REFLEX WHITEPAPER

Building enterprise software that AI agents can operate

PUBLISHED

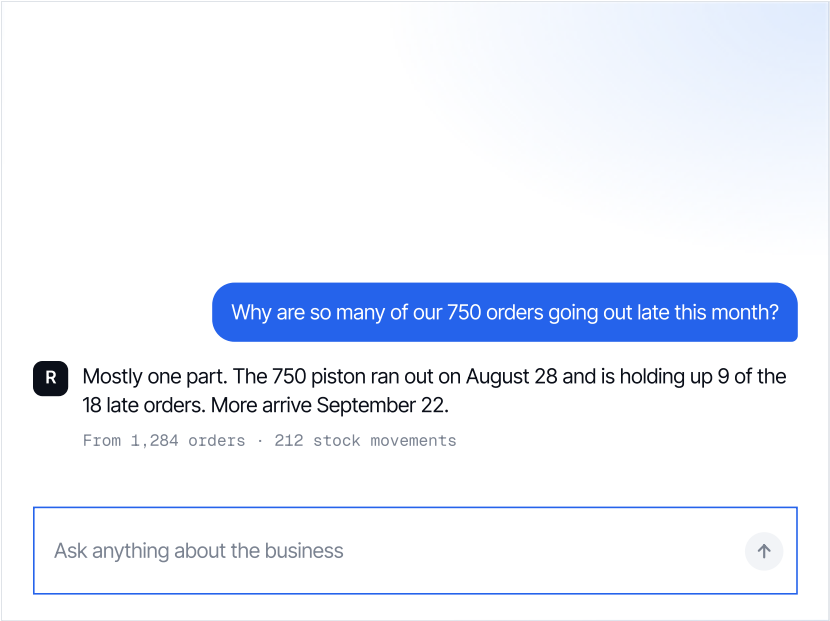
August 2026

SUBJECT

Enterprise architecture

AUDIENCE

Manufacturers



Summary

Enterprise software was built for people filling in forms. On realistic enterprise databases and multi-step workflows, the best available AI models fail most tasks.

The same approach that answers most questions correctly against a clean, well-named database fails against a legacy one. The bottleneck is the system, not the model. Reflex is built to be legible: a data model where every count traces to its transactions, inventory and accounting recorded together, and functions exposed as callable tools instead of screens to be scraped.

Same approach, academic schemas
86%
Spider 1.0 execution accuracy

Same approach, enterprise schemas
2.2%
Spider 2.0

Humans on chained workflows
93.9%
WorkArena++

GPT-4o on the same tasks
2.1%
WorkArena++

Key findings

- **Legibility, not capability, is the binding constraint.** Model performance falls by more than an order of magnitude when the schema stops being designed for comprehension.
- **Naming is a machine interface.** Peer-reviewed work shows identifier naturalness measurably changes how accurately a model can query a database.
- **Bolt-ons multiply the surface.** The average company now runs 101 applications, and manufacturing quoting is routinely sold as a separate product stitched to the ERP after the fact.
- **Most manufacturers have not crossed from pilot to production.** Only 29% run AI at facility or network scale, while 38% are still piloting generative AI.

What legacy ERP costs

Most ERP projects go live. Few change how the business operates.

Outcomes

In a survey of 170 organizations, 30% of ERP projects ran over budget and 22.3% ran over schedule. Productivity and efficiency were realized by 87.3% of organizations live for at least a year; new operating models, the reason transformations get funded, by 40.7%.

