SolvxAI
Data to Decision Powered by AI

SOLVXAI WHITE PAPER SERIES | NO. 02 | APPLICABILITY

The System That Says No

Why refusal is an engineering requirement, and why the model cannot supply it

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

IN BRIEF

Every empirical method in petroleum engineering has a range in which it means something. Outside that range it does not become less precise — it becomes wrong, often while looking entirely convincing. This paper assembles the published evidence on both halves of the problem: where engineering methods stop being valid, and whether AI systems can be relied on to notice. On the first, the record is unambiguous and has been for decades. On the second, it is worse than most buyers assume — the ability to abstain is close to independent of the ability to answer, and the training that makes models better reasoners measurably makes them worse at declining. The conclusion is that refusal has to be built into the system around the model, because it is demonstrably not a property the model supplies.

Contents

1. Executive summary

The most valuable thing an engineering system can tell you is that it cannot answer your question.

This is not a rhetorical flourish. In petroleum engineering, the standard failure is not a method that returns an obviously silly number. It is a method applied slightly outside the conditions it was derived for, which returns a plausible number, with a convincing fit, that is materially wrong. The published record contains clean demonstrations of this, and they are not new.

Three findings frame the argument:

- In a controlled simulation with perfect data, a gas material-balance plot produced a fit of **$R^2 = 0.9998$** — visually a near-perfect straight line — while the resulting estimate of gas in place was **8.2% too high**. The same study found the oil case overestimated original oil in place by **+160%, +90% and +50%** when evaluated at 3%, 7% and 20% depletion respectively. [1]
- On a benchmark built specifically to measure knowing-when-not-to-answer, the highest-accuracy model scored **0.81 accuracy but only 0.46 abstention recall** — near the bottom of the field. The two abilities are close to decoupled, and the authors report that **reasoning fine-tuning degrades abstention by 24% on average**. [5]
- Across seventeen frontier models in four agent configurations, the best achieved only **59.5%** on paired tasks requiring the system to act when it should act and abstain when it should abstain — with a documented failure mode in which agents **execute irreversible actions before** recognising they should have stopped. [10]

The engineering discipline settled its half of this question long ago: reserves standards and securities regulation both require that a method be demonstrated applicable to the formation in front of you, not assumed. What is new is the arrival of systems that will answer anything, deployed against problems where the correct output is frequently a refusal.

> **THE STANDARD ALREADY SAYS THIS**
>
> The Petroleum Resources Management System is explicit that the obligation runs to communicating limits, not just producing numbers: “Regardless of the analytical procedures used, the goal is to communicate the range of uncertainty in the recoverable resources. An underlying principle is that the reliability of the estimates depends on the quantity and quality of the source data.” [3]

2. The failure mode that matters

Engineering software has always been able to produce a wrong answer. What distinguishes the failures that reach a reserves report from the ones caught in review is that the dangerous ones look right.

A method applied outside its range does not usually announce itself. It produces a number in the expected order of magnitude, on a plot that looks like the plots in the textbook, with a correlation coefficient that would pass any casual review. The error is not in the arithmetic. It is in the premise that the method applied at all.

> **A method outside its range does not become less precise. It becomes wrong, while continuing to look exactly like a method that worked.**

This is why “how accurate is it?” is the wrong first question to ask of an engineering system, and why accuracy benchmarks — ours or anyone else’s — do not settle the matter. A system that is right 90% of the time and silent about which 10% is considerably more dangerous than one that is right 80% of the time and tells you when it is guessing.

3. Where engineering methods stop being valid

The petroleum literature is candid about this. Four categories of limit recur, and all four are documented in standards or peer-reviewed work.

3.1 The answer is not unique

Well test interpretation is, formally, an inverse problem. Gringarten states the position plainly: “This is an inverse problem without a unique solution.” He notes that although the catalogue of model components is small, “their combination can yield several thousand different interpretation models to match all observed well behavior,” and that “nonuniqueness decreases as the amount of information increases.” [2]

The practical consequence is that where several models match a pressure response equally well, naming one of them as the answer is a choice being presented as a finding. A system that silently selects the best-fitting model and reports it as the interpretation has concealed the most important thing the analysis discovered.

