Bootstrap v2.1 — Relatório Analítico Completo

Gerado: 2026-02-18 11:15 BRT
Modelo: Claude Haiku 4.5 (thinking: low)
Rounds: 10 (R1-R3 baseline puro, R4-R10 com cross-round patterns)
Ground Truth: Enriched via Sonnet (Stage 0.8, sem sanitização)
Custo total: ~$0.80 (10 rounds × ~$0.08)
Dataset: 100 PRs/round (50 merge, 50 sample estratificada)

---

1. Definição das Métricas

1.1 Merge Prediction (tarefa principal)

O modelo recebe features de um PR e prediz: **merge** ou **no-merge**.

Métrica	Definição	Por que importa
Accuracy	(TP+TN) / Total	Métrica geral, mas enviesada se classes desbalanceadas
Precision	TP / (TP+FP)	"Dos que eu disse merge, quantos realmente foram?" — custo do FP
Recall	TP / (TP+FN)	"Dos que foram merge, quantos eu peguei?" — custo do FN
F1	2×P×R / (P+R)	Média harmônica de precision/recall — métrica principal

Confusion matrix: - **TP (True Positive):** Predisse merge, foi merge □ - **FP (False Positive):** Predisse merge, NÃO foi merge × (sobre-otimismo) - **TN (True Negative):** Predisse no-merge, não foi merge □ - **FN (False Negative):** Predisse no-merge, MAS foi merge × (merge perdido)

1.2 Dedupe Detection (tarefa secundária)

O modelo identifica pares de PRs que são duplicatas (mesmo problema/fix).

Métrica	Definição
Dedupe Precision	Pares corretos / Pares preditos
Dedupe Recall	Pares corretos / Pares reais (ground truth)
Dedupe F1	Harmônica de precision/recall

Métrica	Definição
GT Pairs	Número de pares duplicatas reais no sample
Pred Pairs	Número de pares que o modelo identificou

1.3 Calibration

Confiança do modelo vs accuracy real, em bins de 10%.
- **Overconfident:** confiança > accuracy + 10pp (diz que tem certeza mas erra)
- **Underconfident:** confiança < accuracy - 10pp (acerta mais do que admite)

1.4 Gates

Gate	Critério	Propósito
G1	Accuracy ≥ 70%	Modelo é melhor que random
G2	Delta learning ≥ +3pp	Cross-round patterns melhoram performance

---

2. Resultados Gerais

Round	Fase	Accuracy	Precision	Recall	F1	Dedupe F1	Erros
R1	baseline	70.0%	40.6%	54.2%	46.4%	78.6%	30
R2	baseline	72.0%	44.1%	62.5%	51.7%	82.1%	28
R3	baseline	66.0%	38.6%	70.8%	50.0%	82.4%	34
R4	learning	75.0%	48.6%	70.8%	57.6%	70.3%	25
R5	learning	78.0%	53.3%	66.7%	59.3%	84.6%	22
R6	learning	74.0%	47.4%	75.0%	58.1%	62.1%	26
R7	learning	72.0%	42.9%	50.0%	46.2%	64.3%	28
R8	learning	75.0%	48.5%	66.7%	56.1%	70.6%	25
R9	learning	78.0%	53.1%	70.8%	60.7%	70.0%	22
R10	learning	70.0%	41.2%	58.3%	48.3%	86.5%	30

Médias por fase

Métrica	Baseline (R1-3)	Learning (R4-10)	Delta
Accuracy	69.3%	74.6%	**+5.2pp**
Precision	41.1%	47.8%	+6.7pp

Métrica	Baseline (R1-3)	Learning (R4-10)	Delta
Recall	62.5%	65.5%	+3.0pp
F1	49.4%	55.2%	**+5.8pp**

Gate G1: □ PASS (learning avg 74.6% ≥ 70%)
Gate G2: □ PASS (delta +5.2pp ≥ +3pp)

---

3. Análise de Erros (n=270)

3.1 Distribuição geral

- **FP (falsos positivos):** 185 (69%) — modelo diz "vai ser merged" mas não foi
- **FN (falsos negativos):** 85 (31%) — modelo diz "não merge" mas foi merged

O modelo é sistematicamente sobre-otimista. A cada round, prediz merge demais.

