The numbers behind "never confidently wrong."

qbrin is a trust layer over your company's tools. This report shows every benchmark we run — **including where we trail** — because a trust product that hides its weak spots isn't one. Headline: on the axis that decides whether you can ship AI to a company (not being confidently wrong), **qbrin beats every system we've benchmarked.** On raw answer-accuracy over public Q&A, it trades coverage for calibration and is competitive, not dominant. Every figure below is measured — never modeled unless labelled.

Scorecard

ALL SUITES · MEASURED

BENCHMARK	WHAT IT MEASURES	QBRIN	BEST ALTERNATIVE	
Enterprise RAG Benchmark	500 real company Q&A — decision accuracy	**88.4%**	—	lead
Trap questions (ERB)	Confidently-wrong answers on 281 traps	**0 / 281**	—	lead
Unanswerable safety	False answers when there is no answer (n=301)	**5.3%**	trad RAG 20.3%	~4× safer
Citation trust (ERB)	Answers carrying a verifiable citation	**86%**	raw RAG 6%	lead
Temporal / current-fact	Returns the new fact, not the stale one	**100%**	mem0 25% · Graphiti ~17%	lead
Memory precision	PrecisionMemBench (gbrain's own scorer)	**0.78**	gbrain 0.08	~10× lead
Multi-hop retrieval	vs HippoRAG (published SOTA)	**97**	HippoRAG 96	edge
Compression @8192	vs RAPTOR / GraphRAG / vanilla / community	**33.8%**	community 27.5%	lead at scale
Public Q&A accuracy	FinanceBench / multi-hop raw accuracy	**20–72%**	trad/hybrid higher	trades for safety
Relational-graph retrieval	gbrain's home corpus (P@5)	**21.4%**	gbrain 49.1%	trails
Token economics	Tokens read per answer	**up to 20×**	heavy RAG pipeline	cheaper

How to read this

METHODOLOGY PRINCIPLES

Every figure is from a real benchmark run on qbrin's pipeline; competitor figures are their own published numbers or faithful re-runs on identical data/judge. We label **MEASURED** vs **MODELED**, disclose sample sizes and baselines, and show confidence intervals where the sample is small. "Beats every system we've benchmarked" is scoped to the **trust axis** (lowest false-accept rate + highest citation trust). We do not claim to top every raw-accuracy leaderboard — and the next pages show exactly where we don't.

qbrin Benchmark Report. Suites: Enterprise RAG Benchmark (ERB), multi-hop, RAGBench, FinanceBench, temporal/CTP, PrecisionMemBench, compression SOTA, multilingual. Measured with independent LLM judges; competitor results are published numbers or like-for-like re-runs.

The moat — never confidently wrong MEASURED

TEST	SETUP	QBRIN	BASELINE
False answers on unanswerable questions	301 multi-hop questions with no valid answer; substantive (non-abstain) reply = a fabrication	**5.3%**	trad 20.3% · hybrid 17.6% · rerank 18.9%
Trap questions	281 ERB traps built to bait a wrong answer (131 near-miss + 150 absent-value)	**0 / 281**	—
Citation trust	Share of answers carrying a citation that actually supports the claim (high-stakes mode)	**86%**	raw RAG 6%
ERB decision accuracy	500 real company questions (470 answerable + 30 unanswerable)	**88.4%**	coverage 87.0% · risk 5.3%

Method: ERB is qbrin's own enterprise benchmark on real company Q&A — the most product-representative test we run. 88.4% is in-sample; cross-validated it holds at ≈87.2% (we report both). The safety result is the load-bearing one: when the answer isn't in the sources, qbrin abstains instead of inventing — ~4× less often wrong than traditional RAG.

Head-to-head vs named systems THEIR DATA · THEIR SCORER WHERE APPLICABLE

COMPETITOR	AXIS	QBRIN	COMPETITOR
mem0 / Graphiti	Current-fact accuracy (supersession) on temporal benchmark	**100%**	mem0 25% · Graphiti ~17%
gbrain	PrecisionMemBench — memory precision (gbrain's own scorer, n=77)	**0.78**	0.08
HippoRAG	Multi-hop retrieval recall (matched setup)	**97.0**	96.0
RAPTOR / GraphRAG / vanilla / community	Knowledge-map compression @ 8192-token budget (n=80, bootstrap CI)	**33.8%**	community 27.5 · RAPTOR 18.8 · vanilla 12.5 · GraphRAG 5.0

Honest caveats: the HippoRAG MuSiQue number (qbrin 64.4 vs published 51.9) is **not** a matched run and we don't headline it. On compression there is **no clean sweep** — qbrin leads at the largest budget and on decision questions and clearly beats GraphRAG, but community-summary wins the mid budgets and RAPTOR is competitive throughout; at n=80 the top three overlap in CIs.