3.2 It is too early in the life of the reservoir

Pletcher’s 2002 study in *SPE Reservoir Evaluation & Engineering* is the sharpest published demonstration of a method that fits beautifully and is nonetheless wrong. Working with simulated data — so the true answer was known exactly — he showed that a gas p/z plot under a weak water drive appears linear and invites confident extrapolation:

“The plotted p/z points in Fig. 2 appear to lie on an almost perfect straight line (R^2 = 0.9998 after 10 years), giving the impression that an extrapolation to OGIP could be made with confidence. However, an extrapolation of the points made after 2 years, when 11% of the true OGIP had been produced, would yield a value for OGIP of 109 Bcf, or 8.2% too high. After 5 years, the error would be +6.5%. Even after 10 years and recovery of 54% of the OGIP, the error would still be +4.0%.” [1]

The oil case is starker still. Evaluating at three points in the reservoir's life produced overestimates of original oil in place of 160%, 90% and 50%.

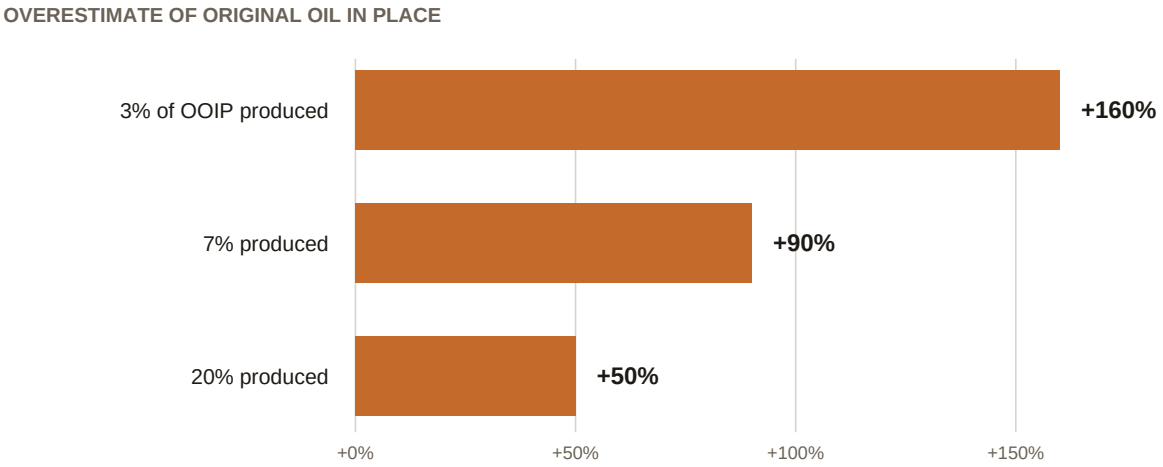

Figure 1. Overestimate of original oil in place from material balance under a weak water drive, evaluated at three stages of depletion in a controlled simulation where the true value was known. Errors of this magnitude arise from data that fits the model well. Source: Pletcher, SPE Reservoir Evaluation & Engineering, 2002 [1].

Two cautions on reading this figure, both stated by the author. It is a single simulated case, not an industry-wide threshold — and Pletcher's own thesis is that this ambiguity is *diagnosable* with the right diagnostic plot, not that material balance is unreliable. That is precisely the point. The information needed to know whether the method applies exists. It simply has to be checked, every time, rather than assumed.

3.3 The correlation was fitted on different rock and different fluid

Empirical correlations carry their provenance whether or not anyone looks it up. Standing's correlation was developed from 105 experimentally determined data points on 22 oil-gas mixtures from California. Vasquez and Beggs drew on over 600 laboratory PVT analyses. Al-Marhoun used 75 bottomhole samples from 62 reservoirs in the Middle East, and its author claims validity only "for all types of gas-oil mixtures that share similar properties as those used in the derivation." [12]

Some correlations document their own failure mode explicitly. The published range for Beggs and Robinson dead-oil viscosity is 16–58 °API and 70–295 °F, with the note that it "tends to overstate the viscosity of the crude oil when dealing in temperature ranges below 100 °F to 150 °F" — a caution attached to the correlation itself, which survives only if the system carrying it preserves the caution alongside the coefficients. [12]

A number produced by a correlation applied outside its fitted envelope is not a low-confidence number. It is an unqualified one, and it propagates cleanly into every calculation downstream.

