SolvxAI
Data to Decision Powered by AI

SOLVXAI WHITE PAPER SERIES | NO. 06 | MARKET & READINESS

The Agentic Gap

The distance between what an agent demonstration proves and what an asset team needs, measured against the published record

Published by the SolvxAI Research Team
July 2026 · Edition 1.0
Contact: sd@nexaconsultinc.com

IN BRIEF

Agent demonstrations are genuinely impressive and mostly honest. They are also measurements of the wrong thing. This paper collects what the published record shows about five specific gaps between a working demonstration and a deployable engineering system: performance that degrades as problems become real, correct answers reached by unreviewable derivations, an inability to stop that is close to independent of capability, memory that does not survive across sessions, and a benefit that shrinks or reverses precisely where expertise is highest. Each gap is evidenced rather than asserted, including the evidence that cuts against the argument. The purpose is not to argue that agents do not work. It is to give a buyer a way to tell the difference between a demonstration and a system.

Contents

1. Executive summary

The demonstration is real. It is measuring something other than what you are buying.

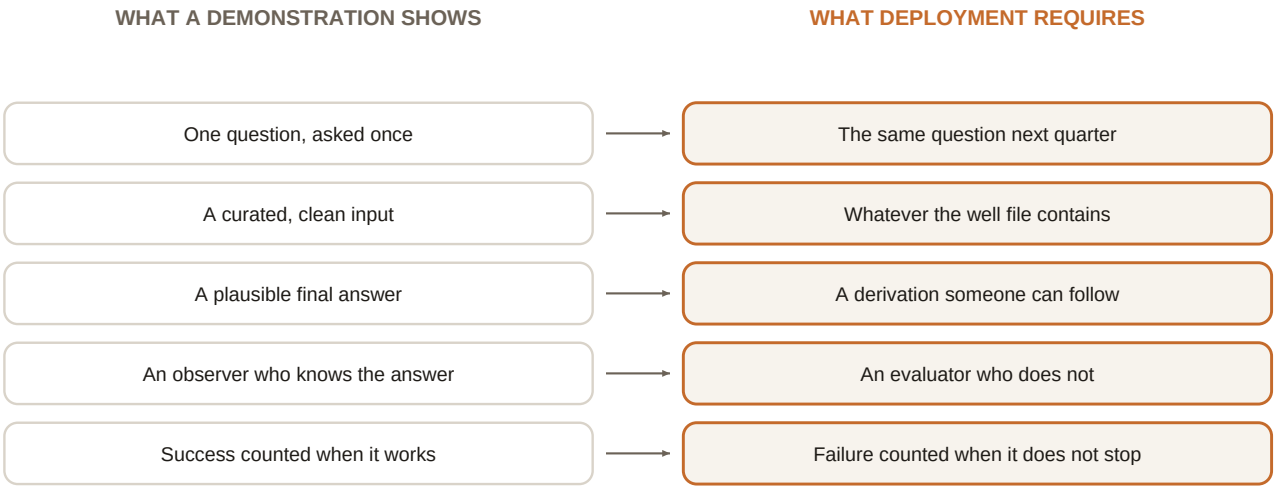
Five gaps recur in the published record. Each is documented, each has been measured by someone with no interest in this argument, and each is invisible in a demonstration by construction.

- **The reality gap.** A published six-agent system for reservoir history matching improved weighted error by **95% on a toy benchmark, 69% on a moderate one, and 13% on a real field**. [1] Performance collapses precisely as the problem becomes the one you have.
- **The trace gap.** On a benchmark spanning nine engineering domains, the best of twenty-seven models reached **65.4% final-answer accuracy but only 42.7% reasoning-trace fidelity**, and **79.5% of frontier-model errors were arithmetic** rather than conceptual. [2] A correct answer does not certify the derivation that produced it.
- **The stopping gap.** Across seventeen frontier models in four agent configurations, the best achieved **59.5%** on paired tasks requiring both acting and abstaining correctly — and abstention was found "largely independent of general task-solving capability." [3] Agents were observed executing irreversible actions *before* recognising they should have stopped. [3]
- **The memory gap.** A peer-reviewed benchmark reports a **30% accuracy drop** for commercial assistants and long-context models on retaining information across sustained interaction. [4]
- **The expertise gap.** Experienced developers working in repositories they knew well were measured **19% slower** with AI assistance, while believing they had been 20% faster. [5]

A demonstration cannot expose any of these. It runs once, on a chosen problem, with a curated input, in front of someone who already knows the answer, and it is judged on whether the output looks right. Every one of the five failures above is invisible under those conditions.