Where qbrin does not lead

WE PUBLISH THIS ON PURPOSE

BENCHMARK	RESULT	WHY
FinanceBench raw accuracy	~20% vs trad RAG 31%	It abstains rather than guess on financial tables; answered-accuracy is actually +9.9pp over trad, but raw accuracy trails.
Multi-hop raw accuracy	72.5% vs hybrid 80%	Same trade — coverage given up for calibration. Strong relative-recovery, lower headline accuracy.
Relational-graph retrieval	21.4% vs gbrain 49.1% (P@5)	Honest loss on gbrain's home corpus — qbrin doesn't traverse a relational graph at retrieval time.
Cross-domain transfer	enterprise-shaped	Verification rules tuned on enterprise text over-abstain on financial-table corpora.

The point: qbrin is built to be **trustworthy and cheap**, not to top every accuracy leaderboard. If your priority is "never confidently wrong + low cost," it wins. If it's raw accuracy on answerable public Q&A, a tuned hybrid RAG is competitive. We'd rather you know that before the pilot than after.

Token economics

MEASURED + MODELED, LABELLED

PATH	TOKENS / ANSWER	VS BASELINE
qbrin — lean / decision path MEASURED	**~687**	**up to 20× cheaper than a heavy multi-query / reranking RAG pipeline**
qbrin — full stack (heaviest) MEASURED	**~3,081**	for the hardest questions
Live routing on a real corpus MEASURED	**~2,876**	~97% of questions routed to the cheap path · ~3× cheaper than the full stack
Traditional RAG (10–12 raw chunks)	~6,000–7,200	modeled from chunk size

Method: "up to 20×" is qbrin's lean path vs a heavy production RAG pipeline (multi-query ×3 / 20-chunk + reranking, modeled ~18,500 tokens); savings are workload-dependent — on a longer-chunk corpus the live router measured ~3× vs the full stack. Cost scales linearly with tokens read, so fewer tokens = lower cost and faster answers.

Retrieval quality OUT-OF-SAMPLE VALIDATED

CORPUS	METRIC	LIFT
Multi-hop	Recall@5 vs the prior default	**+13.75pp**
RAGBench	Recall@5 vs the prior default	**+8.6pp**

Method: measured by split-half / cross-fit on cached per-arm rankings — held-out, parameter-free, and **never worse** than the prior default on any corpus (fails soft when a document carries no metadata).

Multilingual VALIDATION-STAGE — NOT YET SHIPPED

LANGUAGE	BEFORE	AFTER (VS ENGLISH PARITY)
Hindi	11.8%	**~100%**
Telugu	35.3%	**~100%**
Tamil	41.2%	**98%**

Method: cross-language gold-recall retention vs English, recall@6, N=47/language. Validated in isolation; production cutover is staged — this is benchmark evidence, not a live feature yet.

Live run on a real company corpus END-TO-END, ON REAL DATA

PATH	ACCURACY	FALSE-ACCEPT	TOKENS
qbrin full path	**30%**	**3.3%**	~9,000
qbrin cheap router	12.5%	10%	**~2,876**
traditional RAG	20%	6.7%	~2,536

Honest read: qbrin's full path beats traditional RAG by **+10pp accuracy and ~2× fewer fabrications** on real company data — but at higher token cost; the cheap router trades accuracy for ~3× lower cost. Absolute numbers here are **understated**: this run used a lightweight stand-in answer model, so it's a floor, not the production figure. The headline numbers in this report come from the dedicated benchmarks above, not this single live run.

Bottom line. qbrin leads on trust, safety and cost across every system we've benchmarked, and is competitive — not dominant — on raw answer-accuracy, by design. Want to verify on your own data? Run a controlled, admin-approved pilot. **Figures reflect benchmark evidence; results on your data may vary.** · qbrin.com