3.4 The standards already require the limits to be stated

None of this is a novel observation, and the obligation is not merely professional good manners. The Petroleum Resources Management System, in the edition currently governing, is direct about when material balance can and cannot be relied upon:

“In ideal situations, such as depletion-drive gas reservoirs in homogeneous, high-permeability reservoir rocks and where sufficient and high-quality pressure data are available, estimation based on material balance may provide very reliable estimates… In complex situations, such as those involving water influx, compartmentalization, multiphase behavior, and multilayered or low-permeability reservoirs, shales or CBM, material balance estimates alone may provide erroneous results.” [3]

It is equally direct that extrapolation is not permitted on assumption: “Extrapolation of reservoir presence or productivity beyond a control point within a resources accumulation **must not** be assumed unless there is technical evidence to support it.” [3]

United States securities regulation goes further and makes it a condition of reliance. Reliable technology is defined as “a grouping of one or more technologies (including computational methods) that has been field tested and has been demonstrated to provide reasonably certain results with consistency and repeatability *in the formation being evaluated or in an analogous formation*.” [4]

> **WHAT THAT SENTENCE ACTUALLY DEMANDS**
>
> Applicability is not a quality standard a vendor may aspire to. Under Regulation S-X, a computational method may be relied upon only where it has been demonstrated to work *in the formation being evaluated or an analogous one*. A system that cannot say which formations a method was validated against cannot support that determination — and neither can the evaluator relying on it. [4]

PRMS states these thresholds qualitatively rather than numerically. There is no published rule of the form “material balance requires N% depletion,” and any vendor quoting one should be asked for the source. What the standards require is that the evaluator establish applicability for the case in hand and communicate the resulting uncertainty — which is a harder obligation than a threshold would be, not an easier one.

4. Whether AI systems notice

So the engineering half of the problem is well characterised and has been for twenty years. The question is whether a capable general-purpose model, pointed at the same work, recognises the boundary and stops.

The evidence assembled below is recent, largely 2025–2026, and it is not encouraging. We have deliberately included the findings that cut against this paper’s argument, in section 4.5.

4.1 Models are overconfident, but not uniformly

A preregistered study across diverse tasks found that “the current crop of LLMs are, like people, too sure they are right: confidence exceeds accuracy, on average.” The important qualifier follows immediately: “this tendency is moderated by a powerful hard-easy effect, wherein overconfidence is greatest on difficult tests; by contrast, easy tests actually show substantial underconfidence.” [6]

For engineering use this is the worst possible distribution. Confidence is best calibrated where the work is routine and degrades exactly where the problem is hard — which is where an engineer would most want a warning.

A peer-reviewed evaluation across nine models and three factual datasets describes the operational risk in the same terms: “their frequent overconfidence-misalignment between predicted confidence and true correctness-poses significant risks in critical decision-making applications.” It also reports a counter-intuitive wrinkle — that large RLHF-tuned models “can paradoxically suffer increased miscalibration on easier queries.” [7]

A further peer-reviewed result identifies an ownership bias: models assign up to 26% higher confidence to their own responses than to identical content attributed to someone else. [8] A system that grades its own work is, measurably, a generous marker.

4.2 Knowing when not to answer is close to unrelated to answering well

AbstentionBench evaluates twenty frontier models across twenty datasets built from questions that are unanswerable, underspecified, or resting on false premises. Its framing matches the engineering requirement almost exactly: “knowing when not to answer is equally critical as answering correctly.” [5]

The headline result is that “abstention is an unsolved problem, and one where scaling models is of little use.” But the finding that matters most for anyone evaluating engineering software is visible in the paper’s own results table, plotted below: answer accuracy and abstention recall barely track each other.

