ClinicalAI-Trial Analysis — PRD

ClinicalAI-Trial Analysis

Author: Product Team **Date:** 2026-07-27 **Status:** Draft — Solution Review (Stage 4)

Rebuilt from v1.0 to close the gaps identified in the PRD evaluation (21 blockers, 9 improvements). Resolved decisions: model = Anthropic API direct; MOAT disclosed honestly as a strategic risk, not a real moat; problem validation pulled live from the ClinicalTrials.gov API; pricing/revenue included as directional placeholders only.

---

1. Hypothesis

We believe that returning ranked, source-linked, epistemically-labeled failure hypotheses from a single NCT ID will let biotech analysts and competitive-intelligence researchers replace a 30–60 minute manual cross-reference task with a **<20-second automated read**, measured by analyses completed and P95 time-to-result over the first 90 days post-launch.

---

2. Problem

Who: Primary persona — a biotech equity analyst or competitive-intelligence researcher at a fund, bank, or CI team, doing pipeline diligence on a sponsor's clinical programs for an investment, licensing, or competitive read. (Secondary, lower-priority users: academic/regulatory researchers doing landscape review, and science journalists covering drug development — same workflow, lower initial focus.)

How bad, with real data (pulled live from the ClinicalTrials.gov v2 API, 2026-07-27):

- 34,023 trials are currently registered as **TERMINATED** (5.7% of all 595,847 registered studies); combined with WITHDRAWN (16,591) and SUSPENDED (1,748), **52,362 trials (8.8%)** never reached completion as planned.
- In a random sample of 200 TERMINATED trials pulled live via the API: **8.0% have no `whyStopped` value at all**, and another **3.5% cite only a generic "business decision" or "sponsor decision" with no underlying cause — 11.5% combined give literally zero substantive signal.**
- The remaining 88.5% do name *some* cause, but most are single-line labels ("low enrollment," "Covid-19," "increased LFTs") that don't explain the deeper *why* (competitive landscape, site-level

issues, disease rarity, a specific safety signal's severity) without cross-referencing enrollment data, results, and publication history. **This — not just the blank-field case — is the actual gap this product targets: turning a terse or absent stop reason into a structured, source-grounded, honestly-labeled explanation.**

- `[NEED: direct user validation — no analyst/investor interviews or surveys have been conducted yet. The above is registry-level evidence that the data problem is real; it is not yet evidence that the target persona wants this specific solution. Run 5–8 informal interviews with biotech analysts before Solution Review sign-off if this PRD needs to justify spend beyond a solo portfolio build.]`

Current workarounds: Manually cross-referencing ClinicalTrials.gov, PubMed, sponsor press releases, and (for public sponsors) SEC filings, stitched together in spreadsheets or memos — inferred from how CI/analyst workflows are commonly described, not directly observed for this project.

If we don't solve it: No existing tool does causal reasoning over public trial data. ClinicalTrials.gov's own search and the AACT database expose raw fields with no reasoning; PubMed requires manual cross-referencing; Citeline/Trialtrove and GlobalData are comprehensive but enterprise-priced and still largely descriptive rather than causal; generic LLM chat has no structured source-grounding and no fact/inference/hypothesis discipline, so it's prone to inventing specifics with no citation trail.

---

3. Strategic Fit

Why now: ClinicalTrials.gov's v2 API is the only supported version (v1 was fully retired), making structured, reliable data access straightforward for the first time. LLM structured-output reliability has reached the point where a strict, schema-enforced fact/inference/hypothesis contract is achievable without fine-tuning.

Why this approach vs. alternatives considered:

- *Rules-based keyword classification of* `whyStopped` *instead of an LLM* — rejected. Stop reasons and publication abstracts are free-text and highly variable in phrasing ("insufficient population" vs. "low enrollment" vs. "recruitment challenges" all mean the same thing); a keyword/rules engine would need constant manual expansion and would still miss reasoning across multiple weak signals (e.g., enrollment shortfall *combined with* a same-year competitor approval). An LLM's semantic judgment handles this variability and cross-signal reasoning natively.
- *A browser extension overlaying ClinicalTrials.gov* instead of a standalone tool — rejected for MVP. A standalone page gives full control over the four-block labeled-output UX, which is the actual product differentiator; an overlay would fight the host page's layout for that.

Competitive context: ClinicalTrials.gov / AACT (raw data, no reasoning), PubMed (fragmented, manual), Citeline/Trialtrove and GlobalData (enterprise-priced, descriptive not causal), generic LLM chat

(no source-grounding or labeling discipline). None combine free public-data retrieval with explicit epistemic labeling.

