

Bibliography: AI in Reserves Work | Research Sources

[Bibliography: AI in Reserves Work | Research Sources](#)

[SEC REGULATORY STANDARDS & GUIDANCE](#)

[SPE / SPEE & PROFESSIONAL STANDARDS](#)

[LLM LIMITATIONS & AI RELIABILITY](#)

[PHILOSOPHICAL FOUNDATIONS: SYMBOL GROUNDING & TACIT KNOWLEDGE](#)

[PETROLEUM ENGINEERING & RESERVES DOMAIN](#)

[CRUDECODE MCP APPLICATION \(SOURCE CODE ANALYSIS\)](#)

SEC REGULATORY STANDARDS & GUIDANCE

SEC Regulation S-K, Item 1202 — Disclosure of Reserves (17 CFR 229.1202)

<https://www.ecfr.gov/current/title-17/chapter-II/part-229/subpart-229.1200/section-229.1202>

(Text verified against current eCFR, September 4, 2026. Title 17 last amended August 17, 2026.)

How it's referenced: Item 1202 governs what a registrant must disclose about its oil and gas reserves in SEC filings. Four provisions carry the accountability argument, and the subsection numbering matters:

- **Item 1202(a)(7)**

Preparation of reserves estimates or reserves audit. Requires the registrant to disclose and describe the internal controls used in its reserves estimation effort, **and** to disclose the qualifications of the technical person primarily responsible for overseeing the preparation of the reserves estimates. Where a third party conducted a reserves audit, the qualifications of the technical person overseeing that audit must also be disclosed. This is a controls-and-competence requirement, not a numbers requirement.

- **Item 1202(a)(8)(x)**

When a third party prepares or audits the estimates, that party's report must be filed as an exhibit, and the report **must carry the signature of the third party**. This is the provision the presentation's title rests on. The word in the regulation is *signature*.

- **Item 1202(a)(8)(iv)**

The third-party report must state the assumptions, data, methods, and procedures used, and must include a statement that they are appropriate for the purpose served by the report.

- **Item 1202(a)(6)**

Requires a general discussion of the technologies used to establish the appropriate level of certainty for reserves estimates from material properties. **Note the trigger:** this obligation attaches only where the registrant has not previously disclosed reserves estimates in a Commission filing, or is disclosing material additions to its reserves estimates. Directly relevant to AI adoption: a tool that materially adds to booked reserves puts the technology itself on the disclosure page.

Note that probable and possible reserves are permitted but not required under Item 1202 (Instruction 2 to paragraph (a)(2)); only proved reserves are mandatory.

A reserves process that cannot be described as an auditable internal control, or whose methods cannot be attested as appropriate by a named preparer, does not satisfy these provisions.

Why read it: This is the regulatory floor and the precise text where the accountability chain lives. Item 1202(a)(7)'s internal-controls language is the most direct point of contact with the reproducibility argument: a nondeterministic, unauditable process is not an internal control an engineer can describe and stand behind. Reading the actual subsection language rather than a paraphrase is what lets a presenter speak to it accurately in front of an audience that files under it.

SEC Regulation S-K, Item 1203 — Proved Undeveloped Reserves (17 CFR 229.1203)

<https://www.ecfr.gov/current/title-17/chapter-II/part-229/subpart-229.1200/section-229.1203>

(Text verified against current eCFR, September 4, 2026.)

How it's referenced: Item 1203 is four short paragraphs, all specific to proved undeveloped reserves:

- (a) Disclose the total quantity of PUDs at year end.
- (b) Disclose material changes in PUDs during the year, including PUDs converted to proved developed.

- (c) Discuss investments and progress made during the year toward conversion, including capital expenditures.
- (d) Explain why material amounts of PUDs in individual fields or countries remain undeveloped five years or more after first being disclosed as PUDs.

PUDs are the category where forecast judgment has the most room and the least evidence. An undrilled location has no production history at all — level, shape, and timing come entirely from analog selection and professional judgment. That is precisely the work CrudeCode hands to the model: its `deal_forecast_wells` spec has Claude assert the first-production month for DUC and permit wells and source the analog cohort itself.