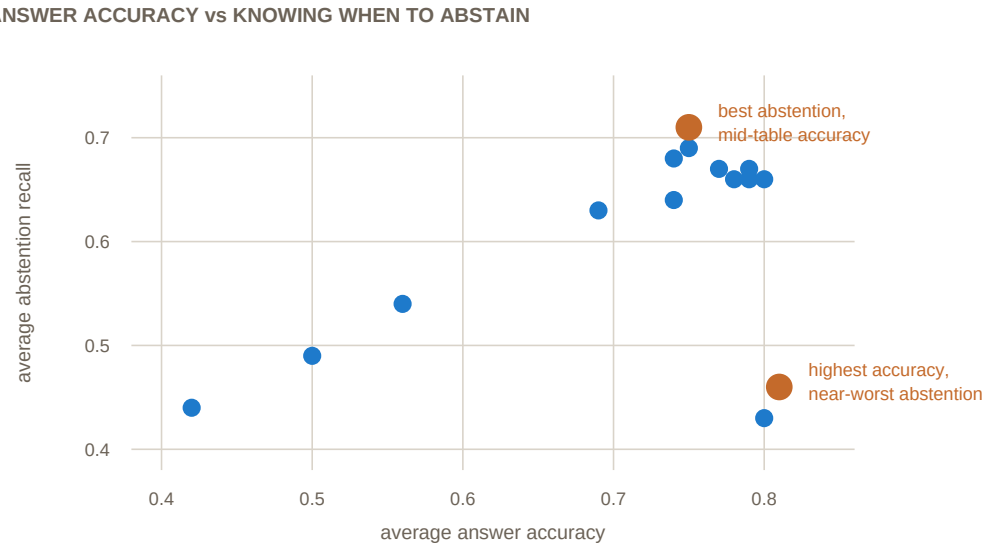

Figure 2. Each point is one of fifteen models: average answer accuracy against average abstention recall. If answering well implied knowing when not to answer, the points would form a rising line. They do not. Source: Kirichenko et al., AbstentionBench, 2025 [5]; preprint — see §6 on grading.

The model at the top of the accuracy column scores 0.81 on accuracy and 0.46 on abstention — third from last on the latter. The best abstainer, at 0.71, sits mid-table on accuracy. Selecting a system on benchmark accuracy tells you close to nothing about whether it will stop when it should. [5]

4.3 The training that improves reasoning makes abstention worse

The most uncomfortable result in the same study concerns reasoning models specifically — the class of system most often proposed for technical work. The authors report, with evident surprise, that “reasoning fine-tuning degrades abstention (by 24% on average), even for math and science domains on which reasoning models are explicitly trained.” [5]

They add that a carefully written instruction can raise abstention in practice but “does not resolve models’ fundamental inability to reason about uncertainty.” [5] Instruction is a mitigation, not a control.

> **Reasoning fine-tuning degrades abstention by 24% on average — even in the domains those models are explicitly trained on.**
>
> Kirichenko et al., AbstentionBench, 2025 [5]

4.4 Agents act before they realise they should not have

Where a model merely answers, an agent does things. A July 2026 study built 263 paired tasks across 42 executable environments, each pair consisting of a task the agent should perform and a variant it should decline, and ran them across seventeen frontier models in four agent configurations. [10]

The best configuration achieved 59.5% paired accuracy — correct on both halves of the pair. The authors' conclusion echoes AbstentionBench: "abstention capability is largely independent of general task-solving capability, indicating that scaling task-solving alone will not close this gap." [10]

They also name the failure mode that should concern anyone contemplating autonomous engineering work: **post-hoc abstention**, "in which agents execute irreversible actions before recognizing abstention triggers." [10] The agent works out that it should have stopped, after it has finished.

4.5 The evidence that cuts the other way

Two findings complicate the picture, and we would rather present them than have a reader find them.

First, calibration is not simply absent. Work presented at EMNLP 2025 reports that "Large Language Models (LLMs) have demonstrated inherent calibration capabilities, where predicted probabilities align well with correctness, despite prior findings that deep neural networks are often overconfident," and identifies "a distinct confidence correction phase in the upper/later layers, where model confidence is actively recalibrated after decision certainty has been reached." [9] Something resembling self-assessment is happening inside these systems.

Second, calibration can be improved substantially without retraining. One study reports error reductions of up to 90% from restructuring the question [7]; another reports a 26% calibration improvement from correcting the ownership bias [8].