Realized benefits

Percent of organizations

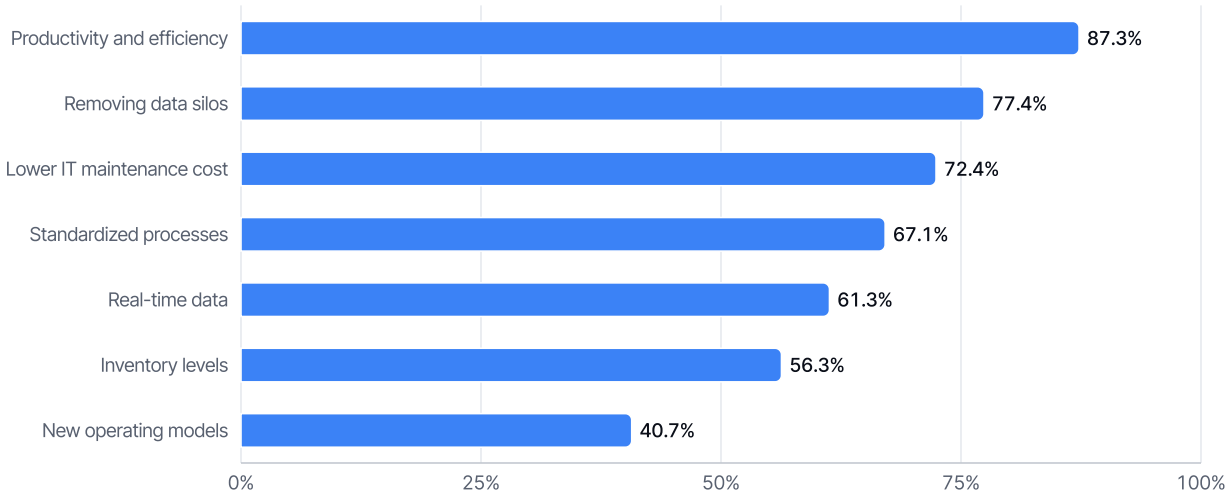


Figure 1. Benefits realized as expected. Share of organizations live at least one year reporting the benefit was realized to the extent expected.

Source: Panorama Consulting Group ERP Report, n = 170.

When it goes wrong

The largest failures follow one pattern: the business customizes the system until it is unique, then cannot operate or upgrade it.

Table 1. Documented ERP failures

COMPANY	REPORTED COST	WHAT HAPPENED
Lidl	≈ €500M	Seven years of SAP work abandoned in 2018 after the retailer customized the system to preserve its own inventory conventions rather than adopt the vendor's model.
Hershey	$112M	A compressed big-bang go-live in 1999 left roughly $100M of orders unfulfilled through the Halloween season; quarterly profit fell 19%.
Revlon	$64M	A 2018 go-live disrupted shipments; the company disclosed unfilled net sales, a material weakness in internal controls, and faced shareholder suits.

Publicly reported figures; see references.

Integration sprawl

The average company ran 101 applications in 2025, the first year that number crossed a hundred. In manufacturing the pattern is specific: quoting and configuration are sold as a

separate layer that sits between CRM and ERP, and every one of those layers is a new integration, a new data model, and a new place for the truth to diverge.

Table 2. Quoting sold as a layer on top of ERP

PRODUCT	ORIGIN	WHERE IT SITS
Vendavo Intelligent CPQ	Acquired Endeavor Commerce in 2017	Between CRM and ERP
Tacton CPQ	Constraint-based configuration for complex products	Between engineering and ERP
Configure One (Revalize)	Marketed to integrate with ERP, CRM, CAD, and eCommerce	Alongside ERP
Epicor CPQ	Add-on module inside an ERP suite	On top of ERP

In manufacturing, a quote is a commitment of capacity and a shipment is an accounting entry. Splitting quoting, scheduling, and the ledger across systems means reconciling them.

What AI-legible means

A system is legible to an agent when the agent can find the right data, understand what it means, and change it through a defined interface. Benchmarks measure each of these separately.

The schema evidence

Text-to-SQL asks a model to answer a question by writing a database query. On Spider 1.0, built from academic databases with small, clean schemas, the leading approach reaches roughly 86% execution accuracy. Spider 2.0 keeps the task and changes the databases to enterprise-realistic ones: large schemas, multiple dialects, multi-step workflows. The same approach scores 2.2%.

