D-01v3 — Análise Pós-Enriquecimento: Clawkeeper Dataset

Data: 2026-02-17 10:51 BRT
Etapa: D-1 (Enrichment Complete)
Dataset: 3.233 PRs históricas | 1.521 enriquecidas (47,0% coverage)
Template: step-report.md

O que fiz

- Processo de enriquecimento (sessão `rapid-sable`) concluído. Arquivo `data/enriched_full.jsonl` gerado com **1.521 PRs únicas**, contendo campos `additions`, `deletions`, `changedFiles`, `comments`, `reviews` por PR individual.
- Análise completa: merge rates por tamanho (labels + LoC reais), cobertura Greptile, efeito top contributors, distribuição temporal.
- Dados comprometidos no repositório `~/repos/clawkeeper`.

O que encontrei

1. Merge Rate Global

Métrica	Valor
Total PRs	3.233
Merged	**780 (24,1%)**
Closed sem merge	2.414 (74,7%)
Open	39 (1,2%)

> ⚠ ***Recalibração:*** *Merge rate era 39% na amostra de 500. Com 3.233 PRs, o valor real é* ***24,1%****. O dataset completo tem proporção muito maior de PRs rejected. Isso é saudável para o modelo — mais massa negativa melhora discriminação.*

---

2. Merge Rate por Tamanho (Size Labels do GitHub)

Label	n PRs	Merged	Taxa
XS	465	143	**30,8%**
S	378	179	**47,4%**
M	206	92	**44,7%**
L	106	50	**47,2%**
XL	167	22	**13,2%**

Padrão: Sweet spot em S-M-L (44-47%). XL despenca (13,2%) — PRs grandes demais são raramente aceitas. XS intermediário (30,8%): pequenas mas talvez triviais ou mal justificadas.

---

3. Merge Rate por Linhas Reais (Enriched: additions + deletions)

Bucket (total LoC)	n PRs	Merged	Taxa	avg_add	avg_del
XS (0–10)	248	35	**14,1%**	3	1
S (11–100)	508	62	**12,2%**	39	6
M (101–500)	411	82	**20,0%**	209	19
L (501–1000)	108	32	**29,6%**	626	90
XL (1000+)	246	17	**6,9%**	33.547	14.392

Divergência com size labels: Labels (auto-calculadas pelo bot) indicam S como sweet spot, mas LoC reais apontam L (29,6%) como pico. Razão provável: XL por LoC inclui auto-gerados (i18n, lock files) que inflam linhas sem sinal real. Ambos os buckets devem entrar no modelo.

---

4. Cobertura Greptile

Categoria	n PRs	Merged	Taxa
Com Greptile review	1.125	158	**14,0%**
Sem Greptile review (enriquecidas)	396	70	**17,7%**

Achado crítico: Greptile **não prediz merge positivamente** — taxa com Greptile (14,0%) é *menor* que sem (17,7%). Possíveis explicações:

1. PRs com review Greptile são mais novas (período de alta rejeição)
2. Greptile review pode sinalizar PRs problemáticas que precisam de atenção
3. Cobertura temporal não-uniforme — Greptile ativo só recentemente

Decisão de modelagem: `has_greptile_review` entra como controle, não como predictor de qualidade. Analisar confounding temporal antes de interpretar coeficiente.

---

5. Top Contributor Comments — Efeito de Merge

Contribuidor	Comentários	PRs	Taxa Merge
steipete (maintainer)	55	44	**75,0%**
Takhoffman	57	49	**59,2%**
thewilloftheshadow	55	48	**27,1%**
HenryLoenwind	58	30	**26,7%**
akoscz	59	4	**25,0%**
mcaxtr	39	17	**17,6%**
tyler6204	54	50	**14,0%**
Glucksberg	62	55	**7,3%**

greptile-apps (bot)	505	390	**13,3%**
openclaw-barnacle (bot)	180	95	**1,1%**
clawdinator (bot)	90	89	**1,1%**
sebslight	299	297	**4,0%**

Efeito steipete — o mais forte do dataset:

- PRs **com** comentário steipete: 47 PRs → **74,5% merge**
- PRs **sem** steipete: 1.524 PRs → **13,3% merge**
- Razão de odds bruta: **~17x**

Features para o modelo:

- `has_steipete_comment` (indicador forte — 17x odds)
- `has_takhoffman_comment` (59,2%)
- `top_trusted_comment_count` (Takhoffman + thewilloftheshadow + HenryLoenwind)

---

6. Engajamento Geral (Quantidade de Comentários)

Faixa de comentários	n PRs	Merged	Taxa
0 comentários	302	44	**14,6%**
1 comentário	586	83	**14,2%**
2–3 comentários	523	74	**14,1%**
4–9 comentários	127	27	**21,3%**
10+ comentários	33	9	**27,3%**

Padrão não-linear: engajamento alto (4+) sinaliza merge. Mas o salto não é linear — o corte em 4 é importante. Criar bucket feature: `high_engagement` (≥4 comentários).