The honest reading is that model calibration is a real, partially understood, improving capability — and that none of that changes the engineering conclusion. Improving on a base of 59–65% correct abstention [11] is welcome and is not a foundation for a reserves estimate. The gap between a research trend and an auditable control is the subject of the rest of this paper.

5. Refusal is a property of the system, not the model

Put the two halves together. Engineering methods have documented boundaries that are checkable from the data in front of you. Models are measurably unreliable at noticing boundaries, get worse at it as they are trained to reason harder, and when wired into agents will sometimes act first. Therefore the boundary check cannot live inside the model.

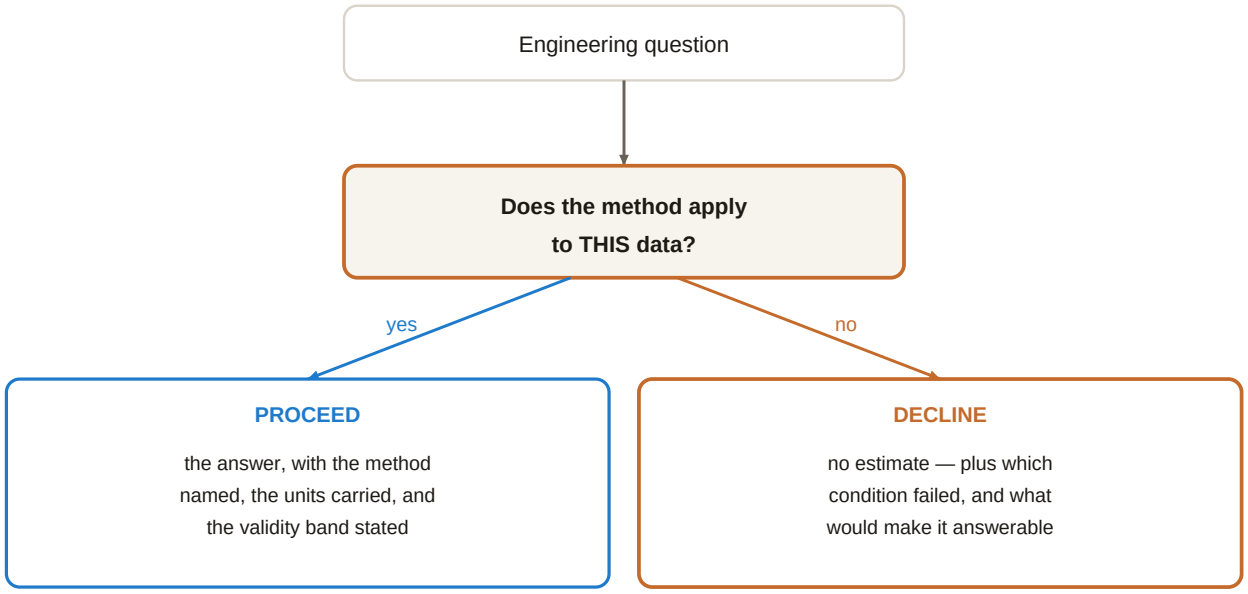

Figure 3. The applicability gate. The check runs before the computation, on the data actually supplied, and it has two legitimate outcomes. A refusal that names the failed condition is a usable engineering result; a silent best-effort answer is not.

What follows is what that requires in practice, described in terms of behaviour rather than construction.

5.1 The criteria are attached to the method, not to the prompt

Each analysis carries its own conditions of validity — the data it needs, the regimes it holds in, the fluid and pressure envelope it was fitted on. Those conditions are properties of the engineering method and are written down once, alongside it. They are not inferred at the moment of asking, and they do not vary with how the question was phrased.

5.2 The check runs before the computation, not after

An applicability test that runs after a number exists is a review step, and review steps get overridden. The check has to be able to prevent the calculation from being attempted — which is the difference between a system that declines and a system that apologises.

5.3 A refusal is a deliverable with content

“I cannot answer that” is not useful. A refusal that earns its place states which condition failed, what the data would have to look like for the analysis to become valid, and what can be said in the meantime — a bounded range, the competing interpretations, or the next measurement worth taking. That output is frequently more valuable than the estimate would have been, because it is actionable.

