Ahead-of-Time Comprehension: Does a Prepared Signal Layer Beat Retrieval and Brute Force on Corpus-Wide Questions?

BuildBetter research · 2026-08-15 · DRAFT — pending blind human spot-check of gold

Abstract. Product teams collect more customer conversations than anyone can read. The workspace in this study holds 6,018 call recordings and 8,533 support conversations. About 101 million tokens. Three strategies compete to answer questions over a corpus that size: extract structured signals once at ingestion (BuildBetter), search the raw text at query time (BM25, embeddings, hybrid), or pay a model to read everything. We measure all three at the data layer. The test is blunt: for each verified piece of customer evidence, does the strategy have it? On calls the current pipeline ingested, the signal layer holds **99.0%** of the evidence at source level (95% CI 98.3–99.7). The best retrieval setup finds **27.9%** — and that is with an oversized budget of 400 chunks per question. Typical budgets find 9–11%. Full context does not fit. A live scan of the whole corpus reached near-equal answer quality at 370–1,100× the cost and hours of latency. The measurement also found two dated pipeline problems: extraction never ran on 2021–2023 calls, and chunk-exact extraction dipped for calls from March–June 2026. All aggregates, hashes, and spend are itemized. Raw text stays in the private run directory.

1. Introduction

"What are the top customer themes about pricing, and how many customers raise each?" Easy question. The answer sits in thousands of calls nobody will read. Every week adds more.

The industry has three answers. **Ahead-of-time comprehension** reads every conversation once, at ingestion, and stores typed extractions — customer problems, requests, questions, praise, each tied to its source. **Query-time retrieval** keeps the corpus raw and searches per question, by keyword match (BM25), embedding similarity, or both fused. **Brute force** reads everything, per question, and pays for it.

Public benchmarks test retrievers and long-context models. None of them answer the operator's question. The operator asks: if the evidence exists in my corpus, which system has it when I ask? End-to-end answer scores cannot settle this. They mix the data question with agent behavior — prompt wording, tool policy, abstention thresholds. In our runs a one-sentence prompt change moved end-to-end scores while the data underneath never moved. So this paper measures the data layer and reports end-to-end results as a labeled appendix.

2. Related work

LIMIT tests compositional retrieval constraints. BrowseComp-Plus tests evidence retrieval over a fixed corpus. Both score retrievers against known evidence, and our coverage framing borrows from that design. Neither includes a build-time extraction layer as a system under test. LongMemEval covers temporal, multi-session, and abstention tasks over conversation history; our negative tasks follow its lead, and our withdrawn trend family failed for reasons it predicts. Mixedbread's Wholembed v3 and Toast 1 report stronger retrieval baselines than the dense and hybrid systems here. Read our retrieval numbers as common production configurations, not the retrieval frontier. Raindrop's Signals 2 is the closest industrial relative — classification at build time — but it is not an open-ended retrieval baseline. No public benchmark we know of measures prepared-extraction coverage against retrieval and full-scan alternatives on the same verified evidence.

3. Corpus and freeze

One full production-shaped workspace: every completed call (6,018) and support conversation (8,533) as of 2026-08-13. Exported read-only. Pseudonymized with salted identifiers. Chunked into 220,698 excerpts of about 2,400 characters — roughly 101M tokens. The workspace's 120,897 signals came from the live extraction pipeline during normal operation, not for this study, and were frozen in the same export with their evidence alignments. Content hashes for objects, chunks, and signals sit in a private manifest. Tracked files carry aggregates only.

4. Gold truth

Gold has to be independent of every system it judges. It comes from raw text only, in four stages. Each stage has a gate.

4.1 Read everything. A batch pass labeled every chunk against a fixed 14-domain taxonomy written before the sweep ran. 220,697 of 220,698 chunks succeeded. No sampling. A prevalence claim needs exhaustiveness, so we paid for exhaustiveness.

4.2 Cluster and verify. Mentions were embedded and clustered into themes. Then a second model family, with a different prompt, re-judged every support pair behind a candidate gold answer: 3,766 of 3,767 pairs, 0.894 mean support rate. Cross-family agreement on 1,000 double-labeled chunks: mean κ=0.538, per-domain κ from 0.12 to 0.74. The weak domains got dropped or repaired downstream.