3.2 Perfil do FP típico

O FP típico é um PR de autor com **alta taxa histórica de merge** (média 99%) mas que desta vez não foi aceito. 109/165 FPs (66%) vêm de autores com merge rate < 25% — indicando que o modelo não está usando bem o autor como feature.

Insight: O modelo confia demais no histórico do autor. Autor com 99% merge rate não garante que *este* PR será merged. Faltam features de qualidade do PR individual.

3.3 Perfil do FN típico

O FN típico é um PR de autor com **zero ou baixíssima taxa de merge** (média 1.0%) — mas que foi aceito. 18 de 85 FNs (21%) são PRs de tamanho L ou XL.

Insight: O modelo descarta PRs de autores "fracos" mesmo quando o PR em si é bom. Autores com 0% histórico podem ter um PR que vale merge — o modelo não distingue.

3.4 Erros por categoria

Categoria	FP	FN	Observação
agents	63	15	forte viés otimista
gateway	20	5	forte viés otimista

Categoria	FP	FN	Observação
None	9	8	
docs	4	12	viés pessimista
channel	11	3	forte viés otimista
commands	9	2	forte viés otimista
fix	8	2	forte viés otimista
cli	6	2	

`agents` **domina os FPs (63/185 = 34%).** É a maior categoria do repo — alta exposição + alta variância de qualidade.

3.5 Evolução dos erros por round

Round	FP	FN	Total	FP%
R1	19	11	30	63%
R2	19	9	28	68%
R3	27	7	34	79%
R4	18	7	25	72%
R5	14	8	22	64%
R6	20	6	26	77%
R7	16	12	28	57%
R8	17	8	25	68%
R9	15	7	22	68%
R10	20	10	30	67%

Observação: Sem tendência clara de melhoria nos erros ao longo dos rounds. Cross-round patterns melhoram accuracy (+5.2pp) mas não resolvem o viés FP estrutural.

---

4. Calibração

R1: 5 bins, 1 overconfident, 1 underconfident
R2: 5 bins, 0 overconfident, 0 underconfident
R3: 5 bins, 2 overconfident, 0 underconfident
R4: 6 bins, 0 overconfident, 1 underconfident
R5: 5 bins, 0 overconfident, 1 underconfident
R6: 5 bins, 0 overconfident, 0 underconfident
R7: 6 bins, 2 overconfident, 2 underconfident

R8: 8 bins, 1 overconfident, 3 underconfident
R9: 5 bins, 2 overconfident, 0 underconfident
R10: 6 bins, 4 overconfident, 1 underconfident ⚠

R10 é preocupante: 4 bins overconfident — modelo ficou mais confiante sem ficar mais preciso. Possível overfitting aos patterns injetados.

---

5. Dedupe

Round	Precision	Recall	F1	GT Pairs	Pred Pairs
R1	69%	92%	79%	12	16
R2	73%	94%	82%	17	22
R3	82%	82%	82%	17	17
R4	81%	62%	70%	21	16
R5	76%	96%	85%	23	29
R6	50%	82%	62%	11	18
R7	64%	64%	64%	14	14
R8	75%	67%	71%	9	8
R9	64%	78%	70%	9	11
R10	84%	89%	86%	18	19

Alta variância (62-86% F1). Cause provável: GT pairs varia de 9 a 23 por round — sample pequeno amplifica flutuações. Dedupe precisa de sample maior ou avaliação agregada.

R6 anomaly (62% F1): Precision caiu pra 50% — modelo predisse 18 pares mas só 11 existiam. Sobre-detecção.

