

# The \$1,032 Research Program: What Fifty Billion Tokens Taught Me About AI-Accelerated Science

Rowan Brad Quni-Gudzinis

2026-08-20

Rowan Brad Quni-Gudzinis — August 2026

*Disclosure and method: this essay was drafted with AI assistance under my direction. Every figure traces to a primary record — the platform’s own usage exports for the token ledger, and the public registries (Zenodo, ORCID, GitHub, the internal knowledge graph, the Cloudflare account) for the output inventory. Nothing here is estimated from memory.*

Between December 1, 2025 and August 20, 2026 — 263 days — I ran a personal research program on **\$1,032.08** of DeepSeek API credits. The program spans ultrametric physics, information theory, laws of form, quantum error correction, and the epistemology of AI-assisted research itself. There is no lab, no grant, no institutional budget. The research staff is me, and the machine. This is the ledger of what the machine did, what it cost, and what it taught me about how science gets done when the marginal cost of a research operation falls by orders of magnitude.

## 1. The ledger: fifty billion tokens

I exported the platform’s usage records for every month of the program (ten archives, November 2025 through August 20, 2026; the November archive contains no billed usage — the key was provisioned and real work began on December 1). Aggregating the per-day cost records and the per-token-type amount records:

### Monthly spend and volume

| Month   | Cost (USD) | Requests | Input tokens | Output tokens |
|---------|------------|----------|--------------|---------------|
| 2025-12 | \$8.89     | 1.0K     | 34.5M        | 3.3M          |
| 2026-01 | \$7.95     | 962      | 30.5M        | 3.2M          |
| 2026-02 | \$0.59     | 127      | 2.5M         | 0.3M          |
| 2026-03 | \$3.22     | 359      | 10.2M        | 1.2M          |
| 2026-04 | \$24.11    | 7.7K     | 483.9M       | 6.2M          |
| 2026-05 | \$110.74   | 50.5K    | 6.92B        | 33.8M         |

| Month          | Cost (USD)        | Requests      | Input tokens  | Output tokens |
|----------------|-------------------|---------------|---------------|---------------|
| 2026-06        | \$108.68          | 45.6K         | 4.87B         | 26.3M         |
| 2026-07        | \$440.16          | 124.3K        | 14.40B        | 76.5M         |
| 2026-08 (1–20) | \$327.73          | 92.7K         | 23.12B        | 97.6M         |
| <b>Total</b>   | <b>\$1,032.08</b> | <b>323.4K</b> | <b>49.86B</b> | <b>248.4M</b> |

*Volumes rounded to display precision; the grand totals below are exact.*

By model, the same window:

| Model                    | Cost (USD) | Requests | Input tokens | Output tokens |
|--------------------------|------------|----------|--------------|---------------|
| deepseek-v4-pro          | \$818.97   | 235.0K   | 30.69B       | 152.1M        |
| deepseek-v4-flash        | \$173.21   | 79.0K    | 18.70B       | 83.3M         |
| chat + reasoner (legacy) | \$39.91    | 9.4K     | 474.3M       | 12.9M         |

Two regimes are visible in that table. December through March was single-threaded conversation on the legacy chat and reasoner models — about \$21 across four months. April onward is the agentic regime: subagents, parallel exploration, adversarial review, and prompt caching as the economic engine. July was the peak month — \$440.16, 124.3K requests (one request every ~11 seconds across a 12-hour working day) — driven by multi-agent red-team orchestration. August is running hotter per day, with the flash model carrying a larger share of the load: a deliberate cost-discipline shift.

#### Grand totals (December 1, 2025 – August 20, 2026):

- **50.11B tokens:** 49.86B input + 248.4M output
- **323,381 API requests** — on average one research operation every ~35 seconds across a 12-hour working day, every day of the window
- **\$1,032.08 total** across 184 billed days: a mean of \$5.61 per billed day, a peak day of \$61.91 (July 16), a peak week of \$173.22 (July 12–18)
- **Blended cost: \$20.60 per billion tokens**

The cache arithmetic is the whole economic story, so it deserves to be stated plainly. **96.4% of input tokens (48.08B of 49.86B) were cache hits:** the system re-reading its own context — canonical prompt stores, registries, plan state — at \$0.0036 per million tokens instead of the \$0.435 miss price. The agentic workflow is inherently a repeated-reading workflow; a 96% hit rate is what turns a stateful research engine from a fantasy into a five-dollar-a-day line item. (Prices from the usage records: v4-pro miss \$0.435/1M, hit \$0.0036/1M, output \$0.87/1M; v4-flash miss \$0.14/1M, hit \$0.0028/1M, output \$0.28/1M.)