MOAT — disclosed honestly: There is **no proprietary data advantage, lock-in, or fine-tuning moat at MVP.** The fact/inference/hypothesis labeling discipline is a trust/UX differentiator today, not a structural moat — a well-resourced competitor (Citeline, GlobalData, or a new entrant) could replicate the same labeling pattern. This is an accepted strategic risk for the MVP stage, not a claimed advantage. The one plausible path to a real moat: if the evaluation benchmark set (Section 6, "AI Quality Bar") grows into a proprietary dataset of trials with validated true failure causes, that dataset — not the labeling UX — becomes the actual defensible asset over time.

Model selection (resolved): Anthropic API, direct. Criteria: (1) reliable structured/JSON output for the strict evidence schema, (2) sufficient context window to hold a full trial record plus multiple PubMed abstracts in one call, (3) cost per analysis low enough to sustain a free public tool, (4) low latency to hit the <20s P95 target. Rejected for MVP: AWS Bedrock — adds IAM/provisioning overhead with no MVP-stage benefit for a solo build; revisit only if a future enterprise tier requires AWS-native integration.

Cost/accuracy tradeoff: A single larger-context call per analysis (full trial record + linked abstracts in one pass) costs more per run than a cheaper/smaller-context model, but avoids the accuracy loss of truncating source evidence. Accepted given low expected MVP volume (hundreds of analyses/day).

Market sizing *(directional estimate, not sourced from third-party market data — flag before using externally)*:

- TAM: order-of-magnitude tens of thousands of biotech equity analysts, CI professionals, and academic/regulatory researchers globally who regularly consult ClinicalTrials.gov. `[NEED: a sourced estimate — e.g., BIO/industry association membership figures or a LinkedIn role-title search — before this number is used in any external-facing document.]`
- SAM: a meaningfully smaller subset who track *specific* trial terminations for investment or CI purposes rather than general research — directionally low thousands.

Development cost: Manpower — ~1 FTE-equivalent solo effort across 8 weeks (opportunity cost only, no cash cost). Infrastructure — ~$20–50/month at MVP traffic (Vercel hobby tier + LLM API usage), scaling roughly linearly with analysis volume.

Directional pricing (future, unvalidated) *(explicitly a placeholder, not a commitment — v1 is free; Non-Goals below excludes billing at launch)*: a future batch/pipeline-report tier priced around **$99–199/month per seat**, positioned well below enterprise platforms like Citeline (which run into five figures annually). Revenue scenarios if pursued: conservative — 50 paying CI teams at $99/month ≈ $59K ARR; target — 200 paying teams at $99/month ≈ $238K ARR. Both figures are directional math, not forecasts — they exist to sanity-check whether the future tier is worth building, not to plan around.

---

4. Solution

What we're building: A stateless web app with a single text input (NCT ID). On submit, the backend fetches and normalizes the trial record, resolves linked PubMed publications, assembles an evidence packet, and runs a structured LLM reasoning pass that returns strict JSON matching a fact/inference/hypothesis schema.

User flow:

```
User pastes NCT ID
  → [System] validates format (NCT + 8 digits) client-side
  → [System] fetches ClinicalTrials.gov v2 record; rejects with "trial not found" if the ID is well-formed but doesn't exist
  → [System] resolves linked PubMed publications via NCBI E-utilities; flags if none exist
  → [System] runs the labeled reasoning pass (taxonomy classification + fact/inference/hypothesis contract + counter-argument requirement)
  → User sees four blocks: Bottom Line, Ranked Hypotheses (confidence + evidence-for + strongest counter-argument), Evidence (source-linked), Guardrail (what's known vs. inferred vs. speculative)
```

Key interactions and states:

- Malformed ID → rejected client-side before any network call, no wasted round-trip.
- Well-formed but nonexistent ID → explicit "trial not found" message.
- Trial exists but `overallStatus` is not TERMINATED/WITHDRAWN/SUSPENDED → no failure hypotheses generated; the tool states plainly that no termination occurred (prevents a fabricated failure narrative on a completed/ongoing trial — see US-4 below).
- API or LLM call fails → retry once, then show a retry-able error state; never render partially-parsed or unlabeled output.
- PubMed enrichment fails → still return a trial-data-only analysis with a note, rather than failing the whole request.

System behavior — autonomy and review: The system operates **fully autonomously with no human-in-the-loop review** before an analysis is shown to the user. This is a stated design choice, not a gap: correctness is enforced entirely through schema validation and the labeling contract, given the free, stateless, high-volume nature of the product. The Guardrail block and persistent disclaimer are the disclosed fallback for cases where the model's confidence is misplaced.

Trust & safety (Responsible AI), addressed explicitly across all four pillars:

- *Transparency* — every claim across all blocks is tagged Fact / Inference / Hypothesis, enforced by schema validation, not just prompt instruction.
- *Reliability* — retry-then-error on technical failure; graceful degradation when enrichment fails; never render unlabeled or partially-parsed output.

