What LLM Document Extraction Actually Costs in 2026 — Methodology

2026 snapshot | July 16, 2026

Objective

Measure what single-document LLM extraction costs at published list prices on real SEC filings, and how much quality signal is available without gold labels, as a reproducible 2026 snapshot.

Corpus

120 public documents from SEC EDGAR, 30 per type: 8-K current reports (excluding Item 2.02-only filings), 10-K annual reports with Item 1A and Item 7 isolated by header parsing, DEF 14A proxy statements validated to contain a Summary Compensation Table, and EX-99 earnings press releases from Item 2.02 8-Ks. Selection is deterministic: a seeded shuffle (seed 20260714) of the top 1,500 registrants in `company_tickers.json`, filled via the SEC submissions API, at most one document per issuer (120 distinct issuers), filing window 2025-07-01 to 2026-06-30. Manifests with source URLs and SHA-256 hashes of raw and cleaned text are committed in the repository. Zero client material; every input is a public document.

Models and harness

Nine models on AWS Bedrock, all called through the same Converse API path with the same per-type system prompt (the written field specification plus its JSON Schema): Claude Sonnet 4.6, Claude Haiku 4.5, Qwen3-235B (VL), Qwen3-32B, GLM 4.7, Llama 4 Maverick, Mistral Large 3, DeepSeek V3.2, and Gemma 3 27B. Configurations: baseline (10-K input routed to Items 1A and 7), wholedoc (full 10-K, routing-lever control, 10-K only), and adaptive thinking (Sonnet 4.6 only; Bedrock rejects it on Haiku 4.5). Five identical repetitions of the baseline ran on the two Anthropic models for the noise floor. Total: 2,280 extraction calls.

Cost basis

Cost equals token usage reported by Bedrock times the model's published hosted list price: the Anthropic API price for Claude models, the Bedrock on-demand US East price for open-weight models (retrieved 2026-07-15, sources committed alongside the figures). This is independent of the credits under which the run was actually billed. Batch figures are a computed 50%-of-list projection on the same token counts, not a measured run.

Validation and outcome accounting

Every output is validated against a strict JSON Schema (`additionalProperties: false`, required fields, non-empty constraints). Validity, parse mode, stop reason, and context overflow are recorded per call and reported per model and type; no failing row is filtered out of the results. Documents exceeding a model's context window are recorded as overflow outcomes without a model call. Truncated generations were re-run at doubled output budgets until none remained except one documented runaway case (a single Sonnet thinking-arm 10-K saturating 65,536 output tokens), which is excluded and disclosed.

Consensus and noise floor

Quality without gold labels is measured as agreement per field group, using rep-1 baseline outputs that validated. Rules: Jaccard for sets (two empty sets agree), relative tolerance 10^{-6} for numerics, histogram overlap for categorical distributions, exact match for scalars; evidence quotes are scored as verbatim-substring grounding against the cleaned source. Cross-model consensus is the mean pairwise score over all valid model pairs per document. The noise floor is the same metric between the five identical repetitions of one model. Adaptations for cross-model alignment (pooled party sets, name-keyed officers, case-folded heading sets) are stated in the analysis document.

Limitations

No accuracy claim against ground truth: consensus measures agreement, and agreed-upon values can be wrong. Validity is schema conformance, not factual correctness. Inputs are cleaned EDGAR text; PDF parsing and OCR are out of scope. The corpus is large-cap-weighted. Prices drift; one price (Llama 4 Maverick) was corroborated via a third-party tracker because the AWS pricing page is script-rendered. The thinking lever exists for one model only. Consensus on low-validity types is measured among surviving outputs.

Infrastructure and reproduction

Corpus assembly, extraction, and rendering all ran as Cloud Build or Cloud Run jobs in an isolated GCP project; AWS credentials were mounted from Secret Manager into jobs only. The full ledger (2,280 rows), raw and cleaned corpus, specs, schemas, harness, worklists, and analysis code are in `github.com/joshuacook/llm-extraction-benchmark` and the project bucket. A third party can re-run the pipeline from the frozen manifests.