Text-to-SQL execution accuracy

Percent

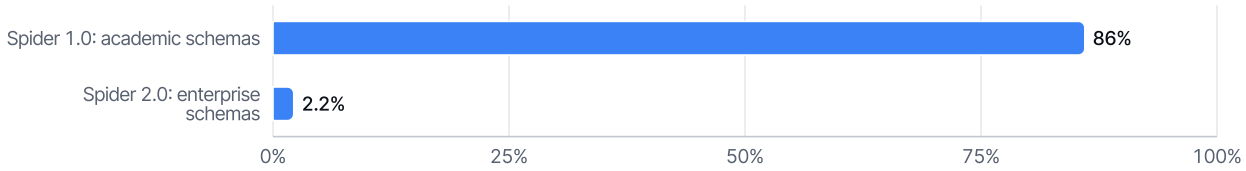


Figure 2. The same approach, two schema regimes. Execution accuracy. The task did not change; the databases did.

Source: Spider 2.0, ICLR 2025.

Purpose-built systems do better on Spider 2.0, but the best scores about 31%.

Systems on Spider 2.0

Percent

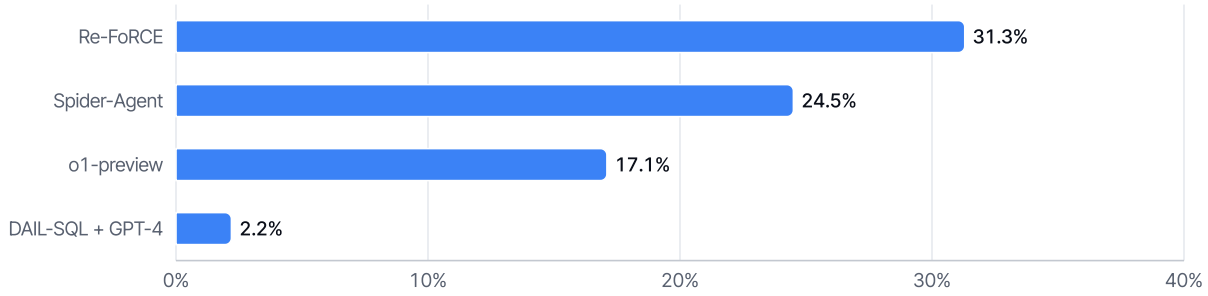


Figure 3. Spider 2.0 leaderboard. Selected systems, execution accuracy on enterprise-realistic databases.

Source: Spider 2.0.

The agent evidence

On WorkArena++, which chains ordinary enterprise tasks into realistic workflows, humans succeed 93.9% of the time and GPT-4o succeeds 2.1%.

Success rate

Percent of tasks completed

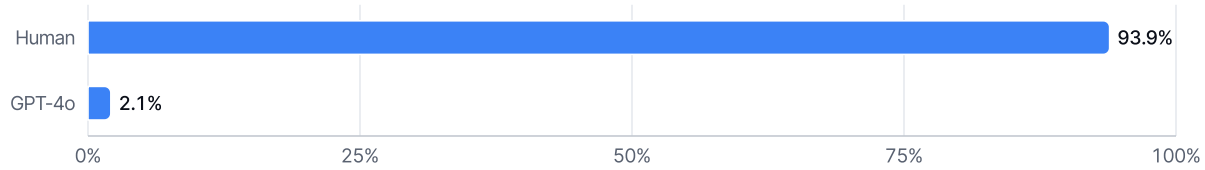


Figure 4. Chained enterprise workflows. Success rate on WorkArena++, which composes ordinary enterprise tasks into multi-step work.

Source: ServiceNow Research, NeurIPS 2024 Datasets and Benchmarks.

Salesforce's CRMArena-Pro shows the same curve inside a business system: agents complete about 58% of single-turn tasks and around 35% once the task becomes a conversation. The failures are routine: filtering a list, following a process across screens, and knowing what not to disclose.