4.3 Remove the vendor's own voice. We ran comparative rounds and audited every row that made no sense. The audits found three ways vendor voice sneaks into gold when you trust frequency. First, whole themes that are the vendor's onboarding chatter — 23 of 49 raw themes, including a 180-source "theme" that is the support team inviting people to Slack. Second, scripted pitch lines repeated near-verbatim across sources; a quote-uniqueness statistic now flags them mechanically. Third, vendor speech inflating real themes pair by pair. We classified all 1,349 support quotes behind task-feeding themes. Only **29.7% carried customer voice.** Prevalence now counts customer pairs only. Each fix is a detector in the pipeline, not a hand edit. Final gold: **24 customer themes, 836 verified customer evidence chunks** (SHA-256 `38fd0c64…`).

4.4 Calibrate the instrument. An oracle — the reader handed gold evidence directly — must score at least 95% on a development split before any comparative number counts. It took nine calibration rounds to get there. The rounds fixed judge-precedence bugs, reader ranking drift, a trap negative question, and a comparison position bias that was handing free wins to position-biased systems. The oracle now scores 100% on every graded metric. One gate remains: a blind 24-item human spot-check of gold (16 verified pairs, 8 refuted, shuffled, sealed key). It has not run yet. This document stays a draft until it does.

5. Measurement

For every verified evidence chunk, one question per strategy: do you have it? Signals are measured standing, two ways. Chunk-exact means a signal aligns to the exact evidence chunk — the lower bound. Source-level means the evidence source produced at least one signal — the upper bound. Retrieval gets the same query per theme for every system and is scored on whether the evidence chunk lands in the top K retrieved, for K of 10, 40, 100, and 400. Reading 400 chunks means ~240K tokens of context per question. Nobody runs that in production. We include it so retrieval gets shown at its best. Full context is scored on feasibility and scan cost. Confidence intervals are 95% cluster bootstraps resampling themes, 10,000 draws — the conservative choice, since evidence within a theme is correlated.

Two frames. The **modern frame** is calls ingested 2025 onward (n=408). It is a time cut, stated openly, and §7 shows why it is the right cut: extraction never ran on the early corpus. The **blended frame** is all eras (n=836). It includes that history and describes the workspace as it stands today.

6. Results: the data layer

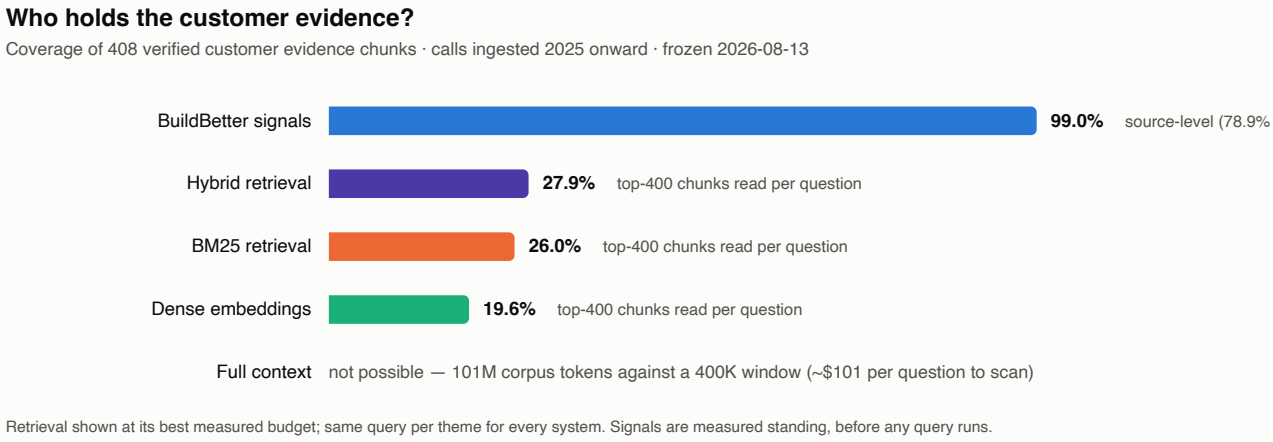

Figure 1. Coverage of verified customer evidence, modern frame (n=408). Signals hold 99.0% at source level before any query runs. The best retrieval setup finds 27.9% after reading 400 chunks. Full context cannot participate.

