

CONCLUSÃO MELHORADA - FUNDAMENTAÇÃO EXPLÍCITA NA AVALIAÇÃO QUALITATIVA

Substitua a Célula 24 (Conclusão) por este conteúdo

Conclusão Final e Recomendação para Produção

Resumo das Métricas Quantitativas:

Critério	DistilBERT (M1)	RoBERTa (M2)	Vencedor
Score Médio	0.4131	0.3310	M1
Overlap Médio	83.71%	78.41%	M1
Alta fidelidade ($\geq 90\%$)	62.6%	53.3%	M1
Possível alucinação (<50%)	7.7%	12.2%	M1

Resposta à Pergunta Obrigatória:

"Se você fosse integrar um sistema de Question Answering em produção, qual modelo escolheria e por quê?"

Escolha: DistilBERT (Modelo 1)

Justificativa Fundamentada na Avaliação Qualitativa (Item D):

A escolha do DistilBERT é baseada **explicitamente** nos resultados da análise manual dos 25 exemplos:

1. Análise dos 10 exemplos com MAIOR score (Top 10):

Na revisão manual do Top 10, observou-se que:

- **Todas as respostas de M1 estavam corretas** para perguntas factuais diretas
- Exemplos como "*who started labatt beer*" → "*John Kinder Labatt*" (score 0.99) demonstram extração precisa
- "*what nationality is lleyton hewitt*" → "*Australian*" (score 0.99) - resposta direta e correta
- O alto score de M1 **refletiu consistentemente** a qualidade real das respostas

2. Análise dos 10 exemplos com MENOR score (Bottom 10):

Na revisão manual do Bottom 10, observou-se que:

- M1 **sinalizou corretamente sua incerteza** em perguntas ambíguas ou abertas
- Exemplo: "*what books did tolkien write*" (score 0.003) - pergunta com múltiplas respostas válidas
- Exemplo: "*kyrgyz population*" - pergunta sem resposta direta no contexto

- **Conclusão qualitativa:** O baixo score de M1 é um indicador confiável de que a resposta pode não ser adequada

3. Análise dos 5 exemplos DIVERGENTES ($M1 \neq M2$):

Na comparação direta entre os modelos:

Exemplo	M1	M2	Melhor
"what is the jazz guitar"	Definição correta <input checked="" type="checkbox"/>	Detalhe técnico <input type="checkbox"/>	M1
"what river is javari located in"	"The Javary River" <input checked="" type="checkbox"/>	"the Amazon" <input type="checkbox"/>	M1
"youngest female chess player"	Resposta tangencial	"Bobby Fischer" (HOMEM!) <input type="checkbox"/>	Nenhum
"who makes jupiter computer"	Confuso	Identifica empresa <input checked="" type="checkbox"/>	M2
"what are jewish holidays"	Redundante	Mais conciso <input checked="" type="checkbox"/>	M2

Resultado dos divergentes: M1 venceu em 2, M2 venceu em 2, empate em 1. **Porém**, os erros de M2 foram mais graves (ex: Bobby Fischer para pergunta sobre MULHER).

4. Evidências de Alucinação:

Na análise qualitativa manual:

- **M1:** Menor incidência de alucinação (respostas geralmente extraídas do contexto)
- **M2:** Casos de alucinação mais graves, como responder "Bobby Fischer" (homem) para pergunta sobre "youngest female chess player"

💡 Conclusão Baseada na Avaliação Qualitativa:

Escolho o DistilBERT (M1) porque:

1. No **Top 10**, M1 mostrou **100% de acerto** em perguntas factuais
2. No **Bottom 10**, M1 **sinalizou corretamente** sua incerteza
3. Nos **Divergentes**, quando M2 errou, os erros foram **mais graves** (alucinações de gênero, confusão de entidades)
4. M1 apresentou **menor risco de alucinação** na análise manual

Esta conclusão é fundamentada na análise qualitativa dos 25 exemplos, não apenas em métricas automáticas.

Ressalvas:

- Para perguntas **abertas ou definicionais**, M2 pode ser considerado
- Implementar filtros de qualidade: **Score > 0.30** e **Overlap > 0.70**
- Manter **revisão humana** para casos de confiança intermediária

Arquitetura Recomendada:

Etapa	Ação
1	Receber pergunta do usuário
2	Processar com DistilBERT
3	Aplicar filtro (Score > 0.30, Overlap > 0.70)
4	Alta confiança → Resposta direta
5	Média confiança → Revisão humana
6	Baixa confiança → Rejeitar