Naming is an interface

In peer-reviewed work published in the Proceedings of the ACM on Management of Data, the naturalness of schema identifiers, real words against abbreviations, measurably changes how accurately a model can query a database.

A column called `qty_remaining` is easier for a model to use correctly than one called `qr3`, even though both hold the same number.

How Reflex is built

Reflex was designed for agents from the start. The properties that make it legible to a model also make it auditable.

The data model

Five design rules keep the data consistent and readable, for people and for models.

01	**Every count has a paper trail**	Stock levels and order quantities are built from the receipts, shipments, and usage recorded against them. When a number looks wrong, you can trace it back to the exact transactions that produced it, and any manual adjustment is recorded with who made it and why.
02	**Inventory and accounting as one record**	When stock moves, its accounting entry is created in the same step. Inventory and the ledger start from one record instead of two systems copying data to each other, so there is less to reconcile at month end and differences are easier to trace.
03	**Products understood from the parts up**	The system knows which parts go into every product, so part availability, reorder points, and lead-time estimates are worked out from those parts and their supplier history instead of guessed at the product level.

04	**Names a person would use**	Data is labeled in plain business language, so a model can understand it the way an employee would, without a translation layer.
05	**Options without duplication**	Product options such as size or finish belong to one product instead of creating separate records, which keeps the catalog, and every question about it, simple.

Answering a question

A manager asks a question in plain language. Reflex works out what the question refers to, gathers the relevant records, and returns a report in which every figure links back to where it came from. Figure 5 shows one question end to end.

