

Step Report: D-1 — Data Snapshot & Homework

Projeto: Clawkeeper **Dia/Etapa:** D-1 — Preparação (lição de casa) **Timestamp:** 2026-02-17T09:35:00-03:00 **Duração:** ~15 min

O que fiz

1. Criei repo `tiuitobot/clawkeeper` (public, MIT) no GitHub
2. Coletei via GitHub API (curl, não gh — gh não estava instalado):
 - 300 PRs abertas (`data/open_prs.json`)
 - 500 PRs históricas merged/closed (`data/historical_prs.json`)
 - 44 issues abertas (`data/open_issues.json`)
3. Baixei governance docs: AGENTS.md (21KB), CONTRIBUTING.md (5KB), CLAUDE.md (10 bytes — aponta pra AGENTS.md)
4. Analisei merge rates por size label, trusted-contributor, top authors
5. Instalação do `gh` CLI ficou pendente (sem sudo), Bruno instalou manualmente depois

O que encontrei

- **Merge rate geral: 39%** (194/500) — variável dependente bem distribuída pro logit
- **Size labels são preditor forte e não-linear:**
 - XS: 33% (N=180) — PRs triviais/incompletas?
 - S: 50% (N=139)
 - M: 55% (N=56) — sweet spot
 - L: 55% (N=38)
 - XL: 16% (N=62) — too big to review
- **trusted-contributor NÃO prediz merge:** 34% vs 39% non-trusted. Counter-intuitive. Label é de confiança no contributor, não na PR.
- **Top authors merged:** arosstale (16), mbelinky (11), JayMishra-source (7), Clawborn (6)
- **Clawborn:** alto volume de PRs closed sem merge — muitas tentativas, taxa mista. Interesting pattern.
- **additions/deletions:** API de listagem NÃO retorna esses campos. Precisa chamada individual por PR. Size labels servem como proxy.
- **Labels mais frequentes em merged:** size:S (70), size:XS (61), agents (60), maintainer (31)

O que pensei

- O logit vai ter poder preditivo razoável só com size + author_history + label combos. Hipótese: $\text{pseudo-}R^2 > 0.5$ é alcançável.
- `trusted-contributor` não predizer merge é um resultado que **vende o Clawkeeper** — mostra que intuições humanas sobre "quem é confiável" não são bons proxies pro que realmente é aceito.
- A curva XS→S→M→L→XL sugere que PRs muito pequenas são noise (typo fixes rejeitados?) e muito grandes são impraticáveis. Há um ótimo no meio.
- Sem additions/deletions per PR, o enriquecimento individual vai custar ~300 API calls (abertos) + ~500 (históricos). Rate limit é 5000/h, então viável em batch.

Decisões tomadas

- Usei curl em vez de gh (não disponível) — funcional, sem loss
- Salvei raw JSON em `data/` (não processado) — permite reprocessamento
- Não enriqueci PRs individuais ainda — espero gh CLI pra fazer batch eficiente
- Criei `HOMEWORK-NOTES.md` com análise resumida

Riscos / Alertas

- **500 PRs históricas pode ser pouco** pro logit se muitas features tiverem missing data. Considerar puxar mais (1000-2000).
- **Additions/deletions missing** é gap real — size labels são proxy mas não substituem LoC exato pro quality gate (1000 LoC rule).
- **Claw Trends (openclawoverview.com)** não renderizou no fetch (SPA). Preciso de browser pra analisar a concorrência direito.

Próximo passo

Enriquecer PRs com additions/deletions/files via `gh` CLI batch. Depois: estrutura do pipeline (Step 1-3 do plano).

Gerado retroativamente. Próximos reports serão via spawn Sonnet + DomainNotify em tempo real.