bagel

# Build vs buy: quality and scale at lower cost

You could build this. But then what?

## 1 What the raw feedback weighs

monthly, low-tier*

| Feedback source | Tokens per unit | Volume | Tokens / month |
|---|---|---|---|
| Call transcript (45 min) | 10,845 | 1,000 | **10,845,000** |
| Support ticket | 410 | 2,000 | **820,000** |
| Feature request (GTM / community) | 250 | 750 | **187,500** |
| CRM note (closed-lost, churn) | 250 | 500 | **125,000** |
| Slack channel (1-day threads) | 300 | 500 | **150,000** |
| Survey response | 200 | 750 | **150,000** |
| **Total** | | **5,500** | **~12.3M** |

## 2 The data-lake problem

A data lake keeps everything, so every query re-reads everything: **12.3M tokens a month, mostly noise**. An hour-long call can hold zero product gaps. Most tickets carry no product insight at all.

**21%**

### Access was never the bottleneck.

Claude's accuracy on Anthropic's own warehouse with no structured layer in front. Thousands of past queries added: accuracy moved **less than one point**. A governed layer took it above **95%**. Their conclusion: **the bottleneck is structure, not access.**

**Anthropic Research**
"How Anthropic enables self-service data analytics with Claude" · June 2026

# 3 What Bagel AI does instead

Extract and normalize **at ingestion, before anyone asks**. Only true product insights enter the queryable set, each compressed to a ~250-token unit linked to its source. A 10,845-token call becomes ~500 tokens - **22x smaller**.

| Feedback source | Volume | True insights (250 tokens) | Ratio | Tokens / month |
|---|---|---|---|---|
| Call transcripts | 1,000 | 2,000 | 1:2 | **500,000** |
| Support tickets | 2,000 | 240 | 8.3:1 | **60,000** |
| Feature requests | 750 | 750 | 1:1 | **187,500** |
| CRM notes | 500 | 100 | 5:1 | **25,000** |
| Slack channels | 500 | 150 | 3.3:1 | **37,500** |
| Survey responses | 750 | 375 | 2:1 | **93,750** |
| **Total** | **5,500** | **3,615** | | **~904K** |

**12X**

## Pay for signal, not noise.

**less tokens for the same feedback set.** 12.3M raw becomes ~904K validated insight tokens. Every query, from Bagel's models or from LLM clients such as Claude via MCP, runs against the small clean set, built by a product-optimized agentic architecture that extracts every insight from **millions of evidence records**. Optimized before you spend your first token.

**Trained on years of this data:** a renewal call reads differently than a sales or support call, and insights land in **your taxonomy** (payments, onboarding, your product tree).

**Bagel eliminates the quality vs. cost tradeoff.** Only validated signal in the set: higher accuracy and 12X lower token load, at once. And it **compounds through memory**: Bagel carries your taxonomy, product context, and every past insight forward, so each query starts informed instead of from scratch, while the full raw feedback stays retained and linked behind every insight. **Nothing is lost, and nothing starts cold.**

## 4 What “we can build it” commits you to

**Build ~10 integrations** for third-party and internal feedback sources, optimize each for product-insight extraction, and maintain them through every API change

**Build in security:** pass SOC 2, enforce role-based access, reduce PII (since your sensitive data will now run through Anthropic directly)

**Re-verify models** on every release (Bagel performs zero-risk upgrades)

**Support governance, enhancements, and customization** for 4-5 internal groups and hundreds of internal customers, each with their own taxonomy and needs

**Staff 1+ dedicated FTE at $180K/year** (conservative), or adoption fails and this is before the first query runs or the first token is spent

**Left alone, it rots in a month.**

Anthropic's own maintenance data: one unmaintained month and accuracy drifted **95% → 65%**. That was a dedicated data team, at Anthropic, on their own model.

**Anthropic Research**
“How Anthropic enables self-service data analytics with Claude” · June 2026

## 5 Let's sum it up

| | Build | Buy |
|---|---|---|
| **Engineering** | $180K+/year dedicated FTE | **Included** |
| **Tokens** | Unoptimized, vs. 12.3M/month lake | **12X-optimized, tier usage** |
| **Integrations, security, upgrades** | You build and maintain | **Included** |
| **Year cost** | **$180K+ plus tokens** | **From $24K plus tier usage** |

### Better, for less.

Your team could build it. You'd pay more to end up with less: fewer sources, noisier data, and a second product to maintain forever.

*Assumptions: mid-market scenario, ~5,500 items/month; token counts per unit as listed; extraction ratios and 250-token insight unit are Bagel production figures; 12,277,500 / 903,750 = 13.6x, stated as 12X; raw feedback retained and linked, 12X applies to the queryable set; $180K/year loaded FTE is conservative. Anthropic figures (21%, 95%+, <1-point gain from raw access, 95→65% drift) from “How Anthropic enables self-service data analytics with Claude,” June 2026.*