Step Report — Stage 0.5: Model Specification

Data: 2026-02-17 19:30 BRT **Executor:** Tiuito (Opus 4.6, main session)
Duração: ~15 min **Commit:** `99a5d36`

Objetivo

LLM atua como econometrista: lê governance docs + dados históricos do OpenClaw e propõe features repo-específicas para o modelo logit de P(merge).

Fontes analisadas

Fonte	Tamanho	Conteúdo
`data/AGENTS.md`	21KB	Estrutura do projeto, CI, coding conventions, multi-agent safety, extensões
`data/CONTRIBUTING.md`	5KB	Lista de 10 maintainers, workflow de contribuição, política de AI PRs
`all_historical_prs.json`	3233 PRs	Metadata: labels, author, additions/deletions, draft, milestone
`enriched_full.jsonl`	3233 PRs	Comments, reviews, files detalhados
`D01v3-enrichment-analysis.md`	Report	Achados estatísticos do Stage 0

Artefatos entregues

Artefato	Path	Conteúdo
Model Spec	`model_spec.json` (v0.5.0)	33 features, split early/mature, estratégia temporal dual

Artefato	Path	Conteúdo
Spec Notes	`docs/model-spec-notes.md`	Racional: insights dos governance docs, exclusões com motivo
Feature Map	`features/feature_map.json`	Mapeamento operacional: feature → regra de extração
Key-Info	`key-info.json` (K010-K013)	4 entries novas

Features especificadas

Early Model (18 features — disponíveis na criação da PR)

Feature	Tipo	Sinal esperado	Fonte
`loc_additions`	numérica	negativo após threshold	PR metadata
`loc_deletions`	numérica	positivo/ neutro	PR metadata
`loc_total`	numérica	negativo após threshold	computado
`files_changed`	numérica	negativo	PR metadata
`size_label`	categórica	não-linear (S/ M/L > XS/XL)	labels
`has_tests`	binária	positivo	file paths (*.test.ts)
`ci_green`	binária	positivo	⚠ GitHub checks API (gap)
`is_draft`	binária	negativo	PR metadata
`category`	categórica	varia por tipo	labels (agents/

Feature	Tipo	Sinal esperado	Fonte
			gateway/ cli/docs...)
`component_area`	categórica	varia	labels (channel:/ *app:*) + paths
`author_prior_prs`	numérica	positivo	computado (temporal guard)
`author_prior_merge_rate`	numérica	positivo	computado (temporal guard)
`author_association`	categórica	MEMBER > CONTRIBUTOR > NONE	enriched
`has_maintainer_label`	binária	**forte positivo (90.7%)**	labels
`has_trusted_contributor_label`	binária	positivo	labels
`has_experienced_contributor_label`	binária	positivo	labels
`weekly_pr_volume`	numérica	negativo (controle)	computado
`release_period`	categórica	varia	GitHub tags API

Mature Model (+14 features — interação acumulada)

Feature	Tipo	Sinal esperado	Fonte
`comment_count`	numérica	não-linear	enriched
`high_engagement`	binária	positivo (4+ comments)	computado
`review_count`	numérica	positivo	enriched

Feature	Tipo	Sinal esperado	Fonte
`has_approval`	binária	positivo	enriched (review state)
`has_changes_requested`	binária	negativo	enriched (review state)
`has_maintainer_comment`	binária	positivo	enriched (maintainers via CODEOWNERS/ frequência histórica)
`has_top_contributor_comment`	binária	positivo	enriched (authorAssociation)
`top_contributor_comment_count`	numérica	positivo	enriched
`has_greptile_review`	binária	fraco/ negativo (controle)	enriched
`greptile_score`	numérica	fraco/zero	enriched
`pr_age_hours`	numérica	positivo → plateau	computado
`touches_multiple_channels`	binária	negativo	labels + paths
`touches_extensions`	binária	varia	labels + paths
`is_fork_pr`	binária	negativo	⚠ GitHub API (gap)
`weeks_since_open`	numérica	positivo (confounding)	computado

Achados repo-específicos

1. **CLAUDE.md é symlink para AGENTS.md** — não são docs separados
2. **`maintainer` label = 90.7% merge** (183 PRs) — sinal mais forte por label, disponível no early model
3. **Labels `trusted-contributor` (96) e `experienced-contributor` (99)** — sinais de status do contribuidor
4. **AGENTS.md adverte explicitamente** sobre refactoring cross-channel → `touches_multiple_channels`

5. **Extensions têm regras de packaging próprias** → `touches_extensions` como sinal distinto
6. **1318 autores únicos, distribuição heavy-tail** — history de autor é esparso para maioria
7. **`requested_reviewers` quase nunca usado** (5 PRs) → excluído

Exclusões (com motivo)

Feature	Motivo
`requested_reviewers`	5 PRs apenas. Zero poder discriminativo
`milestone`	Não explorado. Deferido pra v2
`is_ai_generated`	Precisaria NLP no body. Compliance desconhecida. v2
`pr_template_compliance`	NLP comparativo com template. v2

Gaps pra resolver no Stage 1

1. **`ci_green`** — não está no dataset. Precisa GitHub checks API no ingest
2. **`is_fork_pr`** — não está no dataset. Precisa head.repo vs base.repo no ingest

Correções pós-review

1. **Abstração de identidade hardcoded (`steipete`)**
2. `has_maintainer_comment_steipete` → `has_maintainer_comment`
3. Regras em `feature_map.json` agora detectam maintainers via **CODEOWNERS** ou **frequência histórica de participação** (sem usernames fixos)

4. `has_top_contributor_comment` consolidada como sinal genérico por atividade

5. **Ajuste temporal de `weeks_since_open`**

6. Removida do **early model** para evitar proxy de engagement
7. Mantida apenas no **mature model**

8. Nota adicionada: "*Removido do early pra evitar proxy de engagement. VIF no D2.*"

9. **Cobertura de enrichment corrigida**

10. Verificação executada: `3233/3233`
11. Cobertura atualizada para **100%** em docs e report

Viabilidade

- **100% das features têm mapeamento executável** (feature_map.json)
- **0 feature sem origem identificada**
- **2 features precisam de dados adicionais** (ci_green, is_fork_pr) → resolver no Stage 1
- **Cobertura de enrichment: 100%** (3233/3233) — base completa para estimação

Próximo passo

Stage 0.7 (Bootstrap Sequencial) ou Stage 1 (Ingest PRs abertas) — parallelizáveis.

Report gerado conforme processo definido no PLAN-v4-DIFF.md.