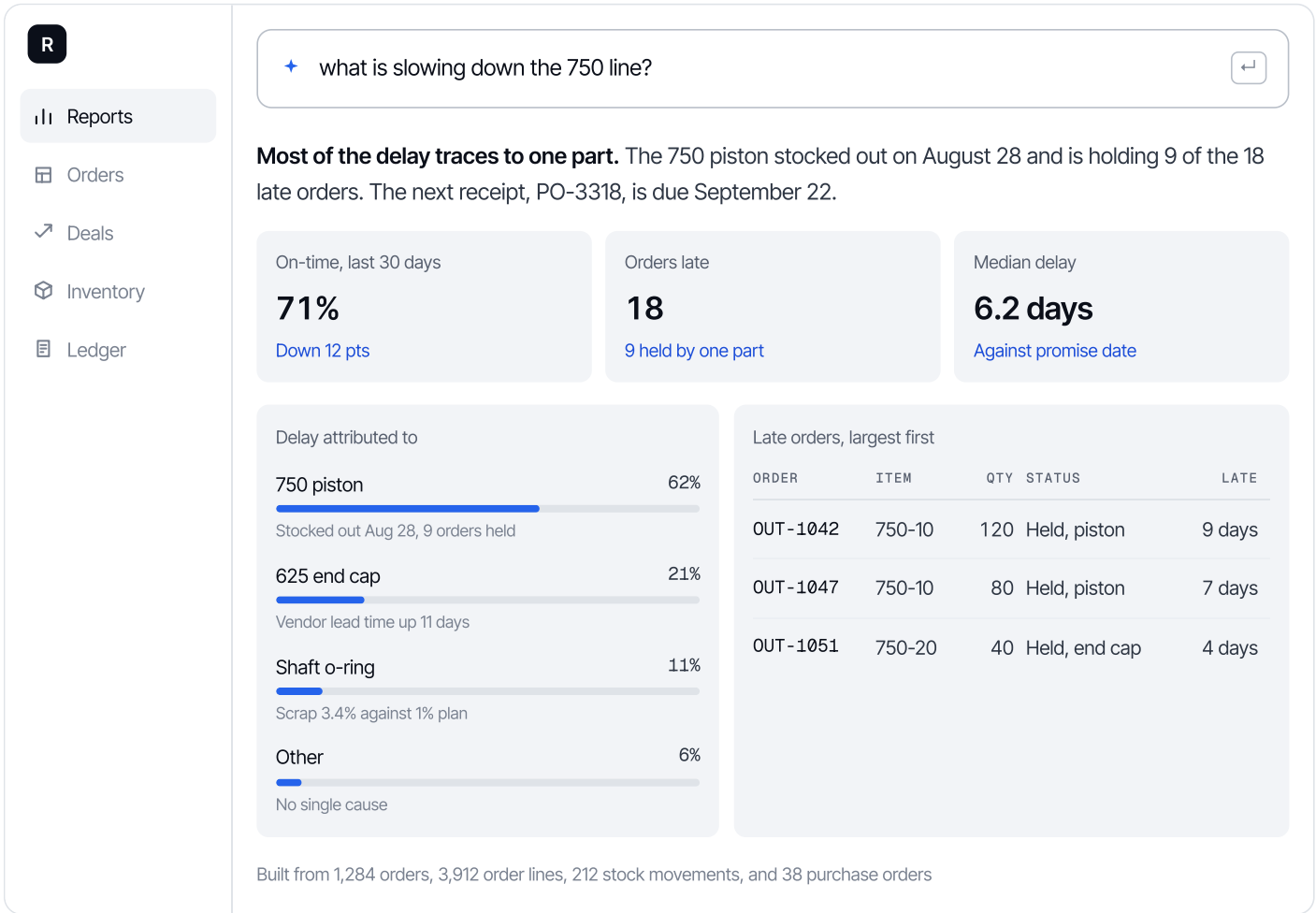


Figure 5. One question, end to end. The question is matched to two products, the relevant records are checked, and each late order is traced to a cause. Drawn illustration with representative data.

The answer in Figure 5 comes from four steps, each made possible by the rules above.

1. **Understand.** "The 750 line" is matched to the two products it refers to, because the data uses the same words the business does.
2. **Gather.** Open orders, the products on them, and the parts those products need are collected in one pass.
3. **Check.** Stock, deliveries, and supplier lead times are read for each part, from the same transactions the rest of the business uses.
4. **Explain.** Each late order is traced to the part that held it, and the report cites the orders, receipts, and purchase orders it used.

The same steps answer questions no report was built for: which vendor's lead time moves the schedule most, which product line consumes the margin, where work in progress is sitting. Quoting, scheduling, inventory, and the ledger are one system, so no integration sits between the question and the answer.

One ledger

In most manufacturers the inventory sub-ledger and the general ledger are separate systems reconciled on a schedule, and inventory misstatements trace overwhelmingly to quantity errors, perpetual records that stopped matching the floor. In Reflex, a movement of material and its effect on the books are recorded together, so month-end reconciliation starts from one shared history instead of two systems that have drifted apart.

Table 3. What the system keeps current

MEASURE	WHAT IT TELLS YOU
Incoming	Stock on order that will replenish inventory, with expected arrival dates
On hand	What is physically in the building, counted from receipts and usage
Committed	Stock already promised to open orders
Projected usage	How quickly stock will be consumed, including demand from the assemblies that use it
Reorder signal	Raised when projected stock will fall below its minimum before new stock arrives
Capacity load	Work scheduled on each machine and line against what it can actually run

Recalculated whenever something changes, so planning and capacity models always start from the current state of the floor.

A callable surface

The Model Context Protocol, published by Anthropic in November 2024 and adopted by OpenAI in March 2025, standardizes how a model calls a system's functions rather than operating its user interface.

Reflex exposes its operations as defined, permissioned functions: enter an order, schedule a release, receive against a purchase, post an adjustment. An agent calls the same function a person triggers, subject to the same validation, and the history records which one did it.