Paragraph (d) is the accountability ratchet. It requires a narrative explanation, years later, of why something booked as proved is still in the ground — a reasoned account that has to survive across reporting periods and connect back to the judgment originally made. That is a demand for persistent, attributable reasoning, not a number.

Why read it: It is the provision where forecast judgment is thinnest and the disclosure burden is heaviest, and where a five-year-old estimate has to be re-explained by whoever owns it now.

SEC Regulation S-X, Rule 4-10 — Financial Accounting and Reporting for Oil and Gas Producing Activities (17 CFR 210.4-10)

<https://www.ecfr.gov/current/title-17/chapter-II/part-210/section-210.4-10> Related SEC staff guidance: Staff Accounting Bulletin Topic 12.

<https://www.sec.gov/interps/account/sabcodet12.htm>

How it's referenced: Rule 4-10 is the foundational definitions-and-reporting rule for oil and gas activities. It sets out the SEC definitions of proved, probable, and possible reserves and the conditions an estimate must meet to qualify — including the "reasonable certainty" standard for proved reserves and the requirement that estimates rest on data considered definitive. Critically, it defines reserves as the **economically producible** quantities under defined price and cost conditions. Reserve reports filed with the SEC state their conformance to this rule (the Ovintiv/McDaniel audit opinion certifies conformance to "SEC's Regulation S-X Part 210.4-10(a)"). The rule is the technical yardstick against which any estimate — human or machine-assisted — is measured.

Why read it: It defines the terms of the debate. "Proved reserves" is not a loose descriptor; it is a defined category carrying a reasonable-certainty threshold *and an economic-productibility test*. Understanding Rule 4-10's definitions is what separates a defensible estimate from a plausible-looking one, and it establishes that the standard of proof is fixed by regulation, not by whatever confidence a tool happens to report.

The economic-productibility requirement is what makes the CrudeCode economics finding decisive rather than incidental: a run-out volume computed with zero operating cost and no economic-limit test is not a reserve under this rule, regardless of how the decline parameters were chosen. See Section 6.

SPE / SPEE & PROFESSIONAL STANDARDS

Petroleum Resources Management System (PRMS), 2018 Revision

Society of Petroleum Engineers, sponsored by SPE, WPC, AAPG, SPEE, SEG, SPWLA, and EAGE; prepared by the SPE Oil and Gas Reserves Committee. Available via OnePetro and SPE.org.

How it's referenced: PRMS is the internationally recognized framework for classifying and categorizing petroleum reserves and resources. It sets out the definitions (proved, probable, possible; reserves vs. contingent and prospective resources) and the principles an evaluation must follow, including that classification rests on the judgment of a qualified evaluator applying consistent, documented criteria. The presentation's use of PRMS rests on this evaluator-centered structure: the framework assumes a responsible professional exercising and recording judgment, not an opaque process producing a number. An estimate whose reasoning cannot be reconstructed or attributed to a qualified evaluator sits outside the way PRMS is designed to be applied.

Why read it: PRMS is the common language reserves estimates are classified in and audited against worldwide. Reading it establishes what the framework assumes about the evaluator's role and judgment.

Cheng, Y., W. J. Lee, and D. A. McVay (2008). "Improving Reserves Estimates From Decline-Curve Analysis of Tight and Multilayer Gas Wells"

SPE Reservoir Evaluation & Engineering, 11(5), 912–920. SPE-108176-PA.

<https://doi.org/10.2118/108176-PA>

Authors (verified against the OnePetro record): Y. Cheng (West Virginia University); W. J. Lee (Texas A&M University); D. A. McVay (Texas A&M University). Received February 20, 2007; revision received April 10, 2008; accepted April 19, 2008; published online October 1, 2008.

How it's referenced: The paper documents why decline-curve analysis produces large, analyst-dependent uncertainty in tight and multilayer gas wells. Such wells frequently exhibit an Arps b-value greater than unity, and analysts handle this differently — some use the matched b, some force b to 1, some switch to exponential decline at a chosen limit — so forecasted

performance differs significantly depending on who analyzes the data. The authors trace the cause to physics: long-duration transient effects in stimulated tight wells and coexisting flow regimes in commingled layers produce abnormal decline behavior ($b > 1$) that a naive fit misreads, leading to forecast errors and affecting reserves classification.

