What LLM Document Extraction Actually Costs in 2026

A reproducible benchmark on public SEC filings: nine models, four document types, 2,280 extractions

2026 snapshot | July 16, 2026

Headline

Extracting structured data from an SEC filing with a large language model costs pennies. At published list prices, the flagship model in this benchmark (Claude Sonnet 4.6) extracts a routed 10-K for $0.20, a proxy statement for $0.25, and an 8-K for $0.02; a fast-tier model runs a third of that, and capable open-weight models run one to three cents per document. The scarce thing is not the extraction, it is knowing whether the extraction is any good: on the hardest document type in this corpus, 199 of 540 outputs across nine models failed schema validation, and the cheapest large model returned well-formed, empty JSON on every long document it saw. Cost is a solved, nearly deterministic quantity. Measurement is the part you still have to build.

What we measured

Scope & approach

A frozen, seeded corpus of 120 public SEC documents (30 each of four types: 8-K current reports, 10-K Item 1A and Item 7, DEF 14A executive compensation, and earnings-release exhibits to Item 2.02 8-Ks) was run through nine models hosted on AWS Bedrock through one uniform API path: Claude Sonnet 4.6 and Haiku 4.5, Qwen3-235B and Qwen3-32B, GLM 4.7, Llama 4 Maverick, Mistral Large 3, DeepSeek V3.2, and Gemma 3 27B. Every call used the same written extraction specification and JSON Schema per document type; every output was validated against that schema; every call's token usage, list-price cost, latency, and stop reason was recorded in a per-call ledger. The run totals 2,280 extractions and $166.71 at list price. Cost figures are token usage times each model's published hosted list price (Anthropic API price for Claude, Bedrock on-demand US East for open-weight models), so they are independent of any credits or discounts we actually paid. Everything is re-derived from public documents; no client material of any kind was used.

The question this benchmark answers is narrow and practical: if you point a current model at a real filing with a real field specification, what does one document cost, and what do you get back? The four document types were chosen to span the difficulty spectrum: short event-driven filings where one document may carry six unrelated events (8-K), long sectioned narrative (10-K), dense recurring tables (DEF 14A compensation), and semi-structured financial highlights (earnings releases).

The cost ladder

Figure 1 is the central result. Each bar is the mean cost per document at published list price for one model on one document type, over the 30 corpus documents of that type (rep 1, baseline configuration).

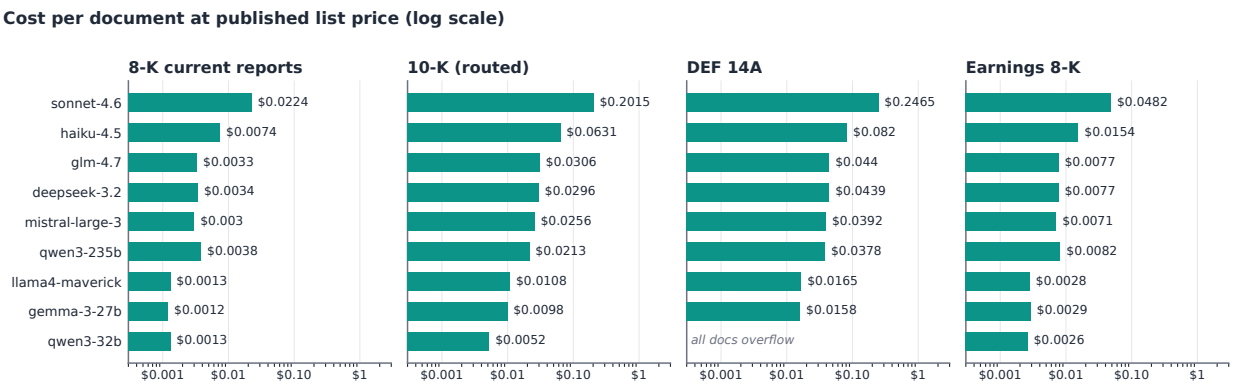

Figure 1: Cost per document by model and type, published list prices, log scale. The 10-K panel is the routed configuration (Item 1A + Item 7 only). Qwen3-32B has no DEF 14A bar because all 30 documents exceeded its 32K-token context window.