- *Accountability* — the tool is advisory-only; there is no human sign-off per analysis, so the user is the accountable party for any decision made from the output. This trade-off is disclosed via a persistent, non-collapsible disclaimer on every results page ("not medical or investment advice").
- *Fairness* — reasoning quality depends on source-data completeness. Trials from smaller or academic sponsors (which tend to report less complete data) may get systematically thinner or lower-confidence analyses than large industry sponsors. Monitored via the benchmark set (Section 6) stratified by sponsor class.
- *Bias (distinct from hallucination)* — the model could pattern-match a failure cause to sponsor size/reputation rather than the specific evidence in front of it. Mitigation: the prompt requires every hypothesis to cite specific evidence from *that trial's own record*; the benchmark set is periodically checked for systematic skew by sponsor class or phase.

Prompting approach: Structured extraction against an explicit 7-category failure taxonomy (recruitment, efficacy, safety/toxicity, funding/business, operational/protocol, strategic/competitive, futility), with few-shot examples of correctly labeled Fact/Inference/Hypothesis output and a hidden chain-of-thought scratchpad (not shown to the user) to reason through evidence before committing to the final ranked JSON.

AI behavior examples *(representative subset — the full 15–25 example set used for prompt development and regression testing lives in the benchmark spreadsheet, Section 6; these illustrate the range)*:

#	Input signal	Expected output behavior
1	`whyStopped`: "Discontinued prematurely due to low enrollment"; actual 34/280 enrolled	Fact: exact enrollment shortfall cited. Inference: low enrollment likely drove the decision. Hypothesis: possible competing trial/standard-of-care shift — flagged only if supporting evidence exists (e.g., a same-indication trial approved nearby in time).
2	`whyStopped` blank; no results section; no linked publications	Bottom Line explicitly states evidence is insufficient for a confident hypothesis; Guardrail leads with "no stop reason was reported and no results were published."
3	`whyStopped`: "Business decision" (generic); results section shows primary endpoint met	Hypothesis: efficacy was *not* the driver (endpoint met) — points toward funding/strategic/business reasons; counter-argument explicitly notes "business decision" is unverifiable without external sourcing.
4	Results section shows adverse events flagged; `whyStopped` mentions "safety"	Fact: specific adverse event data cited. Classified under safety/toxicity with high confidence; counter-argument considers whether AE rate was actually above background.
5	`overallStatus` = COMPLETED (not terminated)	No failure hypotheses generated; Bottom Line states the trial completed as planned.
6	`overallStatus` = RECRUITING (ongoing)	Same as above — no failure narrative; states the trial is still active.

#	Input signal	Expected output behavior
7	Well-formed ID (`NCT` + 8 digits) that doesn't exist in the registry	"Trial not found" error state; zero LLM calls made.
8	Malformed ID (wrong prefix/length)	Rejected client-side before any network call.
9	`whyStopped` mentions "interim analysis... futility"	Classified under futility with high confidence; distinguished from a generic efficacy miss in the Bottom Line language.
10	No linked PubMed publication for a trial with posted results	Evidence block explicitly flags "no publication found for this trial's results" as a Fact, not silently omitted.

Prompt improvement plan: The reasoning prompt is maintained as a versioned file. Any trial a user flags as low-quality (via a lightweight feedback control on the results page) is logged with its NCT ID into a review queue, checked weekly against the benchmark set before any prompt revision ships.

---

5. Non-Goals

- We are **NOT** building user accounts, saved history, or personalization in v1 — the product is fully stateless.
- We are **NOT** supporting batch or multi-trial analysis in v1 — one NCT ID per request, to keep the MVP's reasoning quality bar achievable before scaling scope.
- We are **NOT** integrating any data source beyond ClinicalTrials.gov and PubMed in v1 (no SEC filings, press releases, or paid trial databases) — reserved for a future phase only if the free-source version proves the hypothesis-quality bar is achievable.
- We are **NOT** accepting file uploads of proprietary/non-public trial data — public-registry-only keeps the product legally simple and stateless.
- We are **NOT** building billing, accounts, or a paid tier at launch — pricing (Section 3) is directional only and explicitly deferred pending MVP validation.
- We are **NOT** shipping a native mobile app — responsive web only.

---

6. Success Metrics

North Star metric: Analyses completed with zero unlabeled causal claims — this single metric captures both usage and the product's core credibility promise (a labeled-but-unused product or an unlabeled-but-used one both fail the actual goal).