Why read it: It grounds the b-factor and transient-flow argument in the SPE literature rather than in assertion. The failure mode is not AI-specific in origin; it is a documented hazard of decline-curve analysis that requires physical judgment — recognizing transient behavior, identifying when boundary-dominated flow has been reached — to navigate. Any automated routine that fits a curve without that physical reasoning inherits the hazard.

Collins, P. W., D. Ilk, and T. A. Blasingame (2014). "Practical Considerations for Forecasting Production Data in Unconventional Reservoirs — Variable Pressure Drop Case"

SPE Annual Technical Conference and Exhibition, Amsterdam, 27–29 October 2014.
SPE-170945-MS. Authors: P. W. Collins (Texas A&M University / DeGolyer and MacNaughton), D. Ilk (DeGolyer and MacNaughton), T. A. Blasingame (Texas A&M University).

How it's referenced: The paper addresses a core limitation of traditional time-rate decline-curve analysis: it assumes a constant flowing bottomhole pressure, an assumption unconventional wells routinely violate through managed drawdown, choke changes, and restricted-rate production. Applying the assumption where it does not hold yields misleading EUR values. Two findings bear directly on the presentation's argument. First, the authors note that applying Arps' hyperbolic equation to unconventional reservoirs produces b-values greater than 1, which makes the decline model unbounded and creates the potential to significantly overestimate reserves; the peer-reviewed statement of the transient-flow and b-factor hazard. Second, the work depends throughout on diagnostic judgment (loss-ratio and loss-ratio-derivative plots, pseudopressure transformations, flow-regime identification) that the authors emphasize must be applied with rigorous, well-specific analysis of completion and production data, not by rote fitting.

Why read it: It grounds two of the presentation's claims in authoritative SPE literature. The $b > 1$ overestimation risk is stated plainly by leading practitioners, and the paper's diagnostic workflow illustrates the physical reasoning a defensible forecast requires. Most striking, the authors close by asking whether the rate-decline relations carry any physical meaning or are "statistical in nature" **a question raised from inside petroleum engineering that parallels the symbol-grounding argument: a curve can match production data without any physical model of the reservoir beneath it.** When a leading DCA authority openly asks whether these

models are physics or curve-fitting, **it underscores why the judgment that grounds a forecast cannot be delegated to a system that only pattern-matches.**

SPEE Standards and Code of Ethics

Society of Petroleum Evaluation Engineers. <https://spee.org/>

How it's referenced: SPEE's professional standards and code of ethics govern the conduct of reserves evaluators: competence, independence, and personal responsibility for the evaluations they sign. The presentation uses this as the ethical counterpart to the SEC's legal requirements where SEC rules require a named responsible person, SPEE's code establishes what that person owes: honest, competent, independent judgment they can explain and defend. An automated process cannot hold or discharge an ethical obligation; only the evaluator who signs can.

Why read it: It is the professional-responsibility foundation beneath the legal one. The rules say a person must be named; the code says what that naming means in terms of duty and accountability.

LLM LIMITATIONS & AI RELIABILITY

Kostikova, Aida, Zhipin Wang, Deidamea Bajri, Ole Pütz, Benjamin Paaßen, and Steffen Eger (2025). "LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models"

arXiv:2505.19240. <https://arxiv.org/abs/2505.19240>

How it's referenced: A meta-survey of the limitations literature rather than a benchmark of models. From a corpus of roughly 250,000 papers, the authors identify 14,648 that study LLM limitations and map which limitations the research community focuses on most. Reasoning is the single most-studied limitation, followed by generalization, hallucination, bias, and security. It establishes scale and consensus: concern about LLM limitations is a large, fast-growing field, and the capability reserves work depends on most – reasoning and judgment – is the one the research community most often flags as unreliable.