---

6. Patterns Gerados

R1 (baseline): 17 patterns
R2 (baseline): 13 patterns
R3 (baseline): 16 patterns
R4 (learning): 17 patterns
R5 (learning): 15 patterns
R6 (learning): 15 patterns
R7 (learning): 15 patterns
R8 (learning): 15 patterns

R9 (learning): 18 patterns
R10 (learning): 17 patterns

Todos os rounds geraram 13-18 patterns. Exemplo de pattern do R4 (primeiro com cross-round learning):

> "*First-time contributors with maintainer label have significantly higher merge likelihood regardless of other signals*"
>
> "*Authors with 0% merge rate across 5+ prior PRs face deterministic rejection despite PR quality*"

Insight: Patterns descobrem heurísticas úteis, mas o modelo Haiku não as aplica consistentemente. O delta de +5.2pp accuracy mostra que há sinal, mas a aplicação é ruidosa.

---

7. Diagnóstico e Oportunidades de Melhoria

7.1 Problema principal: viés FP (sobre-otimismo)

185 FPs em 270 erros (68.5%). O modelo prediz merge demais.

Causas prováveis: 1. **Author merge rate domina a decisão.** FPs têm média de 99% author merge rate — o modelo assume "autor bom = PR bom" 2. **Features de qualidade do PR ausentes.** Sem LoC, sem análise de code review comments, sem CI status real 3. **Baseline do repo (~24% merge rate) não é internalizado.** Mesmo com 76% de rejeição base, modelo prediz merge

Ações recomendadas: - Adicionar `review_comments_count`, `time_to_first_review`, `ci_status_real` como features - Criar feature composta: `author_quality = author_merge_rate × pr_specifics_score` - Penalizar predições de merge quando baseline é < 30%

7.2 Feature `is_low_merge_author` subutilizada

Bruno identificou: autores com 5+ PRs e <5% merge rate são determinísticos. Mas só 11/185 FPs têm `is_low_merge_author=True`, e 73/185 têm `None`.

Ação: Garantir que `is_low_merge_author` está populado para 100% do dataset. Hoje 39% é None.

7.3 Dados faltantes degradam precisão

- `size_label`: 35% None/unknown nos FPs
- `loc_total`: maioria None

- `has_tests`: 93% None
- `ci_green`: maioria None

Ação: Enriquecer ground truth com GitHub API para LoC, CI status, test presence. Stage 1 deveria priorizar completude de features sobre novas features.

7.4 Dedupe precisa de sample maior

F1 varia 62-86% com 9-23 GT pairs por round. Instável.

Ação: Avaliar dedupe no dataset completo (não por round), ou aumentar sample para 150+ PRs.

7.5 Calibração degrada no final

R10 tem 4/6 bins overconfident. Modelo fica mais confiante sem justificativa.

Ação: Adicionar "confidence discount" quando features cruciais (LoC, CI) estão ausentes. Forçar incerteza quando dados faltam.

7.6 Category `agents` precisa de tratamento especial

34% dos FPs vêm de `agents`. É a categoria mais populosa e mais variável.

Ação: Feature `category_merge_rate` (taxa histórica por categoria). `agents` pode ter 30% merge rate vs `docs` com 60%.

---

8. Conclusão

Bootstrap v2.1 demonstra que o pipeline funciona: Haiku gera predições, aprende com patterns cross-round (+5.2pp), e identifica duplicatas. **Gates G1 e G2 passaram.**

Mas a análise revela que a melhoria está no teto do que é possível com as features atuais. O salto de 69→75% accuracy veio de patterns heurísticos, não de features ricas. **Stage 1 deve investir em completude e qualidade de features** (LoC, CI, review comments, category_merge_rate) antes de otimizar o modelo.

O viés FP estrutural (68.5% dos erros) é o principal gargalo. Resolver isso requer que o modelo internalize o baseline de ~24% merge rate do repo e pare de confiar cegamente no histórico do autor.