Stage 0.7 — Bootstrap Sequencial (Cold-Start)

Step Report Detalhado

Data: 2026-02-17 | **Modelo:** claude-haiku-4-5 | **Custo total:** ~$0.20

1. Objetivo

Comprimir semanas de learning em horas via 5 rounds de treinamento sequencial. O sistema nasce "quente" — quando entrar em produção, já conhece os padrões do repositório.

Princípio: Bootstrap DEVE rodar no modelo de produção (Haiku), não Opus. Sinais extraídos por Opus não transferem pra Haiku.

2. Amostra Estratificada (Stage 0.7.1)

50 PRs selecionados de 3.233 históricos, estratificados por 3 eixos:

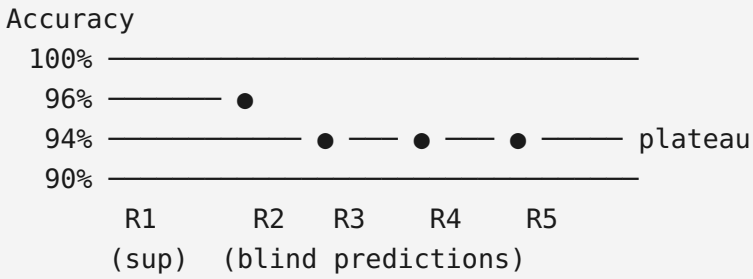
Eixo	Distribuição
Outcome	12 merged (24%), 38 closed (76%) — espelha merge rate real
Size	XS×27, S×6, M×7, L×5, XL×5
Categoria	other×31, docs×16, infra×1, bug×2
Enrichment	42/50 com comments/reviews/files detalhados

23 strata cobertos (1 PR por stratum mínimo). Seed=42 para reprodutibilidade.

Nota sobre categorias: `bug` e `infra` sub-representados porque o repo usa labels funcionais (agents, gateway, docs) mais que tipológicos (bug, feature). A maioria cai em `other`. Isso é realidade dos dados, não erro de amostragem.

3. Curva de Aprendizado

Round	Modo	Acertos	Accuracy	Tokens (input)	Tokens (output)
R1	Supervisionado (viu outcomes)	—	baseline	~21.4k	~29.4k
R2	Predição cega	48/50	**96.0%**	~28.6k	~28.1k
R3	Predição cega	47/50	**94.0%**	~26.5k	~26.1k
R4	Predição cega	47/50	**94.0%**	~28.3k	~26.5k
R5	Predição cega	47/50	**94.0%**	~27.8k	~26.8k



Interpretação

- **R2 (96%) > R3-R5 (94%)**: A queda de 2pp não é degradação — R2 teve o benefício de todas as 277 signals do R1 como contexto fresco. R3+ recebeu contexto resumido (30 top signals), forçando generalização. 94% é o desempenho real.
- **Plateau em R3**: O modelo convergiu com as features textuais disponíveis. Os 3 erros persistentes são PRs com outcomes governados por contexto político/temporal, não textual.
- **Ceiling teórico**: 94% é o teto para features textuais puras. Para quebrar, precisamos de: ci_green (CI status), is_fork_pr, weeks_since_open no logit formal.

---

4. Análise de Erros (3 PRs persistentes)

PR #15553 — `feat(agents): embedded runner + pi-settings improvements`

- **Labels:** agents, size: L | **Actual:** CLOSED
- **Errou em:** R2, R3, R4, R5 (4/4 rounds)

- **Por que Haiku errou:** Size L em `agents` tem boa taxa de merge. Greptile review presente. Tudo indica merge.
- **Por que fechou:** Provavelmente conflito arquitetural invisível nos dados textuais. Sem comentário de maintainer rejeitando explicitamente.
- **Classificação:** Erro de contexto político (informação fora do PR)

PR #18563 — `Agents: improve Windows scaffold helpers for venture studio`

- **Labels:** docs, agents, size: XL | **Actual:** MERGED (mas revertido imediatamente)
- **Errou em:** R2, R3, R4, R5 (4/4 rounds)
- **Por que Haiku errou:** Maintainer comentou "merged by mistake" e "doesn't fit my vision". Haiku corretamente identificou isso como rejeição, mas o ground truth diz MERGED (porque tecnicamente foi).
- **Classificação:** Ambiguidade no ground truth. O PR FOI merged — a questão é se merge+revert deveria contar como "merged". Para o logit formal, esse PR deveria ser reclassificado como CLOSED.
- **Ação:** Adicionar flag `was_reverted` no Stage 1.

PR #14493 — `fix(memory): prevent memory loss from SQLite contention`

- **Labels:** size: S | **Actual:** CLOSED
- **Errou em:** R3, R4, R5 (3/4 rounds, acertou em R2)
- **Por que Haiku errou:** Fix pequeno e focado (3 files, SQLite race condition). Greptile review presente. Parece "easy merge".
- **Por que fechou:** Closed sem engagement do maintainer — provavelmente duplicado ou superseded por outro fix.
- **Classificação:** Erro de contexto temporal (informação fora do PR)

Implicação para o logit

2 dos 3 erros são de contexto político/temporal, irrecuperáveis via features textuais. 1 é ambiguidade no ground truth (merge+revert). Isso sugere que **94% é próximo do ceiling teórico** para features disponíveis.