## 2. What the tokens produced

The corpus, as it stands in August 2026:

- **911 versioned records on Zenodo** under my name, spanning roughly **880 distinct works** (ORCID registers 877 works with versions folded)
- **1,646 paper nodes in the internal knowledge graph** (8,308 nodes total), including drafts and records that predate the window

- **114 GitHub repositories** in the QNFO organization — 109 of them created inside the window
- **20 production Cloudflare Workers** running the platform: paper indexing, semantic search, durable memory, email outreach, lifecycle auditing
- Live public surfaces: [papers.qnfo.org](https://papers.qnfo.org), [qnfo.org](https://qnfo.org), [qwav.org](https://qwav.org), and a set of interactive demos that execute published results in a browser

The methodological core of the program was itself published inside the window: the *Universal Ignorance Audit* (DOI 10.5281/zenodo.21901984) and its companion case study, *Knowing What We Do Not Know* (DOI 10.5281/zenodo.21901983) — a fifteen-question method for auditing what you do not know, plus the pipeline’s own account of how an AI-assisted research pipeline learns from its failure history. A series of program records followed in August 2026, including the *Tyranny of the  $\pm 1$*  essay (DOI 10.5281/zenodo.21939595), which turned the pipeline’s red-team audits into public method.

I will not claim all 911 records are good. They are not — the drafts number in the hundreds, and the pipeline’s own audits have produced hard findings against its own published work (see §4). The honest claim is narrower: the pipeline produced a *verifiable, versioned, attributable* research corpus at a rate no single human could match, and it built the machinery to check that corpus against the literature, against itself, and against the outside world.

### 3. The comparison: what conventional research costs

The baseline, stated as assumptions so the arithmetic can be re-run: a US postdoctoral researcher earns \$50,000–\$70,000 per year — the NIH NRSA postdoctoral minimum was \$50,004 in 2019 [verified via live source], and institutional minima have risen since — and institutions typically budget 1.8–2.2× salary for benefits, overhead, and administration [assumption; standard budgeting practice]. One postdoc, fully loaded: roughly \$90,000–\$155,000 per year.

The measured window is 8.7 months. One postdoc for that period: **about \$65,000–\$112,000**. The machine bill: **\$1,032.08**.

**Cost ratio: ~60× to ~110×** — call it two orders of magnitude. In terms that cut through multiplication: the program’s ~880 registered works cost **about \$1.18 each** in machine time.

Human-equivalent volumes, stated with their conversion factors so they cannot be inflated: the 248.4M output tokens equal roughly 186 million words (at 0.75 words per token) — enough raw generation to occupy a fast typist for about 39 work-years (at 40 words per minute, 2,000 hours per year). The genuinely new text read — cache misses only — is 1.79B tokens, roughly 1.34 billion words: about 45 work-years of continuous reading at 250 words per minute. I state these numbers because they make the scaling concrete, and I flag them because volume is not value: the overwhelming share of that output is intermediate material — code, metadata, tool payloads, discarded drafts — not polished prose.

## 4. What I learned

**1. Production became cheap; judgment became the entire job.** The machine drafts, codes, compiles, deploys, and checks in bulk. What it cannot do is decide what is true, what matters, and what deserves to survive. Every serious session ends the same way: I hold a candidate output and make a kill decision. The pipeline's real inventions were the instruments of that decision — citation audits that verify every reference against live registries, adversarial reviewers that are instructed to attack, publication gates that block known failure modes.

**2. AI errors are systematic, not random — so defenses must be gates, not intentions.** Across nine months the same failure classes recurred: fabricated citations (a hallucinated author list is a research-integrity violation, and it happened), internal reference keys leaking into published bodies, character-encoding corruption propagating from a single encoding slip into PDFs and databases, phantom DOIs, drifting configuration files, files written to the wrong storage bucket. Each class, once seen, became a hard gate in the program's rulebook: scan before commit, verify before cite, check before publish. Intentions did not work. Gates worked.