Why read it: It demonstrates that skepticism about LLM limitations is not a fringe or contrarian position but a peer-reviewed field of thousands of papers, with reasoning at the top of the list. It provides the breadth and legitimacy behind the argument; the targeted papers that follow supply the specific failure modes.

Mirzadeh, Iman, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar (2025). "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models"

International Conference on Learning Representations (ICLR), 2025. arXiv:2410.05229.
<https://arxiv.org/abs/2410.05229>

How it's referenced: Apple's study shows that LLM performance on math problems degrades when only the numerical values are changed, and that variance increases as problem complexity rises. Adding a clause that appears relevant but is not causes performance to drop across all state-of-the-art models. The authors conclude that LLMs rely on pattern-matching rather than genuine logical reasoning.

In reserves terms: every well is the same decline-curve structure with different numbers, and every production history contains anomalies (workovers, shut-ins) that are irrelevant to the underlying decline. A system whose quantitative reasoning is fragile to numerical variation and derails on irrelevant clauses describes the exact failure mode that occurs when it processes real production data.

Why read it: This is the empirical counterpart to the philosophical symbol-grounding argument. Harnad explains why an LLM might manipulate numbers without understanding them; GSM-Symbolic measures it happening on quantitative reasoning tasks.

* *Note on the rebuttal:* This paper has been challenged. Ivanova (2024) argues the statistical evidence for performance differences is weaker than the paper implies—significant at the 5% level but not the 1%, and that some variation across problem instances is expected rather than damning. The rebuttal is available alongside the paper: <https://desirivanova.com/post/gsm-symbolic/>. The conservative claim survives the critique: performance degrades and variance increases as complexity rises.

He, Horace, and Thinking Machines Lab (2025). "Defeating Nondeterminism in LLM Inference"

Thinking Machines Lab: Connectionism, September 2025.
<https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/>

How it's referenced: Documents that LLM inference is nondeterministic; the same input can produce different outputs across runs, and traces the cause to how inference is computed (batch-dependent floating-point behavior), not merely to a sampling temperature that can be set to zero. This is decisive for reserves estimation, where reproducibility is a filing requirement: the

same production data must yield the same reserve estimate today, next quarter, and under audit.

Why read it: It closes the obvious objection that fixing a random seed or averaging repeated runs would solve the problem. The nondeterminism is architectural; a property of how the computation runs, not a bug awaiting a patch. An LLM-based reserves tool cannot guarantee that identical inputs produce identical EURs, so its output fails reproducibility before any question of accuracy is reached. The source is a frontier AI lab with every incentive to present LLMs favorably.

Wang, Zirui, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen (2024). "CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs"

arXiv:2406.18521. Princeton University. <https://charxiv.github.io/>

How it's referenced: CharXiv evaluates how well multimodal LLMs read real charts and finds systematic failures, with models scoring far below humans on questions that require genuinely reading the figure rather than guessing from patterns. This is the kind of visual evidence reserves engineers work from: production-history plots, log-rate-vs-time diagnostics, offset wells mapped by location. A model that cannot reliably read a chart cannot reliably identify a boundary-dominated-flow signature or an anchor point from a rate-time plot.

Why read it: It converts an abstract concern into a measured failure. An AI system instructed to read a production history and identify where the decline begins is performing precisely the chart-reading task current models are shown to fail.

Direct tie to the source code: CrudeCode's own well-forecasting skill concedes that much of the flow-regime judgment is visual. That concession plus this benchmark is the cleanest pairing in the deck; the tool asks the model to do the exact task the literature measures it failing.

Steinacker-Olsztyn, Nicolas, Devashish Gosain, and Ha Dao (2025). "Is Misinformation More Open? A Study of robots.txt Gatekeeping on the Web"

arXiv:2510.10315. <https://arxiv.org/abs/2510.10315>

How it's referenced: The paper examines how robots.txt gatekeeping shapes what content is available for model training and finds systematic asymmetries in what the open web exposes versus withholds. Extended to reserves, the authoritative material: SPE technical papers behind OnePetro, SEC filings, proprietary operator databases, internal reserves workpapers is largely

paywalled, licensed, or private, and therefore underrepresented in the public text on which LLMs train. The model learns reserves methodology from what was freely scrapable, not from the primary sources practitioners rely on. This reserves-specific extension is an inference the paper's findings support, not a claim the paper makes about petroleum data.

