variA/Bly

Scientific Evaluation & Decision Layer for AI Workflows

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

WWW.VARIABLY.TECH

RGB Benchmark - Business Buyer Brief

Compliance, audit, and the math at scale

May 06, 2026

Audience: compliance leads, operations directors, risk officers, heads of customer support, AI program owners. People who buy AI tools but don't necessarily build them. **Length:** 3-4 pages. No AI/ML background required. **Bottom line:** variA/Bly catches the AI failure mode that costs you a compliance breach. We benchmarked head-to-head against RAGAS, the industry's most-cited alternative, and we're **6× cleaner** on the failure that hurts you most - while costing **50–60% less** per evaluation.

1. The headline

Metric	variA/Bly	RAGAS	What it means in business terms
Precision when AI says "this is grounded"	**89.9%**	71.8%	When variA/Bly approves an AI response, it's right **9 out of 10** times. RAGAS is right 7 out of 10.
Wrong-yes rate on hallucinations (FPR)	**6.1%**	38.2%	RAGAS approves ~38 out of 100 hallucinated passages. variA/Bly approves ~6. **6× cleaner on the failure mode that hurts compliance.**
Per-evaluation cost	**$0.015**	~$0.030	variA/Bly is **50% cheaper** than RAGAS's AI-judge cost alone, before adding RAGAS's hosting bill.
SDK integration latency	**<1 ms**	~2.5 sec	variA/Bly evaluates *off* your customer's request path. RAGAS sits on it.
Deterministic scoring	**Yes**	No	Same input → same score, every time. RAGAS uses an AI judge that can drift between runs.
Per-claim audit trail	**Yes**	No	variA/Bly returns a structured record per claim ("this claim failed because of contradiction with policy section 4.2"). RAGAS returns one aggregate number.
Pricing model	**Monthly subscription**	Per-call third-party API + customer's own ops	Predictable spend at any scale.

A quick story to make it concrete.

HealthCheck, a US health-insurance company, runs an AI chatbot for customer support. A customer asks: *"Does my plan cover an MRI for my knee?"* The chatbot reads the customer's policy and replies:

> *"Yes, your plan covers MRI scans. Pre-authorization is not required for MRIs."*

But the actual policy says pre-authorization **is required**. The chatbot got the first sentence right and **made up** the second one. The customer drives to the imaging center, gets turned away, and calls back angry — $7 in agent time, plus a churn risk, plus a complaint that may end up on the regulator's desk.

This is what AI tools call a **hallucination**: the AI made up something that isn't in the source documents.

variA/Bly catches that hallucinated sentence before it ships to the customer. RAGAS lets it through ~38% of the time.

That single number — 6.1% vs 38.2% — is the variA/Bly thesis in one chart: *be right when you say yes*, even on adversarial passages designed to trick the scorer. It's the property regulated buyers refuse to compromise on, and it's the one nobody else is publishing.

---

2. Cost predictability at scale

RAGAS uses a paid third-party AI model (gpt-4o-mini) as its judge. That AI-judge call costs about 3 cents per evaluation — and the customer also has to pay separately to host and run RAGAS itself.

variA/Bly is priced as a monthly subscription with **SEU pricing starting at $0.015 per evaluation, with progressive volume discounts at higher subscription tiers — reaching $0.012 per evaluation at the top enterprise tier** — all-in, no separate hosting cost.

Workload	RAGAS AI-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 AI-judge cost alone — and that's before adding what RAGAS users pay separately to operate RAGAS itself. Including RAGAS hosting and ops, the all-in comparison widens to roughly 70% less.

This matters for two reasons:

1. **AI workloads grow fast.** A customer-support chatbot that started at 100K evaluations/month often hits 1M within a year and 10M within two. With RAGAS, the AI-judge line item alone grows from $36K/yr to $360K/yr to $3.6M/yr on the same product — *before* hosting, *before* the inevitable judge-model upgrade to a stronger (more expensive) model. variA/Bly stays 50–60% below that line as you scale.
2. **Budget approval is easier.** Procurement can plan for a tiered subscription with a known per-evaluation rate. Procurement struggles with per-call third-party API surcharges that depend on OpenAI's published price list and on which judge model the compliance team requires.

> **Reproducibility note.** variA/Bly's scoring is deterministic and the per-claim record is structured and timestamped — same inputs always produce the same score, so any decision can be re-run a year later and you get the same answer. That's the audit-trail property regulators expect. RAGAS's AI judge can drift between runs, and the output is a single aggregate number rather than a structured record.

---

3. Where each tool fits

We're not saying RAGAS is bad. We're saying RAGAS is optimised for a **different kind of AI workflow**: low-stakes search, exploration, and internal-knowledge summarisation, where high recall is the goal and a few wrong "yes" verdicts don't hurt anyone. variA/Bly is built for the *much larger* commercial surface — anywhere a wrong "yes" has compliance, legal, safety, or customer-trust cost.

Use case	Better fit	Why
Healthcare AI assistants (clinical chat, symptom triage, claim Q&A)	**variA/Bly**	Hallucinated medical advice = patient harm + HIPAA exposure. False positives are unacceptable.
Financial-services chatbots (banking support, wealth management)	**variA/Bly**	Wrong financial advice = FINRA exposure + lawsuits. Compliance team needs the audit trail.
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 complaints + regulator escalation.
Pharmaceuticals: regulatory and product info	**variA/Bly**	FDA-grade audit trail is required.
Government / public-sector AI	**variA/Bly**	Citizen-facing decisions need reproducible records.
Customer-facing B2B AI in any industry	**variA/Bly**	Wrong answers cost customer trust and contract churn.
Any AI workflow that needs an audit trail (SOC2, GDPR, HIPAA, FINRA)	**variA/Bly**	Deterministic, reproducible per-claim records.
Internal search assistance over a knowledge base	RAGAS	High-recall retrieval style; the user can spot a wrong answer themselves.
Internal employee knowledge tools (low-stakes lookups)	RAGAS	Mostly internal staff; failure cost is low.
Document summarisation / exploration tools	RAGAS	Recall matters more than precision; reader is in the loop.
Academic research benchmarks	RAGAS	It's the published baseline in academic comparison work.

The pattern is simple: **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 serve buyers with opposite priorities — and the regulated, customer-facing surface is where the budget is.

---

4. Summary

What you get	What it means in business terms
89.9% precision ("right when we say yes")	Fewer compliance breaches reaching production.
6.1% false-positive rate on hallucinations	When the AI's source doesn't support the claim, we catch it ~6× more often than the alternative.
SEU pricing — $0.015/eval entry, $0.012/eval top tier	Predictable monthly subscription bill, **50–60% less than RAGAS's AI-judge cost alone**.
<1 ms SDK integration latency	Scoring is async — it doesn't sit on your customer's request path.
Deterministic scoring	Auditors can reproduce any decision a year later. Internal SLOs are stable.
Per-claim audit trail	The compliance team has the records. The CTO has the explainability.

If you're operating AI in a regulated industry and an auditor's question keeps you up at night, this is the layer you want underneath your AI workflow.

We'd be glad to share a deeper walkthrough on a 30-minute call, or to run a confidential audit on a sample of your live workflow data.

---

5. 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 it's 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.
Distractor	A reference document that does **not** contain the answer. The hard test for a grounding scorer: does it correctly say "no, this passage doesn't support the answer"?
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**