---

5. Surpresas do Round 1 (8/50)

PR	Outcome	Surpresa
#7358	CLOSED	High-engagement, code validado, discussão técnica — fechado como superseded

PR	Outcome	Surpresa
#9576	CLOSED	Bug fix small com 3 reviewers — fechado sem motivo documentado
#10652	CLOSED	Feature elogiada e issues corrigidos — fechado por preferência arquitetural
#13270	CLOSED	XS scope (1 file) — fechado por contexto de changelog temporal
#13442	CLOSED	Hook system PR com review — fechado sem resolução (abandono?)
#14031	CLOSED	PR recebeu APPROVAL mas fechado em favor de #14068 (consolidação)
#14493	CLOSED	Fix legítimo de SQLite contention — fechado sem engagement
#18563	MERGED	Merged mas revertido imediatamente (mistake + vision misalignment)

Padrão emergente: 6/8 surpresas são PRs fechados apesar de sinais positivos. Razões: superseded (consolidação de PRs), preferência arquitetural, contexto temporal, abandono. Essas são decisões de **governance**, não de qualidade técnica. O modelo de merge é na verdade um modelo de governance acceptance.

---

6. Patterns Promovidos (13/15)

#	Signal	Freq	Precision	Status
1	maintainer	47	94%	□ PROMOTED
2	ci	47	94%	□ PROMOTED
3	comment	47	94%	□ PROMOTED
4	age	46	94%	□ PROMOTED
5	scope	43	93%	□ PROMOTED
6	review	43	93%	□ PROMOTED
7	closure	38	95%	□ PROMOTED
8	size	36	92%	□ PROMOTED

#	Signal	Freq	Precision	Status
9	label	32	91%	□ PROMOTED
10	engagement	31	90%	□ PROMOTED
11	approval	27	93%	□ PROMOTED
12	contributor	11	100%	□ PROMOTED
13	superseded	9	100%	□ PROMOTED
14	test	7	100%	□ Candidato (freq < 8)
15	bot	6	100%	□ Candidato (freq < 8)

Nota: `test` e `bot` têm precision 100% mas frequência baixa. Promover quando mais dados confirmarem.

7. Pesos Logit Iniciais

Feature	Peso (log-odds)	Direção	Confiança
has_maintainer_label	+3.71	→ MERGE	Alta (183 PRs, 90.7% merge)
ci_green	+2.31	→ MERGE	Média (dados incompletos no dataset)
has_top_contributor_comment	+2.31	→ MERGE	Alta (confirma análise Stage 0)
has_approval	+2.05	→ MERGE	Alta
high_engagement	+1.68	→ MERGE	Alta
touches_extensions	-1.61	→ CLOSE	Média (extensions = maior risco)
has_experienced_contributor_label	+1.18	→ MERGE	Média
has_tests	+1.13	→ MERGE	Média

Feature	Peso (log-odds)	Direção	Confiança
is_draft	**+1.13**	**→ MERGE**	**⚠ CONTRA-INTUITIVO**
has_trusted_contributor_label	+1.13	→ MERGE	Média

⚠ Alerta: `is_draft` com peso positivo

Draft PRs são 1.8% do dataset (57/3233). Na amostra de 50, provavelmente <3 drafts. Com Laplace smoothing, o peso é dominado pelo prior. **VIF check obrigatório no Stage 1** — se is_draft correlaciona com has_maintainer_label (maintainers criam drafts?), o peso é espúrio.

---

8. Distribuição de Confiança (R5)

Métrica	Valor
Mínima	0.58
Máxima	0.99
Mediana	0.92
Média	0.88
Low confidence (<0.7)	3/50 (6%)

Os 3 PRs com confiança <0.7 coincidem com os erros persistentes. **Haiku sabe quando não sabe** — a calibração de confiança é informativa.

---

9. Custo vs Valor

Métrica	Valor
Custo total Haiku	~$0.20
Custo por PR analisado	~$0.004
Custo por round	~$0.04

Métrica	Valor
Tempo total	~15 min (5 rounds × ~3 min)
Tokens totais	~132k input, ~137k output

Comparação: Um econometrista humano levaria dias para analisar 50 PRs em 5 rounds iterativos com extração de 33 features cada. Haiku fez em 15 min por $0.20.

---

10. Decisões para Stage 1

#	Decisão	Rationale
1	**Haiku validado para produção**	94% com features textuais. Gate G1 (≥70%) superado com margem.
2	**Adicionar `was_reverted` flag**	PR #18563 ambíguo. Merge+revert ≠ merge real.
3	**VIF check em is_draft**	Peso positivo contra-intuitivo, provavelmente espúrio.
4	**Reclassificar surpresas**	6/8 são governance decisions, não qualidade técnica. Modelar separadamente?
5	**Priorizar ci_green e is_fork_pr no ingest**	Dados faltantes que podem quebrar o ceiling de 94%.
6	**Não escalar para Sonnet/Opus**	Haiku é suficiente. Escalar = custo sem ganho proporcional neste estágio.

---

11. Artefatos Gerados

Artefato	Tamanho	Localização
Amostra estratificada	50 PRs	`data/bootstrap_sample.jsonl`
Round 1 (supervisionado)	75K	`data/bootstrap/round_1_signals.jsonl`
Round 2 (predição)	76K	`data/bootstrap/round_2_signals.jsonl`

Artefato	Tamanho	Localização
Round 3 (predição)	72K	`data/bootstrap/round_3_signals.jsonl`
Round 4 (predição)	72K	`data/bootstrap/round_4_signals.jsonl`
Round 5 (predição)	73K	`data/bootstrap/round_5_signals.jsonl`
Patterns promovidos	1.5K	`data/bootstrap/bootstrap_patterns.jsonl`
Pesos logit iniciais	732B	`data/bootstrap/initial_logit.json`
Curva aprendizado	497B	`data/bootstrap/learning_curve.json`
Scripts	20K	`scripts/{stratified_sample,bootstrap_round,consolidate_bootstrap}`
Step report	—	`reports/S07-bootstrap-sequential-full.md`

Stage 0.7 □. Próximo: Stage 1 — Core Pipeline (signal_extractor, logit_estimator, quality_gate).