A BUYER'S FRAMEWORK

How to evaluate *chemical* AI

Most AI evaluations open with the wrong question: **which model is smarter.** Everyone rents the same frontier models, and the badge for "best" rotates between several labs every few months. The real question is *which system knows your chemistry, your catalog, and your supply chain* and can be trusted to act on them.

THE TEN QUESTIONS

Judge any system on the messy field questions, not the demo.

Each question has the same shape: what goes wrong with a horizontal tool, then what good actually looks like.

01 Where does the answer come from?

A folder of uploaded documents goes stale the day after upload, and a 2019 marketing deck can outrank the current spec sheet at random.

THE STANDARD

Answers grounded in a live, governed product database, where every attribute has a source ranked by authority.

02 Is the answer consistent and controllable?

Prior chats and per-user history leak in invisibly, so the same question returns a different answer each time, and no one can see which inputs shaped it.

THE STANDARD

Context is governed, not accidental. Deliberate inputs like region, entitlements, or role shape the answer by design, while the same question under the same context returns the same answer every time.

03 Does it understand chemical data?

A generic system reads every attribute as undifferentiated text, so storage condition and suitable substrate weigh the same, and SKUs collapse into the base product.

THE STANDARD

A data model built for chemistry: typed attributes, units and test methods kept attached, regulatory status as a first-class field.

04 Does it know when it doesn't know?

Frontier models are tuned to sound certain. Ask for a recommendation without naming a decisive constraint, say the resin type, and most will quietly assume one and answer, instead of flagging what they need to know first.

THE STANDARD

Answers come only from retrieved data and cite the governing document, and when a decisive attribute is missing, the system asks for it instead of guessing.

05 What happens when an answer is wrong?

With a horizontal tool there is no path. You can't retrain the model, and a corrections document just adds one more conflicting source. The wrong answer is wrong again tomorrow, for everyone.

THE STANDARD

Kimia gives experts a place to correct the source: fix the attribute, re-rank the document, and the fix holds for every user, permanently.

06 Who controls who sees what?

Horizontal tools inherit file permissions at best. There's no way to let a distributor see application data while margin-sensitive fields stay internal, so you over-share or under-feed the AI.

THE STANDARD

Governance at every grain: workspaces, roles, and access down to the individual attribute, extending across the supply chain.

07 Does it survive past the pilot?

Any LLM demos well on a few hundred documents. At thousands of products and tens of thousands of files, plain retrieval dilutes and accuracy decays quietly, the worst way for it to decay.

THE STANDARD

Retrieval engineered for catalog scale: hybrid semantic + keyword search over structured data, evaluated at full volume.

6% of organizations piloting Copilot expanded to larger-scale deployment.

08 Who keeps it alive?

Statuses, regions, and regulations change constantly. With a document dump, someone has to remember to re-upload, so it doesn't happen. The license is the cheap part.

THE STANDARD

Change flows through the data model and propagates instantly; the platform carries maintenance, including model upgrades, as part of the product.

15–30% of build cost, in maintenance every year, indefinitely.

09 Where does it show up?

An internal chatbot is the easy 20%. On a horizontal stack, every other surface, whether the website, sample intake, distributor assistants, or CRM, is its own months-long integration project.

THE STANDARD

With **Kimia**, form factors are configuration on one governed layer: internal assistant, embeddable website Concierge, distributor-scoped assistants, CRM and API.

10 Does it produce signal, or just answers?

Every question your teams, distributors, and customers ask is market intelligence: demand, competitor mentions, portfolio gaps. Horizontal tools answer it and discard the signal.

THE STANDARD

Queries become a channel of demand signals, competitor mentions, and qualified leads, so the assistant pays for itself twice.

THE NUMBERS

Both alternatives look cheap at the entry point, and get expensive where no one models it.

95%

of enterprise GenAI pilots deliver no measurable P&L impact.

MIT NANDA, 2025

67 / 33

success rate, %: buying from specialists vs. building internally.

MIT NANDA, 2025

6%

of organizations that piloted Copilot moved to scale deployment.

Independent analysis

$66–87

true all-in cost per user / month, not the $30 list price.

Microsoft list pricing

Adopting a horizontal tool

- Requires an M365 E3/E5 base, so the true all-in runs **$66–87 per user / month**.
- **~$79K / year** for 100 seats, before implementation, training, or change management.
- Custom agents meter separately: $200/mo per 25K messages, plus variable Azure consumption.
- **3–6 month** rollout; add 6–12 weeks if data governance needs remediation first.

Building in-house

- **$200K–500K+** for a production-grade build (industry estimate).
- **15–30%** of build cost in maintenance, every year, indefinitely.
- **~$19K/mo** in tokens & infra at 100K queries/day, and retrieval is ~$12K of it.
- **$75K–150K/yr** in engineering per system; 6–12 months to impact.

The expensive part of both paths is the system around the model: data readiness, governance, evaluation, and maintenance. *That is precisely the part Kimia delivers as a product.* Gartner expects 60% of AI projects abandoned through 2026 for want of AI-ready data; S&P Global found 42% of companies abandoned most AI initiatives in 2025, up from 17% the year before.

Out of the box, horizontal AI is genuinely useful, so let your teams use it for drafting, summarising, and brainstorming. The work that carries real technical and regulatory weight, selling on specs and shipping on compliance, needs a system that clears all ten bars. *Run Kimia alongside the tools you already have, and let it own the answers that have to be right.*

Kimia is built to clear *all ten.*

Run the ten questions on your catalog →

Sources: MIT NANDA, *The GenAI Divide: State of AI in Business 2025*; Gartner; S&P Global Market Intelligence; Microsoft list pricing; independent analysis of Copilot adoption. Build and run-cost figures are aggregated industry estimates (2025–26) and are presented as ranges.