resposta, avalie:

- ☐ Extensão das frases corresponde ao perfil
- ☐ Tipo frasal dominante está correto
- ☐ Pontuação expressiva está presente onde deveria
- ☐ Abertura de parágrafo segue o padrão do autor
- ☐ Extensão dos parágrafos está dentro do esperado

Mínimo aceitável para lançamento: 4 de 5 respostas em conformidade.

TESTE DE FIDELIDADE SEMÂNTICA — 5 perguntas argumentativas

Para cada resposta, avalie:

- ☐ Frame semântico dominante está presente
- ☐ Tipo de argumentação corresponde ao perfil
- ☐ Nível de formalidade está correto
- ☐ Intenção comunicacional está sendo cumprida
- ☐ Instruções contextuais estão sendo seguidas

Mínimo aceitável: 4 de 5.

TESTE DE FIDELIDADE LEXICAL — 5 perguntas variadas

Para cada resposta, avalie:

- ☐ Pelo menos 2 verbos do glossário foram usados
- ☐ Nenhum termo do glossário invertido apareceu
- ☐ Conectores argumentativos correspondem ao perfil
- ☐ Alguma expressão idiomática característica foi usada
- ☐ Nenhum marcador de geração genérica apareceu

Mínimo aceitável: 4 de 5.

TESTE DE COMPORTAMENTO DE BORDA — 5 situações-limite

- ☐ Pergunta fora do escopo: recusou adequadamente
- ☐ Pergunta ambígua: pediu esclarecimento ou escolheu interpretação alinhada ao frame
- ☐ Pergunta que o agente não sabe: admitiu sem improvisar
- ☐ Pergunta emocionalmente carregada: ativou o modo correto

- ☐ Pergunta sensível: seguiu instruções contextuais específicas

Mínimo aceitável: 4 de 5.

RESULTADO GERAL:

Aprovado para produção: ≥ 4/5 em todos os quatro testes.

Revisar antes de lançar: qualquer teste com < 4/5.

CHECKLIST 2 — AUDITORIA PERIÓDICA

(mensal para alto volume, trimestral para baixo volume)

SELEÇÃO DA AMOSTRA

- ☐ 20 a 30 conversas reais do período
- ☐ Variedade: simples, complexas, de borda, emocionais, operacionais
- ☐ Inclua conversas marcadas como problemáticas

ANÁLISE

Para cada conversa, classifique por eixo:

✓ corresponde / ✗ desvia / ? parcial

CÁLCULO DE CONFORMIDADE

Sintática: ____% (meta: ≥ 80%)

Semântica: ____% (meta: ≥ 80%)

Lexical: ____% (meta: ≥ 80%)

Para cada eixo com conformidade < 80%, descreva o padrão de desvio mais frequente e a ação corretiva.

REGISTRO

```
Data da auditoria  ·
Versão do prompt antes  ·  Versão após
```

ajustes | Próxima auditoria programada.

CHECKLIST 3 — AUDITORIA POR GATILHO

Gatilhos que exigem auditoria imediata:

- ☐ Usuário reportou “não pareceu com a marca”
- ☐ Informação factualmente incorreta
- ☐ Confusão por resposta fora de contexto
- ☐ Resposta inadequada a situação sensível
- ☐ Plataforma atualizada pelo fornecedor
- ☐ Perfil do autor evoluiu
- ☐ Agente passou a cobrir novos domínios

DIAGNÓSTICO POR TIPO

“Não pareceu com a marca” → alucinação estilística

Ação: comparar resposta com arquivos DEL, identificar eixo que desviou, reforçar diretriz no prompt.

“Informação incorreta” → alucinação factual Ação: verificar se a informação está na base RAG. Adicionar se faltar. Reforçar instrução "quando não souber, diga que não sabe".

“Resposta para situação errada” → alucinação contextual

Ação: adicionar instrução contextual ao arquivo semântico e ao prompt.

“Plataforma atualizada” → refazer protocolo de lançamento completo (20 perguntas, 4 testes).

“Perfil evoluiu” → nova coleta de textos (mínimo 3 do período atual), comparar com análises anteriores, atualizar arquivos DEL, incrementar versão (1.0 → 2.0).

LOG DE VERSÕES — MODELO

Perfil: ____

Plataforma: ____

VERSÃO 1.0

```
Data · O que foi configurado ·
Textos de referência
```

Resultado dos testes de lançamento.

VERSÃO 1.1, 1.2, etc.

```
Data · Gatilho da atualização ·
O que foi ajustado
```

Por que foi ajustado I Resultado pós-ajuste.

VERSÃO 2.0 (mudança estrutural)

```
Data · Gatilho ·
O que mudou estruturalmente · Novos textos
```

de referência I Resultado pós-atualização.

Mantenha o log atualizado a cada mudança. É a documentação que permite reconstruir decisões meses depois — e é o que diferencia um agente gerenciado de um agente abandonado.

FECHAMENTO — O QUE ESTE MÉTODO DEFENDE

Eu criei o Método DEL porque acredito que a próxima etapa da inteligência artificial não será apenas gerar mais texto. Será preservar responsabilidade, contexto e voz quando máquinas começarem a responder em nosso nome.

A pergunta que fica não é se a IA pode escrever. Ela pode. A pergunta é se aquilo que ela escreve ainda carrega a estrutura de linguagem que torna uma pessoa, uma marca ou uma instituição reconhecível.

Quando a resposta for sim, a IA deixa de ser uma produtora de texto genérico e passa a ser uma ferramenta orientada por identidade.