5.4 Ambiguity is surfaced, not resolved by preference

Where several models match the data, the system's obligation is to present them as competing interpretations with what distinguishes them, not to rank them and report the winner. Gringarten's several thousand candidate models [2] do not collapse into one because software chose one.

5.5 The refusals are enumerable in advance

A buyer should be able to ask what a system will decline to do and receive a list — not a philosophy. If the answer is a description of the system's general caution rather than named analyses and named triggering conditions, there is no gate, only a disposition.

> **THE TEST WE WOULD APPLY TO OURSELVES**
>
> Ask any vendor, including us, to name three analyses their system refuses to perform and the exact conditions that trigger each refusal. Then supply data that should trigger one, and watch. A gate that cannot be demonstrated on request is not a gate.

6. A note on our own sources

This paper argues that systems should state the limits of their evidence, so it should state the limits of its own.

The petroleum engineering evidence in section 3 is old. The strongest quantitative source [1] dates from 2002 and the governing standard [3] from 2018. That is partly because the underlying physics has not changed, and partly an artifact of access: a large portion of the recent SPE literature on interpretation limits sits behind a paywall we did not cross for this review. Readers with institutional access will find more, and more recent, material than is cited here.

The correlation provenance in section 3.3 is drawn from commercial software reference documentation [12] rather than the original papers. The underlying facts are uncontroversial and long established, but the citation is weaker than we would prefer and is graded accordingly.

On the AI side, AbstentionBench [5] and AgentAbstain [10] are preprints, not peer-reviewed publications, as is the preregistered calibration study [6]. The calibration results at [7], [8] and [9] are peer-reviewed. The frontier abstention rates quoted in section 4.5 come from a model developer's own safety report [11] — a vendor self-report, and treated as one. That same report notes that measured abstention differed significantly depending on whether the model appeared to recognise it was being evaluated, which is a caveat on every self-reported abstention figure including that one.

None of these caveats overturn the argument, and we would rather you weigh them yourself than discover them later.

7. Questions worth asking

These are diagnostic. They separate a system with applicability control from one with a cautious tone, and they can be asked of us with the same force as of anyone else.

- **1.** Name three analyses the system refuses to perform, and the exact conditions that trigger each refusal.
- **2.** Show me a refusal. Supply data that should trigger one and let me watch it happen.
- **3.** When a method is applied, what validity range does the output carry, and where did that range come from?
- **4.** When several interpretations fit the data equally well, what does the system return?
- **5.** Does the applicability check run before the computation or after it? Can it stop the work?
- **6.** For a correlation, can the system tell me the fluid and pressure envelope it was fitted on, and whether my case sits inside it?
- **7.** What does the system do when the data is sufficient for a bounded range but not a point estimate?
- **8.** If an autonomous run encounters a condition it should decline, does it stop before acting or after?
- **9.** How would an auditor establish that a method was applicable to this formation, using only what the system produced?
- **10.** What has the system refused to do for your other customers in the last month?

8. Conclusion

A system that will answer anything is not a more capable system. It is an uncalibrated one.

The petroleum engineering profession worked out decades ago that method applicability is not a detail. It is written into the resources standard, and into the securities regulation that governs what may be reported to a market. The requirement is that reliance be demonstrated for the formation in front of you — not that a method be generally respectable.

The published record on AI systems is that they are not reliable at supplying this. The ability to abstain is close to independent of the ability to answer, it does not improve with scale, it degrades with the reasoning training that improves everything else, and in agent settings the recognition can arrive after the action. These are measured findings, not speculation, and several come from the model developers themselves.

The engineering response is not to wait for models to improve, though they will. It is to put the boundary check where it can be specified, tested, audited, and shown to a buyer on demand — outside the model, attached to the method, running before the computation, and returning a refusal with enough content to act on.

The useful question to ask any vendor is not how often the system is right. It is what the system will not do, and whether they can show you.

How SolvxAI has already resolved this

The applicability gate in Figure 3 is not an aspiration. It is built, and it works the way section 5 says it must.

