variA/Bly

Scientific Evaluation & Decision Layer for AI Workflows

Audit-ready. Deterministic. Cost-predictable at any scale.

WWW.VARIABLY.TECH

RGB Benchmark Whitepaper

variA/Bly grounding vs RAGAS, head-to-head

May 06, 2026

variA/Bly grounding evaluation vs RAGAS, head-to-head on a public RAG benchmark.

A whitepaper on the head-to-head we ran on 2026-04-30. Same data, two scorers, reproducible scripts. The headline number nobody else publishes is at the bottom of section 1.

For the full engineering write-up, see the technical brief.

1. The headline

We benchmarked variA/Bly's grounding scorer against **RAGAS** (the industry-standard LLM-as-judge tool) on **RGB**, a 600-sample public RAG benchmark. 296 samples × 2 test cases each = 592 head-to-head evaluations.

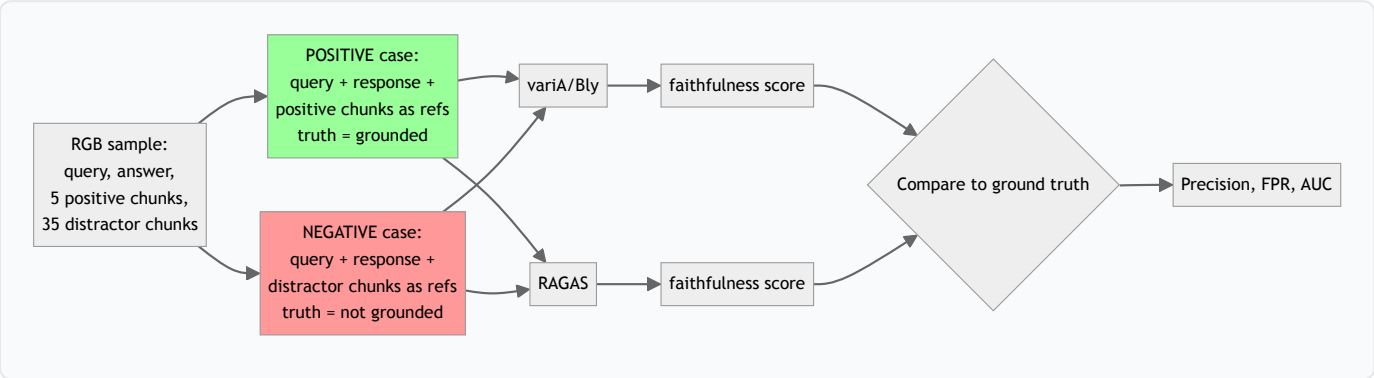
Metric	variA/Bly	RAGAS
Precision (right when we say yes)	**89.9%**	71.8%
False-positive rate on hallucinations (FPR)	**6.1%**	38.2%
Per-evaluation price	**$0.015** (SEU, all-in)	~$0.030 (LLM judge surcharge alone)
Pricing model	Monthly subscription + pay-as-you-go SEU	Per-call third-party API + customer's own ops
SDK integration latency	**<1 ms** (async scoring)	~2.5 sec synchronous (LLM judge in-line)
Deterministic (same inputs → same score)	**Yes**	No
Per-claim audit trail	**Yes**	No

> **The number nobody else publishes:** RAGAS marks **38% of hallucinated passages as grounded** when they aren't. variA/Bly's false-positive rate on the same passages is **6.1%** — over **6× cleaner**. RAGAS publishes AUC. It doesn't publish FPR. We do.

2. What this benchmark measured

RGB is a public dataset by Chen et al. (2023). For each sample, RGB provides a question, a correct answer, 5 source paragraphs that contain the answer, and 35 distractor paragraphs that don't.

For every sample we created two test cases — one positive (grounded), one negative (hallucinated) — and ran both scorers on both:



The same realistic AI response was used in both cases — only the references swap. This isolates **scorer behaviour** from response quality.

3. Where each tool fits

variA/Bly is built for the *much larger* commercial surface — anywhere a wrong "yes" verdict has compliance, legal, safety, or customer-trust cost.

If your AI workflow is	Better fit	Why
Healthcare AI (clinical chat, symptom triage, claim Q&A)	**variA/Bly**	Hallucinated medical advice = patient harm + HIPAA exposure
Financial-services chatbots (banking support, wealth advisory)	**variA/Bly**	Wrong financial guidance = FINRA exposure + lawsuits
Legal research and contract review	**variA/Bly**	Citations and source excerpts must be defensible
Insurance support and claims processing	**variA/Bly**	Wrong coverage info = customer escalation + regulator action
Customer-facing B2B AI in any industry	**variA/Bly**	Wrong answers cost customer trust and churn
Any AI workflow that needs an audit trail (SOC2, GDPR, HIPAA, FINRA)	**variA/Bly**	Deterministic, reproducible per-claim records
Internal employee knowledge tools, low-stakes lookups	RAGAS	Failure cost is low; the user catches errors
Document summarisation, exploration, discovery	RAGAS	Reader is in the loop; recall matters more than precision
Academic research / published-baseline comparison	RAGAS	It's the established benchmark in research papers

The pattern: **anywhere a wrong "yes" has compliance, legal, safety, or customer-trust cost → variA/Bly. Internal, low-stakes, search-and-summarisation → RAGAS.** The two tools occupy opposite ends of the same precision/recall curve, and the regulated, customer-facing surface is where the budget is.