Strategy	Modern frame (2025+, n=408)	Blended (all eras, n=836)
Signals, source-level	**99.0%** [98.3, 99.7]	72.2% [55.9, 85.8]
Signals, chunk-exact	78.9% [74.5, 83.5]	59.3% [45.8, 70.0]
Hybrid retrieval, top-400	27.9%	26.8%
BM25, top-400	26.0%	24.4%
Dense embeddings, top-400	19.6%	20.0%
Best retrieval, top-100	11.3%	11.1%
Full context	does not fit: 101M corpus tokens vs a 400K window; ~$101 per question to scan uncached	

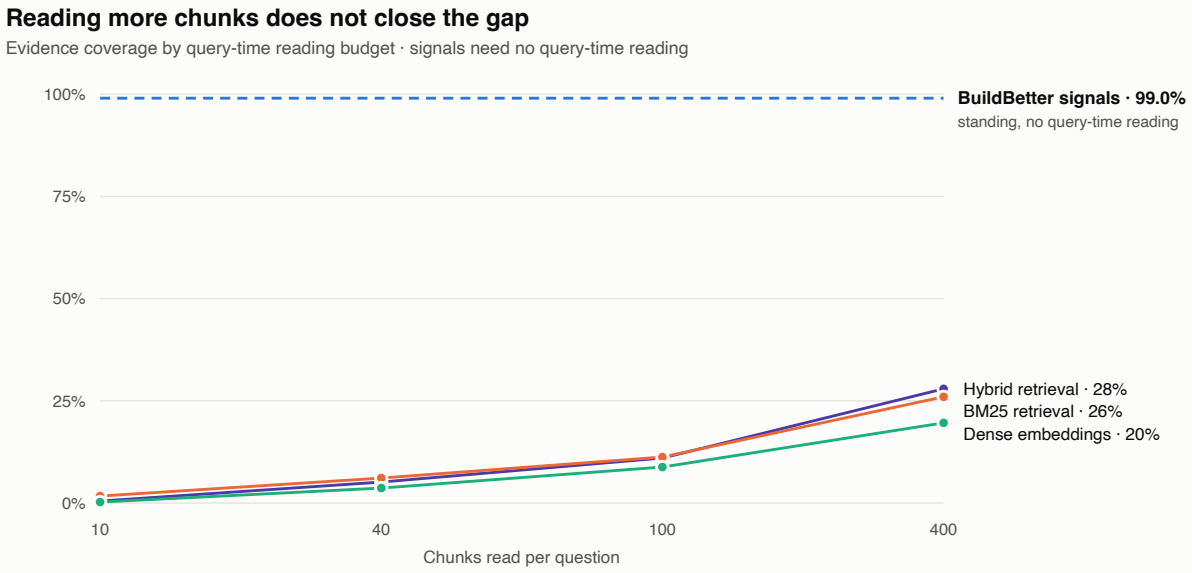

Figure 2. Coverage by reading budget, modern frame. Retrieval climbs with budget and stays a factor of ~3.5 below the signal layer at the largest budget measured. The signal layer reads nothing at query time. Extraction already ran.

No interval overlaps, in either frame. Support conversations hit 100% source-level coverage. The blended shortfall lives almost entirely in recordings — 332 of 340 uncovered chunks. Section 7 explains why.

7. Why signals are not 100%

The theory says every conversation passes through extraction, so coverage should approach 100%. In the modern frame it does. The gap is history.

The backlog. Of the 340 uncovered blended-frame chunks, 68% come from sources with zero signals. Not weak extraction — no extraction. Corpus-wide, 100% of 2021–2023 recordings carry no signals. 2024: 41%. 2025–2026: 7–11%. Extraction arrived during 2024 and nobody backfilled the past. The uncovered calls are ordinary full-length calls, median ~23 chunks. The fix is a backfill over roughly 1,000 early recordings. After the backfill, this measurement re-runs for free and the blended number converges toward the modern one.

The spring dip. Inside 2026, chunk-exact coverage ran 100% in January and 79% in February. Then 36–54% for March through June. Then 86–100% in July and August. Source-level stayed near 100% the whole time. Monthly samples are small — 5 to 14 chunks — but four consecutive depressed months with a clean recovery brackets a window where extraction got coarser and then got fixed. We are handing that window to the pipeline team to match against deploys. The benchmark was not built to find regressions. It found one anyway, with dates.

8. Money no object: a live full-corpus scan

The strongest competitor on paper reads everything. We ran it, live. One batched map over all 220,698 chunks with a question-conditioned prompt and strict schema — 8.5% of chunks failed to parse, counted against it. One scan served three questions. A reduce step grouped 18K–52K evidence notes per question into answers. The same blind judge scored them against the same gold.