Three readings of the ladder:

- **The ceiling is twenty-five cents.** The most expensive cell in the benchmark, the flagship on a proxy statement, is $0.2465 per document. A full pass over all 120 corpus documents on Sonnet 4.6 costs about $15.56; the same pass on Haiku 4.5 costs $5.04; on Llama 4 Maverick it costs $0.94.
- **The floor is a tenth of a cent.** Gemma 3 27B extracts an 8-K for $0.0012 and Llama 4 Maverick for $0.0013. Whether those outputs are usable is a separate question, and Section 3 is about exactly that.
- **Batch halves everything.** Batch-mode pricing is 50% of list on the same tokens. Applied to this ladder it puts the flagship 10-K at $0.10 and the Haiku 10-K at $0.03 per document.

How to read the batch column

The batch figures are a computed projection: 50% of list price applied to the measured token counts. Batch mode was not run as a separate measured configuration in this benchmark.

Cost is nearly deterministic

Across five identical repetitions of the full baseline on the Anthropic pair, the standard deviation of mean per-document cost never exceeds $0.0006 (Sonnet on the 10-K). Cost estimates from a single run are reliable; error bars on the ladder are smaller than the printed digits. The same is not true of content, which is the subject of the noise floor in Section 4.

Measurement is the scarce part

Every output in this benchmark was validated against a strict JSON Schema derived from the written field specification. Validity and context overflow are reported as first-class results for every model and type (Figure 2); no failing row was dropped from the analysis.

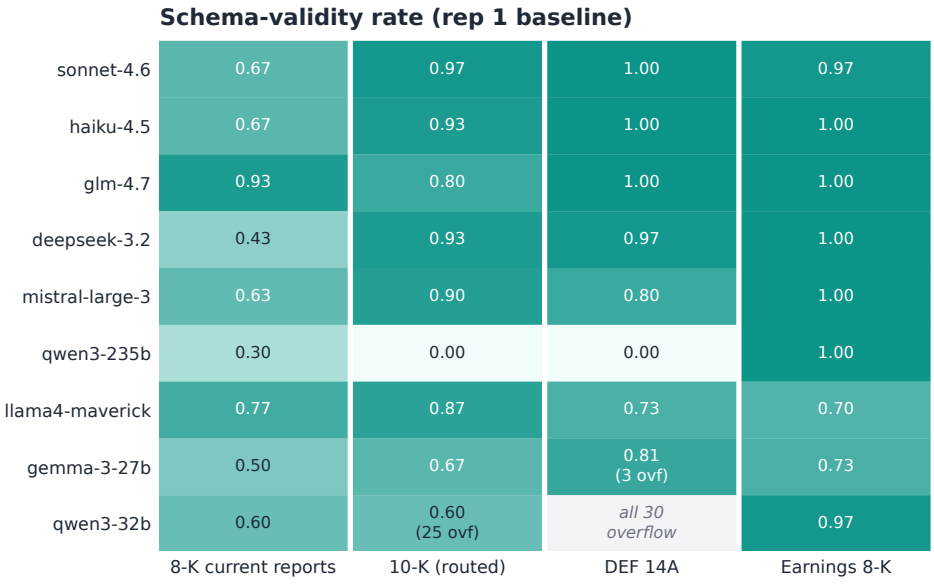

Figure 2: Schema-validity rate per model and document type, computed over the documents that fit the model's context window; (n ovf) annotations count the documents of the 30 that did not fit and are reported alongside, never dropped. Qwen3-32B's DEF 14A cell has no rate because nothing fit.

Four findings anchor the claim that measuring extraction quality, not running extraction, is where the difficulty lives.

The multi-event 8-K is the hardest type in the corpus

Across nine models, 341 of 540 8-K extractions validated (63%), the worst of the four types, and even the frontier pair sits at 0.67. The 8-K schema requires one typed record per Item heading, and a single filing can carry one event or six: the failure mode is variable cardinality, models emitting the wrong number or shape of event records. By contrast the earnings release, a semi-structured but single-object type, validates at 516 of 540 (96%). Difficulty tracks output structure, not document length.