**3. The most valuable output is knowledge of the pipeline's own ignorance.** The Universal Ignorance Audit is the program's signature result, and it is a meta-result: a method for mapping what is not known, including what the machine does not know it does not know. A research system that only produced papers would be a paper mill. One that produces an audited map of its own unknowns is an instrument.

**4. Publish, then audit — every time.** The program's standing rule is that every published artifact receives an adversarial audit after publication. The first audit of the flagship pair surfaced 13 hard findings — unresolvable citations, title mismatches, missing deposits — and each finding became the next cycle's remediation item. Publication is a checkpoint in a loop, never an endpoint.

**5. Most failures were self-inflicted, and the fix was usually a rollback of my own last change.** When the email pipeline failed on every address of one domain while ten sibling domains worked, the cause was not the platform: it was my own routing rules, added days earlier. When a publication mirror landed in the wrong storage bucket, it was my script. The single most useful diagnostic habit the program taught me: ask “what did I change?” before asking “is the platform down?”

**6. Architecture is what makes tokens cheap.** Fifty billion tokens for a thousand dollars is not a pricing accident; it is an engineering result. Byte-identical prompt stores across five systems, canonical registries, a knowledge graph that makes context retrievable instead of re-explained, subagents with bounded prompts that demand evidence. The discipline is unglamorous, and it is the entire margin.

## 5. What \$1,032 did not buy

Nothing in the ledger above is external validation. Citations by other researchers were not purchasable here; neither were peer acceptance, venue acceptance, or correspondence. All of those remain priced in human currency, and all remain pending. The program's own rules require that traction signals be external only. By that standard, the efficiency claim in this essay is a supply-side claim awaiting a demand-side verdict.

The falsifiable statement, per the program’s own method: **if, by August 2028, this corpus has produced no external citations, no substantive correspondence, and no venue acceptance, then the efficiency demonstrated here was hollow — volume without a reader.** I register that kill criterion now, in advance, so that this essay is a prediction and not a retrodiction.

One further honesty: the human’s time is not in the denominator. I directed every session, wrote every kill criterion, made every publication decision. The correct sentence is not “an AI did the work of a lab.” It is: one researcher’s judgment, multiplied by a machine that never sleeps, at the cost of a laptop — with the judgment itself the irreplaceable and unpriced input.

## 6. The asymmetry

For the individual researcher, the cost structure of doing science has changed by two orders of magnitude while the cost structure of being right has not moved at all. What is scarce is no longer compute, labor, or even time. It is taste, honesty about error, and the patience to audit. Institutions that adapt will, I expect, multiply their people rather than replace them; individuals who adapt get to run programs that used to be the size of a small lab. I am one of them, and the ledger above is the receipt.

## Appendix: data provenance and method

| Claim                                                         | Source                                                                                                                                                                        |
|---------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Token, request, and cost totals; monthly and model breakdowns | Platform usage exports (per-day cost CSV + per-token-type amount CSV, Nov 2025 – Aug 20, 2026), aggregated by a purpose-written script; November archive verified header-only |
| Model prices (v4-pro, v4-flash)                               | price column of the amount records                                                                                                                                            |
| 911 Zenodo records                                            | Zenodo API fielded search, creators.name:"Quni-Gudzin" (verified working query; 400-on-metadata-field variant documented)                                                     |
| 877 ORCID works                                               | ORCID public API works endpoint (num-found)                                                                                                                                   |
| 1,646 Paper nodes / 8,308 nodes                               | Internal knowledge-graph stats endpoint                                                                                                                                       |
| 114 GitHub repositories (109 in window)                       | GitHub org API, created_at dates                                                                                                                                              |
| 20 Cloudflare Workers                                         | Cloudflare account Workers API                                                                                                                                                |
| Live surfaces (papers.qnfo.org, qnfo.org, qwav.org)           | HTTP GET, 200                                                                                                                                                                 |
| NIH NRSA postdoctoral minimum \$50,004 (2019)                 | Wikipedia, “Postdoctoral researcher” (live-fetched August 2026)                                                                                                               |
| Load factor 1.8–2.2×                                          | Assumption (standard budgeting practice), stated as such                                                                                                                      |

| Claim                                                                                   | Source                                                                  |
|-----------------------------------------------------------------------------------------|-------------------------------------------------------------------------|
| Word conversions (0.75 words/token;<br>250 wpm reading; 40 wpm typing;<br>2,000 h/year) | Stated conventions; volume figures are<br>conversions, not value claims |