Step Report â€” Stage 0.5: Model Specification

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

Objetivo

LLM atua como econometrista: lÃª governance docs + dados histÃ³ricos do OpenClaw e propÃµe features repo-especÃficas para o modelo logit de P(merge).

Fontes analisadas

Fonte	Tamanho	ConteÃºdo
`data/AGENTS.md`	21KB	Estrutura do projeto, CI, coding conventions, multi-ag safety, extensÃµes
`data/CONTRIBUTING.md`	5KB	Lista de 10 maintainers, workflow de contribuiÃ§Ã£o, tica de AI PRs
`all_historical_prs.json`	3233 PRs	Metadata: labels, author, additions/deletions, draft, milestone
`enriched_full.jsonl`	1521 PRs	Comments, reviews, files detalhados
`D01v3-enrichment-analysis.md`	Report	Achados estatÃsticos do Stage 0

Artefatos entregues

Artefato	Path	ConteÃºdo
Model Spec	`model_spec.json` (v0.5.0)	33 features, split early/mature, estratÃ©gia temporal dual
Spec Notes	`docs/model-spec-notes.md`	Racional: insights dos governance docs, exclusÃµes com motivo
Feature Map	`features/feature_map.json`	Mapeamento operacional: feature â†’ regra de extraÃ§Ã£o
Key-Info	`key-info.json` (K010-K013)	4 entries novas

Features especificadas

Early Model (19 features â€” disponÃveis na criaÃ§Ã£o da PR)

Feature	Tipo	Sinal esperado	Fonte
`loc_additions`	numÃ©rica		PR metadata

Feature	Tipo	Sinal esperado	Fonte
		negativo apÃ³s threshold	
`loc_deletions`	numÃ©rica	positivo/ neutro	PR metadata
`loc_total`	numÃ©rica	negativo apÃ³s threshold	computado
`files_changed`	numÃ©rica	negativo	PR metadata
`size_label`	categÃ³rica	nÃ£o-linear (S/M/L > XS/ XL)	labels
`has_tests`	binÃ¡ria	positivo	file paths (*.test.ts)
`ci_green`	binÃ¡ria	positivo	âš ï¸ GitHub checks API (gap)
`is_draft`	binÃ¡ria	negativo	PR metadata
`category`	categÃ³rica	varia por tipo	labels (agents/ gateway/ cli/docs...)
`component_area`	categÃ³rica	varia	labels (channel:/ *app:*) + paths
`author_prior_prs`	numÃ©rica	positivo	computado (temporal guard)
`author_prior_merge_rate`	numÃ©rica	positivo	computado (temporal guard)
`author_association`	categÃ³rica	MEMBER > CONTRIBUTOR > NONE	enriched
`has_maintainer_label`	binÃ¡ria	**forte positivo (90.7%)**	labels
`has_trusted_contributor_label`	binÃ¡ria	positivo	labels
`has_experienced_contributor_label`	binÃ¡ria	positivo	labels

Feature	Tipo	Sinal esperado	Fonte
`weeks_since_open`	numÃ©rica	positivo (confounding)	computado
`weekly_pr_volume`	numÃ©rica	negativo (controle)	computado
`release_period`	categÃ³rica	varia	GitHub tags API

Mature Model (+14 features â€” interaÃ§Ã£o acumulada)

Feature	Tipo	Sinal esperado	Fonte
`comment_count`	numÃ©rica	nÃ£o-linear	enriched
`high_engagement`	binÃ¡ria	positivo (4+ comments)	computado
`review_count`	numÃ©rica	positivo	enriched
`has_approval`	binÃ¡ria	positivo	enriched (review state)
`has_changes_requested`	binÃ¡ria	negativo	enriched (review state)
`has_maintainer_comment_steipete`	binÃ¡ria	**forte positivo (74.5%, 17x)**	enriched
`has_top_contributor_comment`	binÃ¡ria	positivo	enriched (authorAssociation)
`top_contributor_comment_count`	numÃ©rica	positivo	enriched
`has_greptile_review`	binÃ¡ria	fraco/ negativo (controle)	enriched
`greptile_score`	numÃ©rica	fraco/zero	enriched
`pr_age_hours`	numÃ©rica	positivo â†’ plateau	computado
`touches_multiple_channels`	binÃ¡ria	negativo	labels + paths
`touches_extensions`	binÃ¡ria	varia	labels + paths
`is_fork_pr`	binÃ¡ria	negativo	âš ï¸ GitHub API (gap)

Achados repo-especÃficos

1. **CLAUDE.md Ã© symlink para AGENTS.md** â€” nÃ£o sÃ£o docs separados
2. **`maintainer` label = 90.7% merge** (183 PRs) â€” sinal mais forte por label, disponÃvel no early model
3. **Labels `trusted-contributor` (96) e `experienced-contributor` (99)** â€” sinais de status do contribuidor
4. **AGENTS.md adverte explicitamente** sobre refactoring cross-channel â†’ `touches_multiple_channels`
5. **Extensions tÃªm regras de packaging prÃ³prias** â†’ `touches_extensions` como sinal distinto
6. **1318 autores Ãºnicos, distribuiÃ§Ã£o heavy-tail** â€” history de autor Ã© esparso para maioria
7. **`requested_reviewers` quase nunca usado** (5 PRs) â†’ excluÃdo

ExclusÃµes (com motivo)

Feature	Motivo
`requested_reviewers`	5 PRs apenas. Zero poder discriminativo
`milestone`	NÃ£o 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`** â€” nÃ£o estÃ¡ no dataset. Precisa GitHub checks API no ingest
2. **`is_fork_pr`** â€” nÃ£o estÃ¡ no dataset. Precisa head.repo vs base.repo no ingest

Viabilidade

- **100% das features tÃªm mapeamento executÃ¡vel** (feature_map.json)
- **0 feature sem origem identificada**
- **2 features precisam de dados adicionais** (ci_green, is_fork_pr) â†’ resolver no Stage 1
- **Cobertura de enrichment: 47%** (1521/3233) â€” suficiente para estimaÃ§Ã£o

PrÃ³ximo passo

Stage 0.7 (Bootstrap Sequencial) ou Stage 1 (Ingest PRs abertas) â€” parallelizÃ¡veis.

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