Question	Full scan	Signal-backed agent	Oracle ceiling
Top 5 themes, whole workspace	0.20	0.00	1.00
Top 3 integration themes	0.33	0.33	1.00
Prevalence comparison of two topics	1.00	1.00	1.00
Measured cost per question	$33.55 cold; $11.21 amortized over three	~$0.03	—
Latency	~14 hours (batch)	~1 minute	—

Total evidence access bought one extra gold theme, on one question, for 370–1,100× the cost. The scan's enumerations also grouped evidence into broad umbrella buckets instead of gold's specific themes. Every non-oracle system fails the same way. Above the data layer, the binding constraint is aggregation quality, not evidence access.

9. Appendix result: end-to-end answers measure the agent, not the data

Six systems answered 16 corpus-wide tasks under a shared contract with blind cross-family judging. The oracle scores 1.00 on every graded metric. Every real system scores 0.05–0.29 on enumeration completeness. A one-sentence contract change — strict abstention versus best-estimate — moved several systems' scores by whole tiers. Two findings held steady across every round. Single-shot retrieval abstains on every corpus-wide enumeration; forty chunks cannot support top-N-with-counts claims. And agentic raw retrieval fabricates themes for verified-absent topics in every round, while the signal-backed agent under the strict contract declines correctly. The numbers live in the tracked aggregates. They stay out of the headline because they measure agent design, which is revisable, not data, which is not.

10. Threats to validity

- **Gold shares a lens with retrieval.** Gold comes from raw text. Retrieval searches raw text. The signal layer is measured against both. The verification, agreement, and voice gates bound this asymmetry. They do not remove it. The bias also runs the other way — gold's granularity was audited partly through disagreements the measured systems surfaced.
- **Coverage is a bracket.** Chunk-exact (79%) and source-level (99%) bound true topical coverage in the modern frame. We report both and claim only what each bound supports.
- **Our retrieval is not the frontier.** Queries are theme labels. No query rewriting, no reranking, no late-interaction models. Stronger stacks would score higher and should be added before any external comparison claim.
- **The frame is a choice.** The modern frame excludes un-backfilled history. Both frames appear with intervals. The blended frame is the honest description of the workspace today.
- **The inference is scoped.** One organization. One freeze. 24 themes, 836 evidence chunks, three scan questions. The task set sits below its 30-task target. The regression window rests on 5–14 chunks per month. The findings are strong for this workspace and directional beyond it.
- **The human gate is pending.** The blind spot-check of gold has not run. Draft until it does.
- **The instrument missed twice, on record.** Batch cost projections under-counted reasoning tokens twice: the sweep came in at $34.17 against $26.11 projected, the scan at $33.51 against $20.73. Hard budget gates contained both. Total external spend: $86.74 under a $100 cap, itemized.

11. Reproducibility

Every number traces to a tracked file in this directory. `data-coverage.json` holds coverage, budgets, splits, intervals, and scan results. `gold-manifest.json` holds gold statistics, agreement, and hashes. `complete-v1/v2/v3-summary.json` hold the end-to-end runs — v1 kept with its position bias as a cautionary record. Charts regenerate from `corpus_v2_study.py charts`. The private run directory holds the frozen corpus, gold with quotes, task wording, and per-row records under content hashes. Authorized reviewers can reproduce the full run from it. Independent external reproduction needs a public corpus and independent human labels.

12. Conclusion

On a 101-million-token customer corpus, prepared extraction is the only strategy that is near-complete, fast, and cheap at once. 99.0% source-level coverage on processed calls, about a minute per question, about $0.03. Retrieval at production budgets finds a ninth of the evidence. Retrieval at an extreme budget finds a quarter. Reading the whole corpus per question costs three orders of magnitude more and answered no better. The scoped claim: if the pipeline processed the conversation, the evidence is in the signal layer — and nothing near its cost comes close.

Next: the blind spot-check that lifts the draft status, the pre-2024 backfill that closes the blended gap, and the deploy-log check on the spring-2026 window. Each one re-measures against this same harness for free.

References

- Google DeepMind. LIMIT: benchmarking compositional retrieval constraints.
- Texttron. BrowseComp-Plus: fixed-corpus evidence retrieval.
- Wu et al. LongMemEval: temporal, multi-session, and abstention evaluation over conversational memory.
- Mixedbread. Wholembed v3 and Toast 1: late-interaction and agentic-search baselines.
- Raindrop. Signals 2: frontier classification at build time.

Companion artifacts: the executive one-pager (`2026-08-14-report.html`), the markdown paper, and PR #5691 with the full decision and spend log. Models: extraction sweep and scan on `gpt-5.4-nano` ; answers and reduction on `gpt-5.6-luna` ; verification, voice classification, and judging on `claude-sonnet-5` . External spend: $86.74 under a $100 hard cap.