Where the market is

AI in manufacturing

Percent of respondents

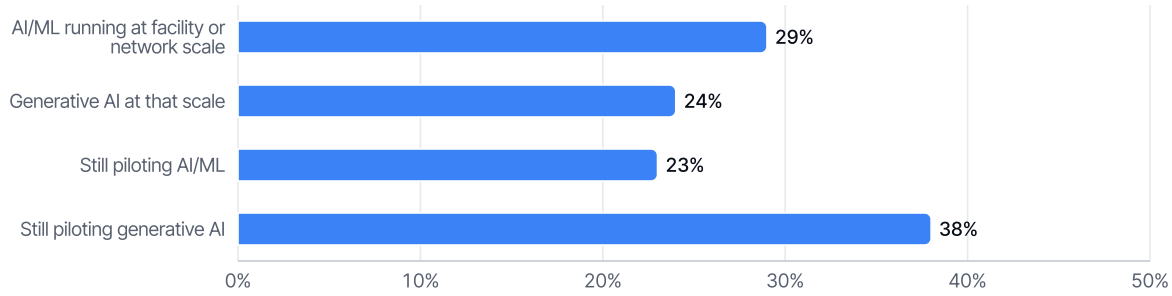


Figure 6. Production versus pilot. Share of 600 manufacturing executives at companies above $500M revenue.

Source: Deloitte, 2025 Smart Manufacturing and Operations Survey.

More manufacturers are piloting generative AI than running AI at scale. Pilots stall where the benchmarks do: reading a real database and acting through a real interface.

Notices & references

On the evidence

Benchmark results move quickly; the figures here reflect published leaderboards and papers as of September 2026 and should be re-checked before reuse. Vendor-published claims about semantic layers improving model accuracy are directionally consistent with the peer-reviewed schema-naming work but are not independent, and are therefore not cited as evidence here. Widely repeated ERP failure statistics that could not be traced to their claimed source are omitted.

Trademarks

SAP is a trademark of SAP SE. Salesforce and CRMArena are trademarks of Salesforce, Inc. ServiceNow and WorkArena are trademarks of ServiceNow, Inc. All other marks belong to their owners and are used for identification only.

Image credits

The cover animation and the report view in Figure 5 are drawn mock-ups of the Reflex interface, not screenshots.

References

1. Panorama Consulting Group, ERP Report (n = 170, survey period January 2025 – January 2026) (2026). www.panorama-consulting.com/resource-center/erp-report/
2. Computer Weekly, Lidl dumps €500m SAP project (2018). www.computerweekly.com/news/252446965/Lidl-dumps-500m-SAP-project
3. Computer Weekly, SAP disruption leads to Revlon class action (2019). www.computerweekly.com/news/252464278/SAP-disruption-leads-to-Revlon-class-action-lawsuit
4. CIO, Supply Chain: Hershey's Bittersweet Lesson (1999 events). www.cio.com/article/270245/supply-chain-management-supply-chain-hershey-s-bittersweet-lesson.html
5. Okta, Businesses at Work 2025 (March 2025). www.okta.com/newsroom/articles/businesses-at-work-2025/
6. Anthropic, Introducing the Model Context Protocol (November 2024). www.anthropic.com/news/model-context-protocol
7. MCP, Model Context Protocol specification (2025). modelcontextprotocol.io/
8. Spider 2.0 (ICLR 2025), Spider 2.0: evaluating language models on real-world enterprise text-to-SQL workflows (2024–2025). spider2-sql.github.io/
9. Proceedings of the ACM on Management of Data, SNAILS: Schema Naming Assessments for Improved LLM-Based SQL Inference (2025). doi.org/10.1145/3709727
0. ServiceNow Research (arXiv:2403.07718), WorkArena and WorkArena++: benchmarking web agents on enterprise knowledge work (2024). arxiv.org/abs/2403.07718
1. Salesforce AI Research (arXiv:2505.18878), CRMArena-Pro: holistic assessment of LLM agents across business scenarios (June 2025). arxiv.org/abs/2505.18878
2. Deloitte, 2025 Smart Manufacturing and Operations Survey (n = 600) (May 2025). www.deloitte.com/us/en/insights/industry/manufacturing/2025-smart-manufacturing-survey.html
3. Vendavo, Vendavo acquires Endeavor Commerce (July 2017). www.vendavo.com/company/press/vendavo-acquires-endeavor-commerce/