2. What a demonstration actually measures

This is not an accusation of dishonesty. Most demonstrations are faithful representations of the system doing the thing it was asked to do. The problem is structural: the conditions of a demonstration remove exactly the variables that determine whether a system is deployable.



Nothing in the left column is dishonest. It is simply not a measurement of the right column.

Figure 1. The five substitutions a demonstration makes. Each is reasonable in isolation; together they remove every condition under which the failures in section 3 would become visible.

The most consequential substitution is the last. A demonstration is scored on whether it produced a good answer. A deployed engineering system has to be scored on whether it produced a bad one *without saying so* — because that is the failure that reaches a reserves report.

> **A demonstration is judged on its successes. An engineering system has to be judged on the shape of its failures.**

3. The five gaps, with the evidence

3.1 Performance degrades as the problem becomes real

The most directly relevant study available built a six-agent system for reservoir history matching, on the architecture the market now describes as agentic: an orchestrating planner, specialist agents, retrieval over simulator documentation, a non-model simulator in the loop, and human checkpoints. It is competent, honest work by authors with no stake in this paper's argument.

The results are the point. Weighted normalised error improved by 95% on a toy benchmark, 69% on a moderate one, and **13% on a real field**. [1] The gradient runs in the direction of realism, and it is steep.

A second study makes the same point about consistency rather than accuracy: on graduate-level computational mechanics tasks with five attempts each, a leading model solved 30 of 33 tasks *at least once*, but only 26 of 33 *on every attempt*. [6] A demonstration shows you the first column. Reserves reporting requires the second.

3.2 A right answer does not certify the derivation

On a contamination-resistant engineering benchmark spanning nine domains and twenty-seven models, the best model reached 65.4% final-answer accuracy while achieving only 42.7% fidelity on the reasoning trace — and 79.5% of frontier-model errors were arithmetic rather than conceptual. [2]

For most commercial uses the answer is the product and the derivation is incidental. In reserves work the relationship is inverted: the derivation is the deliverable, because it is what a qualified evaluator reviews and what an auditor follows. A system that reaches good answers by unreviewable routes is not partially suitable for that work.

3.3 The inability to stop is not fixed by capability

This is the gap most likely to be missed, because it is the one a demonstration is structurally incapable of showing — nobody demonstrates a system declining.

A benchmark built specifically to measure abstention across twenty frontier models found the problem "unsolved, and one where scaling models is of little use," with the counter-intuitive finding that reasoning fine-tuning **degrades** abstention by 24% on average. [7] The model with the highest answer accuracy in that study scored near the bottom on knowing when not to answer.

In agent settings it is worse, because an agent acts. Across 263 paired tasks in 42 executable environments and seventeen models, the best configuration reached 59.5% paired accuracy, and the authors report that abstention capability is "largely independent of general task-solving capability, indicating that scaling task-solving alone will not close this gap." [3] They name the failure mode precisely: **post-hoc abstention**, where agents "execute irreversible actions before recognizing abstention triggers." [3]

> **WHY THIS MATTERS MORE IN AUTONOMY THAN IN CHAT**
>
> A chat assistant that should have declined produces a bad answer someone may catch. An autonomous run that should have declined produces a bad answer, and everything downstream of it, and the recognition arrives after the actions do. The published rate at which current systems get both halves right is under 60%. [3]

3.4 Memory does not survive the session

Agent demonstrations run inside one conversation. Engineering work does not.

A peer-reviewed benchmark evaluating long-term interactive memory across five abilities — information extraction, multi-session reasoning, temporal reasoning, knowledge updates and abstention — reports a 30% accuracy drop for commercial assistants and long-context models on retaining information across sustained interaction. [4] A second peer-reviewed benchmark, built because prior work had evaluated no more than five conversation sessions, finds models struggle with long-range temporal and causal dynamics and that retrieval augmentation, while it helps, still lags human performance substantially. [8]

The practical consequence for an asset team is that the second study on the same field starts from nothing, and the system cannot tell you what it concluded last quarter or why.

3.5 The benefit is smallest where the expertise is highest

The fifth gap is covered at length in the fourth paper in this series, and summarised here because it belongs in this list.

Sixteen experienced developers, averaging five years on the repositories concerned, were measured across 246 randomised tasks. They forecast a 24% reduction in completion time; afterwards they estimated a 20% reduction. Measured, completion time **increased by 19%**. [5] In a separate randomised clinical trial, fifty physicians given a language model showed no significant improvement in diagnostic reasoning — while the model working alone outperformed the physicians using conventional resources. [9] The tool was capable; the combination was not.