Why read it: It offers a mechanism for a familiar pattern: LLMs are often confidently wrong about reserves methodology in a specific direction: fluent in the vocabulary, unreliable on the standards. If SPE-PRMS, the SEC definitions, and audited workpapers are not in the training data, the model reconstructs the field from its shadow on the open web. Cheng paper in this bibliography is an example of information that would be crucial for training sitting behind a paywall, hurting model training before it begins.

Hinton, Geoffrey (2025). "Will AI Outsmart Human Intelligence?"

Friday Evening Discourse, The Royal Institution, London, 2025. Video.

<https://www.youtube.com/watch?v=lkdziSLYzHw>

How it's referenced: A source on the inscrutability of LLM reasoning from one of AI's strongest proponents. Hinton is not a skeptic; he holds that these systems are genuinely intelligent, argues they may surpass human capability, and left Google to warn that AI is too powerful rather than too limited. In this lecture he nonetheless states that it is hard to analyze what LLMs learn and therefore hard to know whether they truly understand what they are saying. If the field's most prominent proponent grants that the reasoning is inscrutable, the inference follows directly: reasoning that cannot be inspected cannot be audited, reproduced, or signed for.

Why watch it: It answers the charge that the argument is anti-AI or built on selectively chosen pessimists. The concession comes from the "godfather of AI" himself, on the specific point that matters, whether AI's reasoning can be inspected and certified. The relevant admission is the inscrutability point; Hinton poses the "glorified autocomplete" framing only as a question he then argues against.

PHILOSOPHICAL FOUNDATIONS: SYMBOL GROUNDING & TACIT KNOWLEDGE

Harnad, Stevan (1990). "The Symbol Grounding Problem" *Physica D: Nonlinear Phenomena*, 42(1–3), 335–346.

[https://doi.org/10.1016/0167-2789\(90\)90087-6](https://doi.org/10.1016/0167-2789(90)90087-6)

How it's referenced: Harnad argues that manipulating symbols by their shape and their relationships to other symbols is not the same as understanding what those symbols mean; the symbols must be grounded in something beyond more symbols. His illustration is learning Chinese from a Chinese-to-Chinese dictionary: you could shuffle definitions forever and never connect a single word to the world. An LLM processing "boundary-dominated flow," "b-factor," or "4,300 bbl/mo" is doing exactly this: the tokens mean only their statistical proximity to other tokens. The model has never seen a reservoir, watched pressure decline, or observed a well water out. The CrudeCode skill instructs the model to read a production history and judge "what it is evidence of"—but evidence about a physical reservoir the model has no access to.

Why read it: This is the first rung of the argument and the foundation the others build on. It explains why an AI can generate a fluent, confident, syntactically correct estimate that is semantically empty—the numbers have no grounded connection to the physical reservoir the estimate claims to describe. Keep the claim narrow and defensible: the point is not that AI can never understand anything (a contested philosophy-of-mind position), but that these particular symbols, in this particular workflow, are not grounded in the physical system the estimate is supposed to represent.

Hume, David (1748). *An Enquiry Concerning Human Understanding*, Section IV: "Sceptical Doubts Concerning the Operations of the Understanding"

How it's referenced: Hume's problem of induction holds that no number of past observations can logically guarantee a future outcome; we assume the future will resemble the past, but that assumption cannot be justified without already presupposing it. Decline-curve analysis is induction in its purest form: past production is projected forward on the assumption the curve holds. **The point is not that DCA is invalid, but that something must close the gap between observed data and a justified future claim.** A human engineer closes it with reservoir physics, geological context, and professional judgment about why the curve should

hold and when it will break. An LLM closes it with pattern-matching over text; Humean habit with no physical grounding beneath the numbers.

Why read it: This is the philosophical spine of the P90 argument. Proved reserves require a justified 90% confidence bound, and induction alone cannot license a confidence level; that warrant must come from outside the data, from physics and professional standards. The engineer's signature certifies that the inductive leap was justified, which is precisely the step Hume showed cannot be skipped.