You can pay for empty output, and only validation catches it

Qwen3-235B is among the cheaper large models on the ladder ($0.021 per routed 10-K) and returned syntactically clean, schema-shaped JSON on every long document, with nothing in it: `risk_factors: []`, `risk_factor_count: 0`. It validated 0 of 30 routed 10-Ks and 0 of 30 proxy statements, while validating 30 of 30 short earnings releases. A cost benchmark without schema validation would have scored those runs as successes at an attractive price.

Small-context models drown in SEC filings

Qwen3-32B (32K context) could not fit 85 of its 150 assigned documents: all 30 DEF 14As, all 30 whole-document 10-Ks, and 25 of 30 routed 10-Ks. An overflow is not an error to retry; the model cannot do the job in one shot at any price, and the benchmark records it as its own outcome.

Some fields are not well-posed enough to measure

The weakest field group in the benchmark is the 10-K MD&A driver list: cross-model agreement 0.19, and same-model repeat agreement only 0.71 for Sonnet and 0.36 for Haiku. When a model cannot agree with itself across identical runs, the field specification, not the model, is underdetermined. Extraction specs need the same engineering discipline as schemas: fields whose answers are not unique should be redesigned or dropped.

Consensus and the noise floor

With no gold labels, quality is measured as agreement: cross-model consensus per field group (do independent models extract the same values?) scored against a noise floor of same-model repeat agreement (does one model extract the same values twice?). Figure 3 shows both.

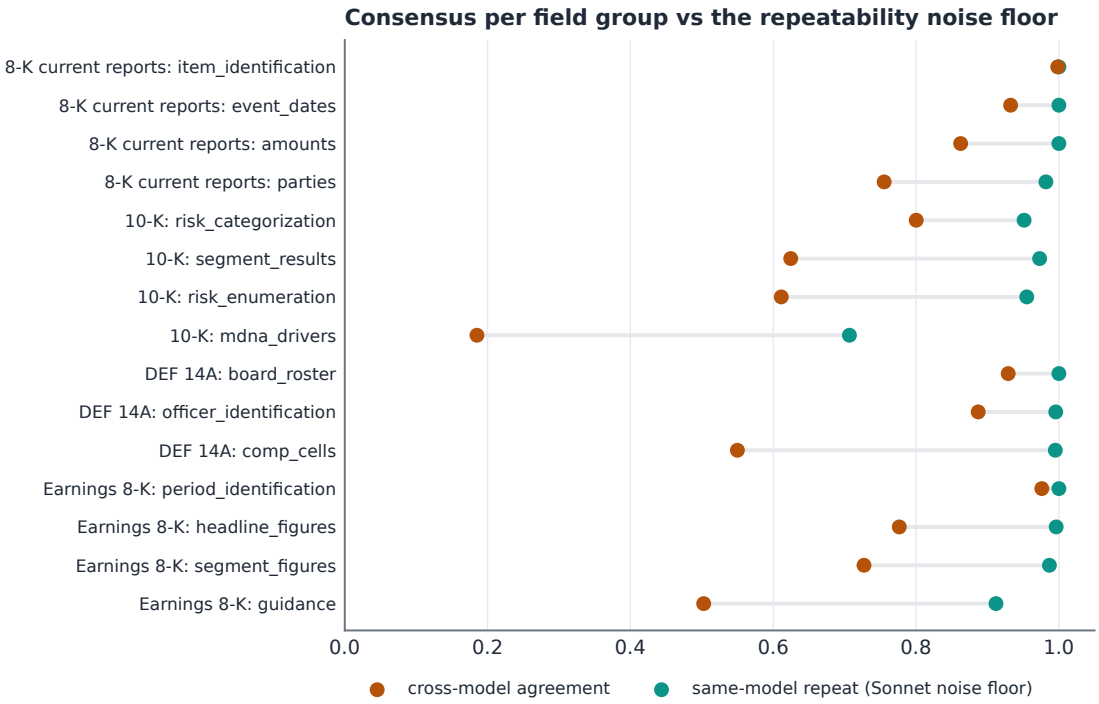

Figure 3: Amber: mean pairwise cross-model agreement (valid outputs, rep 1 baseline). Teal: same-model agreement across five identical Sonnet 4.6 repetitions. The gap between the dots is real model disagreement; the distance of the teal dot from 1.0 is inherent run-to-run noise.