---

7. Reviews Formais (Approved/Requested Changes)

Grupo	Mediana Reviews	Média Reviews
PRs merged	**1**	**1,73**
PRs não merged	**1**	**1,31**

Merged PRs têm **32% mais reviews** em média. Feature `review_count` é relevante mas não dominante.

---

8. Distribuição Temporal (por Semana ISO)

Semana	PRs abertas	Merged	Taxa
2025-W48–W49	2	1	50,0%

2026-W02–W04	6	2	33,3%
2026-W05	283	19	**6,7%**
2026-W06	1.310	221	**16,9%**
2026-W07	1.293	353	**27,3%**
2026-W08	339	184	**54,3%**

Padrão crítico: Merge rate crescendo monotonicamente: 6,7% → 16,9% → 27,3% → 54,3%.

Duas interpretações simultâneas:

1. **Backlog clearing:** PRs antigas sendo finalmente mescladas (viés de sobrevivência)
2. **Seleção progressiva:** PRs abertas recentemente são mais selecionadas (apenas as boas chegam a abrir)

Para o modelo: `weeks_since_open` como controle é essencial. Modelo sem essa variável subestimará PRs recentes.

O que pensei

Convergência vs. divergência entre métricas de tamanho

A divergência entre size-label e LoC-real é um achado relevante para o modelo. Size labels são calculadas pelo próprio GitHub bot com lógica de peso por tipo de arquivo — são possivelmente melhores preditores que LoC bruta porque já filtram auto-gerados. Vou incluir **ambas as representações** e deixar o modelo escolher via regularização.

Greptile: predictor de risco, não de qualidade

O efeito negativo de Greptile é contraintuitivo, mas faz sentido: Greptile comenta em PRs complexas que atraem revisão automática — mas complexidade nem sempre converte em merge. É um proxy de "PR não trivial". No modelo logit, isso pode entrar como controle ou feature de complexidade.

steipete como sinal de merge mais forte do dataset

17x odds é impressionante. Mais do que qualquer variável de tamanho. Isso é capturável mas requer cuidado: `has_steipete_comment` é um signal que só aparece tarde no ciclo de vida da PR (steipete comenta quando já considera aceitar). Para o modelo "early" (features estáticas), essa feature estará ausente. Para o modelo "mature", é a feature mais poderosa.

Decisões tomadas

1. **Merge rate real = 24,1%** (não 39%). Plano atualizado.
2. **Greptile como controle** (não predictor de qualidade). Feature: `has_greptile_review`.
3. **steipete = proxy de aprovação real** — incluso no modelo mature como feature binária.
4. **`weeks_since_open`** como regressor obrigatório (viés temporal forte).
5. **Two-tier feature set confirmado:**
   - *Tier 1 (early, D-2 usável agora):* size_label, size_loc, has_greptile, pr_age_hours
   - *Tier 2 (mature):* has_steipete_comment, top_trusted_comment_count, review_count, high_engagement

Riscos

Risco	Severidade	Mitigação
47% coverage no enriquecimento	Médio	PRs não enriquecidas ficam sem features de LoC — usar size_label como fallback
Greptile confounding temporal	Alto	Controle explícito `weeks_since_open` + teste de robustez
steipete endogeneidade	Alto	Modelo early exclui steipete por design; modelo mature documenta limitação
Merge rate 24,1% vs 39% anterior	Médio	Amostra anterior enviesada. Usar 24,1% como ground truth.

Próximo passo (D-2)

1. **Feature engineering:** construir matrix (3.233 × features) para fitting do logit
2. **Dual model:** quality (nov-jan) vs. triage (fev)
3. **Weighted regression:** `weekly_pr_volume` como peso
4. **AUC + pseudo-R²:** comparar Modelo 1 (baseline, sem steipete) vs Modelo 2 (enriched)
5. **Deadline D-4 = sexta 21/02** — modelo deve estar fitted

Gerado: 2026-02-17 10:51 BRT | Fonte: enriched_full.jsonl (1.521 PRs) + all_historical_prs.json (3.233 PRs)