Applicability criteria are attached to the engineering functions themselves — the data each method needs, the regimes it holds in, the envelope it was fitted on — written down once with the method, not inferred from the phrasing of a question. The check runs before the computation and can stop it. A refusal comes back as a deliverable: which condition failed, what the data would have to look like for the analysis to become valid, and what can honestly be said in the meantime. Where several interpretations fit the data equally well, they are returned as competing interpretations — not silently ranked.

Question 2 in section 7 is the one we most want to be asked: supply data that should trigger a refusal, and watch what comes back.

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] **Pletcher, J. L.** “Improvements to Reservoir Material-Balance Methods.” *SPE Reservoir Evaluation & Engineering*, February 2002, pp. 49–59 (SPE 75354). Peer-reviewed. Simulated cases with known true values.

[2] **Gringarten, A. C.** “Well Test Analysis in Practice.” *The Way Ahead*, Society of Petroleum Engineers, 14 June 2012. SPE professional magazine; not peer-reviewed. https://jpt.spe.org/twa/well-test-analysis-practice

[3] **SPE / WPC / AAPG / SPEE / SEG.** *Petroleum Resources Management System*, revised June 2018 (v. 1.03), sections 2.4, 4.1 and 4.2. Multi-society industry standard; voluntary. https://www.spe.org/media/filer_public/0c/83/0c835db9-501f-4ce7-97f1-a1d6bb4e3331/prmgmtsystem_v103.pdf

[4] **US Securities and Exchange Commission.** Regulation S-X, Rule 4-10, 17 CFR §210.4-10(a)(24) and (a)(25), defining reasonable certainty and reliable technology. Binding regulation. https://www.law.cornell.edu/cfr/text/17/210.4-10

[5] **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 time of writing. 20 models, 20 datasets. https://arxiv.org/abs/2506.09038

[6] **Michael, N., BenShushan, D., Bien, J., Moore, D. A.** “Confidence Calibration in Large Language Models.” arXiv:2605.23909, 3 April 2026. Preprint; preregistered study. https://arxiv.org/abs/2605.23909

[7] **Chhikara, P.** “Mind the Confidence Gap: Overconfidence, Calibration, and Distractor Effects in Large Language Models.” *Transactions on Machine Learning Research*, 2025 (arXiv:2502.11028). Peer-reviewed. https://arxiv.org/abs/2502.11028

[8] **Sanz-Guerrero, M., Mager, M., von der Wense, K.** “Large Language Models Are Overconfident in Their Own Responses.” arXiv:2606.03437, 2 June 2026. Accepted to ACL 2026 Findings; peer-reviewed. https://arxiv.org/abs/2606.03437

[9] **Joshi, A., Ahmad, A., Modi, A.** “Calibration Across Layers: Understanding Calibration Evolution in LLMs.” arXiv:2511.00280, 31 October 2025. Accepted at EMNLP 2025 (main); peer-reviewed. Cited here as evidence against this paper’s framing. https://arxiv.org/abs/2511.00280

[10] **Liu, X., Zhang, Y. E., Kasprova, V., Rabbani, P., Zahraei, P. S., Zhang, T., Ebrahimpour-Boroojeny, A., Chandrasekaran, V.** “AgentAbstain: Do LLM Agents Know When Not to Act?” arXiv:2607.10059, 11 July 2026. Preprint. 263 paired tasks, 42 environments, 17 models. https://arxiv.org/abs/2607.10059

[11] **Model developer safety report.** Correct-abstention rates measured on AbstentionBench for current frontier models, reported at 59.1% and 64.1–65.0%. arXiv:2606.12429, May 2026. **Vendor self-report** — the reporting party is the model developer, and the same report notes measured abstention varied significantly with apparent evaluation awareness. https://arxiv.org/abs/2606.12429

[12] **Correlation provenance and fitted ranges** for Standing, Vasquez and Beggs, Beggs and Robinson, and Al-Marhoun, as published in commercial reservoir-engineering software reference documentation. **Vendor-authored documentation**, not the original papers; the underlying ranges are long established. https://www.ihsenergy.ca/support/documentation_ca/Harmony/content/html_files/reference_material/calculations_and_correlations/oil_correlations.htm