Upstream engineering is expert work. This is the profile in which naive assistance performs worst.

4. Why the market signal is unusually noisy

Three structural features make this market harder to read than most.

The category is crowded with claims. A major analyst firm estimates that of the thousands of vendors claiming agentic capability, only around one hundred and thirty offer anything substantially agentic, and forecasts that more than 40% of agentic AI projects will be cancelled by the end of 2027 on cost, unclear value, or inadequate risk controls. [10] We note that this is an analyst forecast rather than a measurement, and that the accompanying vendor estimate derives from the firm’s own market analysis.

The supporting literature is vendor-shaped. In assembling the research for this series, searches on agent memory and retrieval failure returned vendor blogs — companies selling the remedy — ahead of peer-reviewed work in effectively every query. Two widely circulated statistics we attempted to trace turned out to be fabricated or misattributed, including one citing a benchmark that does not exist under the attribution given.

Self-reported evidence points the wrong way. This is the most practically damaging feature for a buyer running a pilot. Participants who were 19% slower believed they were 20% faster. [5] Danish workers self-reporting roughly 3% of hours saved showed no measurable effect on recorded hours in administrative data. [11] A pilot evaluated by asking participants whether it helped will return a positive answer close to regardless of what occurred.

THE HONEST READING OF THE CANCELLATION FORECAST

A projection that 40% of agentic projects will be cancelled is not evidence that the technology fails. It is consistent with a market where the gaps in section 3 are real, unmeasured at purchase, and discovered afterwards. That is a procurement problem before it is a technology problem — which is the more useful way for a buyer to hold it.

5. What closes each gap

Stated as requirements rather than as architecture, and each tied to the gap it addresses.

Gap	What has to be true instead
Performance degrades on real problems [1]	The computation is fixed, tested and cited rather than improvised per run, so behaviour does not depend on how hard the instance is
A right answer with an unreviewable derivation [2]	Every figure carries the method that produced it, its units and its source, so a qualified evaluator can follow it without rerunning it
The system will not stop [3][7]	Applicability criteria attached to each method, checked before the computation, with refusal as a legitimate output that names the failed condition
Knowledge does not survive the session [4][8]	A durable record of what was concluded and on what evidence, with superseded findings retained rather than overwritten
Benefit shrinks with expertise [5][9]	The value has to come from making verification cheap, not from asking an expert to trust an opaque suggestion

These are not exotic requirements, and none of them is a model capability. They are properties of the system built around the model — which is the argument the first paper in this series makes at length, and which the five gaps here are, in effect, the empirical case for.

6. How to run an evaluation that would find these

If the gaps are invisible in a demonstration, an evaluation has to be designed to expose them. Five tests, each targeting one gap.

- **Run it twice.** Same inputs, different days. Compare the numbers, not the impressions. This tests 3.1 directly and takes an afternoon.
- **Ask for the derivation, not the answer.** Give the output to someone qualified who was not in the room and ask them to check it. Time how long that takes. That duration is the real cost of the system.
- **Supply data that should be refused.** Too little history, a correlation outside its range, a question the data cannot settle. A system that answers anyway has failed the test that matters most.

- **Come back in a month.** Ask what it concluded and why. Ask what has been superseded since.
- **Measure elapsed time on real work, with a control.** Not satisfaction, not perceived speed, and preferably not collected by whoever sponsored the pilot.

Every one of these is cheap. None of them is standard practice.

7. Limits of this paper

Two of the studies most central to the argument are preprints rather than peer-reviewed work: the abstention benchmark [7] and the agent abstention study [3]. We rely on them because they are the only systematic measurements of knowing-when-to-stop that exist, and we grade them as preprints where they appear.

The reservoir history-matching result [1] is a single study of a single system on a single field. It is the most directly relevant evidence available and it is also one data point; a different system might degrade less. Our argument requires only that the gradient exists, not that its exact magnitude generalises.

The market forecast in section 4 is an analyst projection with an accompanying vendor count derived from the same firm's market analysis. We report it as such and do not build any part of the argument on it.

The expertise result [5] involves sixteen participants, and its authors explicitly do not claim their developers or repositories represent most software work. We have not found an equivalent study in petroleum engineering, and no such study appears to exist.

Finally, and most importantly: **this paper is published by a vendor in the category it is describing.** Every criticism in section 3 applies to us as much as to anyone, and the tests in section 6 are ones we would expect to be put to us. That is why they are written as tests a buyer runs rather than as claims a vendor makes.