Primary metrics:

Metric	Baseline	Target (90 days post-launch)
Analyses completed	0	1,000
P95 time-to-result	N/A	< 20 seconds

Secondary metrics:

- % of analyses on terminated/withdrawn/suspended trials (core use case): target ≥ 70%
- Return-user rate (analyzed ≥ 2 distinct NCT IDs): target ≥ 20%

Guardrail metrics (must not get worse):

- Analysis error rate (invalid ID, API failure, malformed LLM output): baseline N/A, must stay **< 3%**
- % of rendered claims missing a Fact/Inference/Hypothesis label: must stay at **0%** — this is a hard schema-enforced gate, not a soft target.

AI Quality Bar (evaluation plan):

- **Ground truth set:** a 30–50 trial benchmark of TERMINATED trials with well-documented, high-confidence stop reasons (explicit `whyStopped` plus corroborating published post-mortem or news coverage), built before Week 5 (Reasoning engine complete) so that milestone has a real acceptance bar, not just schema validity.
- **Accuracy threshold:** top-ranked hypothesis matches the documented real reason in **≥ 70%** of benchmark-set trials; 100% of claims carry a Fact/Inference/Hypothesis label (hard requirement, restated here as the Honest/Helpful bar).
- **Eval tracking:** a spreadsheet with columns — NCT ID, run date, prompt version, top hypothesis, documented true reason (if known), reviewer plausibility score (1–5), notes.
- **Pre-ship experimentation:** before promoting a revised prompt/taxonomy to production, run it against the full benchmark set and compare plausibility scores against the current production prompt; ship only if scores are equal or better.
- **Component-level quality bar:** ≥ 80% of hypotheses for well-documented trials (has `whyStopped` + a results section) rated "plausible" by a domain-knowledgeable reviewer in manual QA; trials with sparse source data are expected to produce lower-confidence output, not false precision.

---

7. Rollout Plan

Gradual rollout with two gated checkpoints, not a big-bang launch:

Stage	Timing	Scope	Go/no-go criteria
Data + enrichment	Weeks 1–3	Internal only	Correctly normalizes 20+ real terminated-trial IDs across different data-completeness levels; PubMed linkage correctly resolves for

Stage	Timing	Scope	Go/no-go criteria
layer			known-published trials and correctly flags unpublished ones.
Reasoning engine	Weeks 4–5	Internal only	100% of benchmark-set analyses pass schema validation with zero unlabeled claims; top-ranked hypothesis hits the ≥70% accuracy bar (Section 6) against the benchmark set.
Private beta	Week 7	~10 target users (biotech-focused contacts, relevant online communities)	Beta users can complete an analysis end-to-end without confusion about the labeling scheme; no critical mislabeling incidents reported. This turns the earlier GTM "beta" mention into an actual phased checkpoint, retiring the risk that public users misinterpret hypotheses as confirmed fact (Section 4, Fairness/Accountability).
Public launch	Week 8	Public	Live, error rate < 3%, all guardrail metrics holding.
Post-launch iteration	Weeks 9–12	Public	Incorporate beta/launch feedback on hypothesis quality; evaluate promoting P1 items (evidence-packet caching, shareable permalinks) based on usage; decide whether the directional paid batch-analysis tier (Section 3) is worth pursuing based on real return-user data.

Rollback plan: The product is stateless with no data migrations, so "rollback" means reverting to the previous prompt/taxonomy version (kept in version control per Section 4's prompt improvement plan) or, in the case of a critical labeling failure, taking the public site down until fixed — there is no user data to preserve or migrate.

Post-launch monitoring: Monthly, re-run the benchmark set against production and flag any drop in plausibility score > 10% for prompt review. This is in addition to (not a replacement for) the existing operational monitoring for ClinicalTrials.gov/NCBI API deprecation and per-analysis LLM cost.

---

8. Open Questions

- `[NEED: informal validation interviews]` — 5–8 conversations with biotech analysts/CI researchers to confirm the *solution*, not just the data problem, resonates. Owner: solo builder. Target: before public launch (Week 8).
- Is a shareable permalink per NCT ID (P1/P2) worth pulling into the public-launch scope given it plausibly aids organic sharing/distribution, or does it stay in the Week 9–12 iteration phase as currently scoped?
- Should lightweight per-IP rate limiting ship at public launch to bound LLM cost exposure, given there are no accounts to gate on? Current default: no, monitor cost in the first weeks and add only if needed.

- `[NEED: sourced TAM/SAM figures]` — the Section 3 market-size estimate is directional only; needs a real source (industry association membership, LinkedIn role-title search) before it's used outside this internal document.
- The Section 3 pricing/revenue figures ($99–199/month tier; conservative/target ARR scenarios) are directional math with no validation behind them — before treating them as anything more than a sanity check, anchor them to real signal (e.g., observed return-user rate or explicit inbound interest post-launch).