Structured, closed-form field groups reach near-perfect consensus: which items an 8-K contains (1.00), the reporting period of an earnings release (0.98), the board roster (0.93). Numeric-cell and list-valued groups sit in the middle: compensation cells agree at 0.55 across models against a 0.995 same-model floor, meaning models are highly self-consistent and genuinely disagree with each other, mostly on which cells they cover. Narrative groups are weakest, and the MD&A driver group fails the well-posedness test above.

What consensus does and does not claim

Agreement is computable only where both models produced schema-valid output, so consensus on the 8-K is measured among the surviving 63%; the validity figure carries the failure mass. Consensus also measures agreement, not correctness: nine models can agree on a wrong value. The claim sold here is cost and relative quality signal, not accuracy against gold labels.

A related audit result: evidence quotes, which the 8-K and earnings specs require verbatim, ground in the source document at rates between 0.56 (Qwen3-32B) and 0.94 (Sonnet 4.6 on earnings releases). Quotes are mostly real and mostly close, and not reliably verbatim for any model; a pipeline that presents them as clickable audit pointers needs its own verification pass.

The levers

Section routing halves 10-K cost at unchanged validity

Sending the whole 10-K instead of the two target sections costs 2.0x on both Anthropic models (Sonnet $0.20 routed vs $0.40 whole; Haiku $0.06 vs $0.13) and 2.6x on Qwen3-235B, with validity unchanged (Sonnet 0.97 routed vs 0.93 whole, Haiku 0.93 vs 0.97). Routing is also what makes the 10-K possible at all for a 32K-context model: 5 of 30 routed documents fit where 0 of 30 whole documents did. Never pay to send sections you do not want extracted.

Thinking buys validity on the hardest type, at about 3x

With adaptive thinking enabled, Sonnet 4.6's 8-K validity rises from 0.67 to 0.87 while the other three types stay flat, and cost rises 1.5x (DEF 14A) to 3.4x (earnings), driven by five to six times the output tokens. Thinking is a targeted spend: worth it where output structure is the failure mode, wasted where validity is already at ceiling.

The thinking lever is measured on one model

Bedrock rejects adaptive thinking on Haiku 4.5, so the thinking comparison exists for Sonnet 4.6 only. No thinking claim is made for any other model. One thinking-arm run (a single 10-K) is excluded as a documented runaway-generation case that saturated a 65,536-token output budget.

What this means for buyers

At these prices, per-document extraction cost is no longer a budgeting question: a thousand routed 10-Ks on the flagship is about $200 at list, $100 in batch, $63 on the fast tier, and $11 on Llama 4 Maverick. The budgeting question is the one the validity and consensus tables answer: which model, on which document type, produces output you can trust, and how you will know. The benchmark's own machinery, schema validation, per-field consensus, a repeat-run noise floor, and grounded evidence quotes, is a working outline of that measurement layer.

Reproducibility

The corpus is a frozen, seeded document-id list (seed 20260714) committed as manifests with SHA-256 hashes of every raw and cleaned document; prices are committed with source URLs and retrieval dates; every extraction call is a ledger row with token counts, cost, stop reason, and validity; specs, schemas, harness, and analysis code are in the repository (`github.com/joshuacook/llm-extraction-benchmark`). The companion methodology page states the selection, comparison, and exclusion rules.

Boundaries of the claim

Costs are list prices as of 2026-07-15 and will drift; the Llama 4 Maverick Bedrock price was corroborated through a third-party tracker and should be re-verified before quoting. Inputs are cleaned EDGAR text, so PDF parsing and OCR costs are out of scope. The corpus is large-cap-weighted (top 1,500 registrants). Validity is schema conformance, not factual accuracy; consensus is agreement, not ground truth. All extraction ran through the Bedrock Converse API; provider-native APIs may price or behave differently.