Note: Hume never used the word "induction" himself, and the memorable illustrations (the turkey, the sun failing to rise) are later additions by other philosophers; induction is merely a paraphrase for purposes of simplification and understanding for the audience to understand the argument.

Polanyi, Michael (1966). *The Tacit Dimension*

Doubleday & Company, Garden City, New York.

How it's referenced: Polanyi's thesis "we can know more than we can tell" holds that much expert knowledge is tacit: real and reliable, but impossible to fully articulate as stateable rules. You can recognize a face among thousands yet cannot write down how. Reserves judgment is tacit in exactly this way: an engineer knows when a decline curve doesn't look right, when an offset is truly analogous, when a transient has ended. The CrudeCode skill concedes this on its own terms, instructing the model that "much of this is visual; that's fine" and that "you look and you decide." The document's authors were too honest to pretend the decisive judgments were mechanical, and in admitting it they named the problem: the skill asks for tacit expertise on every judgment line.

Why read it: This explains why documentation and prompts are not enough. The moment tacit knowledge is written down completely, it stops being the thing that made the expert an expert. A human engineer's tacit judgment is earned by years of watching real wells behave; when an LLM executes the same "look and decide" instruction, it has no tacit expertise to draw on and substitutes pattern-matching, then produces a rationale that reads like earned judgment – in some computer science and AI research circles some people consider this hallucination or approximation of a positive output. Automating reserves estimation would require codifying tacit judgment into an instruction set, which is precisely what cannot be done. Although this work will not provide the evidence for "AI will and can hallucinate" it helps to propose the question of is there knowledge that AI just can't know? That question can serve as a guideline to where AI should, and should not be applied.

Dahl, Roald (1953). "The Great Automatic Grammatizator"

How it's referenced: Dahl's short story imagines a machine that mass-produces fiction, and the writers who are gradually pressured into leasing their names to its output. It is a prophetic study

of what happens when human professionals are replaced by an automatic system: the machine has fluency but no taste, no judgment, and—crucially—no one to hold responsible. The final turn is a moral one: the narrator watching writers surrender their names to the machine, and the reader left to weigh what is lost when authorship and accountability are decoupled from the work.

Why read it: This is the closing rung; a humanistic argument rather than a technical one, and the emotional payoff of the title. It frames the reserves problem in terms of professional responsibility and institutional trust: the question is not only whether the machine can produce an estimate, but who is accountable when the estimate is wrong. The engineer's name on the reserve report is the accountability the machine cannot supply. (Publication note: "The Great Automatic Grammatizator" first appeared in Dahl's 1953 collection *Someone Like You*; it is sometimes cited via the later anthology *The Umbrella Man and Other Stories*.)

PETROLEUM ENGINEERING & RESERVES DOMAIN

Ovintiv Inc. Reserve Audit Opinion — McDaniel & Associates Consultants Ltd. (2022)

Exhibit 99.1 to SEC Form 10-K, filed January 25, 2022. Signed by David M. Sawyer, P.Eng., Executive Vice President, McDaniel & Associates Consultants Ltd. (APEGA Permit P3145). https://www.sec.gov/Archives/edgar/data/1792580/000156459022006978/ovv-ex991_711.htm

How it's referenced: This is what professional accountability looks like as a physical artifact. The audit opinion carries a Certificate of Qualification: a named engineer (David M. Sawyer), his degree (Petroleum Engineering, University of Calgary), his license (registered Professional Engineer with APEGA Permit Number P3145), his firm and permit number, his years of experience, and his signature—attesting that the assumptions, data, methods, and procedures are appropriate, and that the estimates conform to SEC Regulation S-X 210.4-10(a) and Regulation S-K 229.1202. The document also states that the firm is independent, has no financial interest in the properties, and that its compensation is not contingent on the reserve estimates.

Every element the presentation showcases human professional attributes AI cannot supply are present on this one page: a name, a credential, a standard, an independence attestation, and personal responsibility for the number.