8. Conclusion

The gap is not between what agents can do and what they claim. It is between what a demonstration measures and what an engineering deliverable requires.

Agents work. They plan, call tools, write code, and improve on a schedule no single vendor controls. Nothing in the published record contradicts that, and this paper does not attempt to.

What the record does show is that the properties an asset team actually depends on — the same number next quarter, a derivation an evaluator can follow, a system that stops when the method does not apply, knowledge that survives the session, and value that holds up in expert hands — are not the properties a demonstration exercises, and are close to independent of the capability a demonstration displays.

The useful response is not scepticism about the category. It is a different evaluation. Run it twice, ask for the derivation, supply something it should refuse, come back in a month, and measure the time rather than the impression. Any system worth deploying will survive that, and the five afternoons it costs are considerably cheaper than the alternative.

How SolvxAI has already resolved this

The table in section 5 is not a wish list. It is a description of decisions SolvxAI made before this paper was written, one per gap.

Performance does not degrade with realism, because the physics is never improvised: established computation runs in deterministic, cited, versioned engineering functions. The derivation is reviewable, because every figure carries its method, units and sources. The system stops, because applicability criteria are checked before the computation and a refusal is a legitimate output. The work survives the session, because conclusions live in a durable decision memory with superseded findings retained. And the expert case holds, because the output is built to be verified cheaply rather than trusted blindly.

Which is why section 6 exists. Run it twice. Ask for the derivation. Supply something it should refuse. Come back in a month. Measure the time. We wrote the tests because we expect to pass them.

ABOUT SOLVXAI

SolvxAI is the agentic operating system for upstream oil and gas. To see this working against your own asset, or to request the technical papers in this series, contact sd@nexaconsultinc.com.

References

Sources are listed with publication year; preliminary, self-reported, or vendor-authored work is flagged as such.

[1] **PetroGraph: a multi-agent framework for reservoir history matching.** arXiv:2605.15028, May 2026. **Preprint.** Benchmarks two synthetic cases and one real field. https://arxiv.org/abs/2605.15028

[2] **EngTrace: A Symbolic Benchmark for Verifiable Process Supervision of Engineering Reasoning.** arXiv:2511.01650, November 2025; revised June 2026. **Preprint.** 1,350 instances, three engineering branches, nine domains, twenty-seven models. https://arxiv.org/abs/2511.01650

[3] **Liu, X., et al.** “AgentAbstain: Do LLM Agents Know When Not to Act?” arXiv:2607.10059, 11 July 2026. **Preprint.** 263 paired tasks, 42 sandbox environments, 17 frontier models, 4 agent harnesses. https://arxiv.org/abs/2607.10059

[4] **Wu, D., et al.** “LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory.” ICLR 2025. **Peer-reviewed conference.** https://arxiv.org/abs/2410.10813

[5] **METR.** “Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity.” arXiv:2507.09089, July 2025. **Preprint**; randomised at task level; 16 developers, 246 tasks. https://arxiv.org/abs/2507.09089

[6] **FEM-Bench.** arXiv:2512.20732, December 2025; revised May 2026. **Preprint.** Graduate-level computational mechanics; five attempts per task. https://arxiv.org/abs/2512.20732

[7] **Kirichenko, P., Ibrahim, M., Chaudhuri, K., Bell, S. J.** “AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions.” arXiv:2506.09038, 10 June 2025. **Preprint**; no venue noted on the arXiv record at the time of writing. 20 models, 20 datasets. https://arxiv.org/abs/2506.09038

[8] **Maharana, A., et al.** “Evaluating Very Long-Term Conversational Memory of LLM Agents.” ACL 2024. **Peer-reviewed.** https://aclanthology.org/2024.acl-long.747/

[9] **Goh, E., et al.** “Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial.” *JAMA Network Open* 7(10):e2440969, 28 October 2024. **Peer-reviewed randomised trial**; 50 physicians; clinical vignettes. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395

[10] **Gartner.** “Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027.” Press release, 25 June 2025. **Analyst forecast, not a measurement**; the accompanying vendor estimate derives from the firm’s own market analysis. https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

[11] **Humlum, A., Vestergaard, E.** “Still Waters, Rapid Currents: Early Labor Market Transformation under Generative AI.” NBER Working Paper 33777, May 2025, revised March 2026. **Not peer-reviewed.** 25,000 workers across 7,000 workplaces, linked to national administrative registers. https://www.nber.org/system/files/working_papers/w33777/w33777.pdf