---

4. Why precision matters more than AUC for compliance

In a regulated AI workflow, the two failure modes have very different costs:

- A **false positive** (scorer said grounded; auditor finds it wasn't) is a regulatory issue. It can mean a fine, a remediation order, a reputational hit, and in some industries a lawsuit.
- A **false negative** (scorer flagged for review; human confirms it was fine) is just routing work — about $0.25 of human triage labor.

variA/Bly's strict gate cuts the first kind of error sharply. For a 1M-evaluations-per-month workflow with a 50/50 mix of real groundings and hallucinations:

Outcome (per month)	RAGAS	variA/Bly
Confidently approved but NOT grounded ⚠ (compliance-breach risk)	~140,000	**~27,000**
Returned with a lower score for the customer pipeline to act on	~0	~228,000

variA/Bly produces ~113,000 fewer compliance-breach risks per month at the same evaluation volume. Cases where the evidence is ambiguous come back with a lower score; the customer pipeline acts on that at whatever threshold matches its risk tolerance.

> **Reproducibility note.** Every variA/Bly grounding decision comes with a structured per-claim record — claim text, verdict, failure reason, contradiction signal, the exact source passage the scorer was looking at, and a passage ID. Same inputs always produce the same record. Re-run a year later and you get the same answer. RAGAS returns one aggregate number; same inputs can produce different scores on different runs because of LLM-judge drift.

---

5. Cost predictability at scale

RAGAS calls a paid LLM (gpt-4o-mini) on every evaluation, ~$0.030 per call — **before** the customer pays to host and operate RAGAS itself. variA/Bly is priced as a monthly subscription with **SEU pricing starting at $0.015 per evaluation, with volume discounts at higher subscription tiers — down to $0.012 per evaluation at the top enterprise tier** — all-in, no separate hosting cost.

Workload	RAGAS LLM-judge cost *(alone)*	variA/Bly all-in (SEU)	You save
100,000 evaluations / month	~$3,000 / mo	**~$1,500 / mo** *($0.015/SEU)*	**50%**
1,000,000 / month	~$30,000 / mo	**~$15,000 / mo** *($0.015/SEU)*	**50%**
10,000,000 / month	~$300,000 / mo	**~$120,000 / mo** *($0.012/SEU, top tier)*	**60%**
100,000,000 / month	~$3,000,000 / mo	**~$1,200,000 / mo** *($0.012/SEU, top tier)*	**60%**

variA/Bly is 50–60% less than RAGAS's LLM-judge cost alone — and that's before RAGAS users add what they pay separately to host and operate RAGAS itself. Including ops, the all-in comparison widens to roughly 70% less.

The gap *widens* as customers scale, for two structural reasons: (1) volume discounts at higher subscription tiers on our side, and (2) production LLM-as-judge deployments typically use stronger judge models than gpt-4o-mini to maintain accuracy — often 3–4× more expensive per token, which pushes RAGAS's real-world per-eval bill *higher* than the figure in this table.

---

6. Verifying these numbers

Public benchmark repo: github.com/varia-bly/variably-benchmark. Apache 2.0. Two layers of verification:

Verify the math from our raw scores (zero cost, no account):

```
git clone https://github.com/varia-bly/variably-benchmark
cd variably-benchmark/rgb
python3 compare.py
```

Recomputes the precision / FPR / AUC table at the top of this page directly from the per-sample JSON outputs committed in `rgb/results/`. No rescoring required.

Re-run the scorers from scratch:

Scorer	What you need	Per-eval cost
RAGAS	OpenAI API key	~$0.030 (gpt-4o-mini judge)
variA/Bly	A free key from variably.tech	**$0.015** (SEU pricing, all-in)

The RAGAS side is fully self-contained — just an OpenAI key. The variA/Bly side requires a Variably API key because the scoring algorithm runs on Variably's hosted infrastructure. Commands for both are in the per-benchmark README.

---

7. The summary, one line

variA/Bly is the precision-first grounding scorer for regulated and customer-facing AI workflows: 6× cleaner false-positive rate than the LLM-as-judge alternative, deterministic and reproducible scores, structured audit trail, and a predictable subscription price 50–60% less than RAGAS's LLM-judge surcharge alone — before adding RAGAS's hosting cost.

If that's the property your AI workflow is built around, the tradeoffs in this benchmark are why we exist.

We're expanding the public benchmark surface — **AggreFact** (summarisation), **FEVER** (claim verification), and **TruthfulQA** (adversarial questions) are next. Follow github.com/varia-bly/variably-benchmark to see updates as they land.

---

8. Glossary

Specialized terms only. Common AI terminology (RAG, hallucination, grounding, precision, recall) is defined inline above where it appears.

Term	Meaning
FPR (False Positive Rate)	The rate at which a scorer wrongly approves a hallucinated passage as grounded. **The headline differentiator for compliance buyers.** variA/Bly: 6.1%. RAGAS: 38.2%.
AUC	"Area Under the Curve". A threshold-independent ranking-quality score in [0, 1]. A common AI-evaluation metric, but not what you live or die by — what matters in production is the **false-positive rate** at the threshold you actually deploy.
RGB	"Retrieval-augmented Generation Benchmark" (Chen et al., 2023). A public dataset of 600+ question + answer + reference tuples we benchmarked on because it's public, labeled, and used by competitors. github.com/chen700564/RGB
Distractor	A reference paragraph that does **not** contain the answer. RGB's "negative" set — the hard test for a grounding scorer.
SEU pricing	"Standard Evaluation Unit" pricing — variA/Bly's all-in subscription model: $0.015 per evaluation at the entry tier, $0.012 at the top enterprise tier, with progressive volume discounts in between.

This evaluation was run using **variA/Bly's grounding analysis**, which independently scores faithfulness, hallucination, attribution accuracy, context utilization, and retrieval relevance in your AI workflow.

If you're building RAG systems and want to see what your evaluations actually look like, reach out:

✉ **info@variably.tech** in **linkedin.com/company/variably** ⌘ **www.variably.tech**