Why read it: For PEs and QREs, this is the concrete reference point for "defensible." It shows what a credible reserves estimate is required to carry before it can be filed with the SEC and, by direct contrast, exactly what an AI-generated estimate cannot carry. When evaluating any AI-assisted workflow, hold its output against this page and ask what is missing. The answer is:

the accountability chain. This is the single highest-value visual in the deck because it makes the accountability argument tangible rather than abstract.

Cheng, Y., W. J. Lee, and D. A. McVay (2008). "Improving Reserves Estimates From Decline-Curve Analysis of Tight and Multilayer Gas Wells"

SPE Reservoir Evaluation & Engineering, 11(5), 912–920. SPE-108176-PA.

<https://doi.org/10.2118/108176-PA>

How it's referenced: This paper documents a problem at the heart of the CrudeCode critique: for tight and multilayer gas wells, decline-curve analysis frequently yields an Arps b-value greater than 1, and different analysts handle that differently; some use the matched b, some force b to unity, some switch to exponential decline at a chosen limit. The authors state plainly that forecasted performance differs significantly when different analysts analyze the same data, producing large uncertainty in the reserve estimate. The cause is physical: long-duration transient effects in stimulated tight wells, and coexisting flow regimes across commingled layers, produce abnormal decline behavior ($b > 1$) that a naive fit misreads. This is the peer-reviewed statement of the exact anchor-point and b-factor problem the presentation raises, that fitting through a transient overstates b, and overstated b overstates reserves.

Why read it: It grounds the anchor-point argument in the SPE literature rather than in assertion. The failure mode is not an AI problem in origin: it is a known, documented hazard of decline-curve analysis that requires physical judgment (identifying boundary-dominated flow, recognizing transient behavior) to navigate. That is precisely the judgment an LLM cannot supply, which is why an autofit routine (human-configured or AI-driven) inherits the hazard. (Note: this is the same paper cited as SPE-108176-PA in Section 2; provided here to match use case)

Kim, Sungil, Tea-Woo Kim, and Suryeom Jo (2025). "Artificial intelligence in geoenergy: bridging petroleum engineering and future-oriented applications." *Journal of Petroleum Exploration and Production Technology* 15:35. <https://doi.org/10.1007/s13202-025-01939-3>

Surveys on AI in Petroleum Engineering (honest-broker sources) to showcase positive applications of AI

- Xue et al. "Artificial Intelligence Methods for Oil and Gas Reservoir Development: Current Progresses and Perspectives." Peer-reviewed assessment of where AI is applied in upstream workflows and where its limitations lie.

- "Artificial Intelligence in Geoenergy: Bridging Petroleum Engineering and Future-Oriented Applications." Domain review of AI integration opportunities and constraints.
- "Research Status and Development Directions of Artificial Intelligence in Reservoir Protection." Analysis of where AI capability maps onto reserves-adjacent tasks.

How they're referenced: These are the honest-broker sources: balanced, peer-reviewed views of where AI genuinely helps in upstream work (seismic processing, drilling optimization, equipment-failure prediction, history matching) versus where the evidence does not support it (SEC-grade reserves estimation). The presentation uses them to draw the line the vendor marketing blurs: AI's documented successes cluster in pattern-rich, physically-constrained, tolerant-of-error tasks, not in the reproducible, auditable, single-number judgment that reserves reporting demands.

Why read them: They are what makes the talk credible rather than adversarial. The argument is not anti-AI; it is pro-standards and pro-defensibility. These surveys allow me (the presenter) to demonstrate genuine command of where AI adds value, establishing that the reserves critique is a targeted, evidence-based boundary, not a blanket dismissal.

Citing them is what allows me to re-direct attention from reserved workflows to other areas where AI investment may be genuinely justified.

CRUDECODE MCP APPLICATION (SOURCE CODE ANALYSIS)

CrudeCode Contributors. "CrudeCode MCP Application"

GitHub repository. <https://github.com/crude-code/mcp-app> Archive analyzed: `mcp-app-main.zip`. Re-audited September 4, 2026.

⚠ THIS ENTIRE SECTION IS REWRITTEN. The v1 section contained a factual error and cited a filename that no longer exists. Do not use the v1 text.

How it's referenced: The open-source package for CrudeCode is examined to show the design–execution gap. The well-forecasting skill is a conceptually sound engineering procedure. The implementation assigns execution of every judgment call in that procedure to Claude — a nondeterministic language model — and confines the deterministic code to arithmetic downstream of those judgments.

The central finding: there is no optimizer

- **SciPy is not a dependency.** `requirements.txt` lists numpy and pandas. There is no optimizer package anywhere in the environment.
- `server/valuation/forecast.py` states in its own module docstring that there is no fitting and no parameter selection anywhere in the module or the server; decline parameters are asserted upstream by Claude and arrive as literals. The module evaluates the Arps hyperbolic relation and applies a terminal-decline switch. That is its entire scope.
- **This was a deliberate removal, and it is documented.** `docs/research/2026-07-23-forecasting-repositioning-ideas.md` records the decision to delete the prior engine — described as a classification ladder, `scipy curve_fit`, and type-curve construction — on the reasoning that well forecasting is not a good deterministic problem and that the old approach compiled judgment too early into rules. The design thesis is stated in the document as: **"Python does deterministic work; Claude does probabilistic work."** Parameter selection is characterized as judgment, and moved to Claude.

The argument is therefore not that a regression is wrapped in a nondeterministic shell. It is that **there is no regression**. The model asserts three numbers per stream and the calculator evaluates them.

Key files examined

File	What it establishes
<code>requirements.txt</code>	No SciPy. No optimizer of any kind.
<code>docs/research/2026-07-23-forecasting-repositioning-ideas.md</code>	The design record: the fitting engine was deleted on purpose, and parameter selection was reassigned to the model. Contains the "Python does deterministic work; Claude does probabilistic work" thesis, and the statement that there is no optimizer anywhere in the backend. Also states that guardrails and hindcast grading were deferred, and that the tool is not being used to book reserves yet.

<code>skills/well-forecasting/SKILL.md</code>	The design specification — a six-question engineering procedure covering history interpretation, trust in the well's own record, anchor and q_i selection, D_i , b , and consequence review. Methodologically sound. Explicitly concedes that much of the work is visual. Its b -doctrine independently restates the transient/ $b > 1$ hazard found in the SPE literature.
<code>prompts/outer/system_prompt.md</code>	Governs the model's posture. Instructs that its job is judgment. No enforcement mechanism binds it to the skill's procedure.
<code>prompts/outer/tool_deal_forecast_wells.md</code>	The parameter-assertion contract. States that the server never chooses a parameter and that the judgment belongs to the model. Documents the accepted payload and the bounds.
<code>server/valuation/orchestrator.py</code>	Bounds validation. $q_i > 0$; D_i strictly inside $(0, 1)$ nominal monthly; b inclusive on $[0, 2]$; uptime factor on $[0.5, 1.0]$. The tool spec characterizes these as sanity rails rather than judgment: anything inside the rails commits.
<code>server/valuation/config.py</code>	The defaults the model never sees. See below.
<code>server/valuation/econ.py</code>	Revenue $\rightarrow$ cashflow $\rightarrow$ NPV. Contains no economic-limit test, no shut-in logic, and no negative-cashflow truncation.
<code>server/valuation/forecast.py</code>	Hyperbolic evaluation and the terminal switch. Deterministic given parameters — and given parameters only.

The defaults finding

`server/valuation/config.py` holds every economic parameter as a frozen dataclass. The shipped defaults:

- **Terminal decline: 5% per year.** One global value, applied to every well in every basin and formation. The code comments describe it as belonging to the calculator rather than to judgment — it is the one forecast constant the model does not set and cannot see.
- **Flat price deck: \$70.00/bbl oil, \$3.50/MMBtu gas.** Not the SEC 12-month first-day-of-month average. Not a strip.
- **Operating cost: zero.** `opex_per_bbl_usd`, `opex_per_well_per_month_usd`, and `capex_per_well_usd` all default to 0.0.
- **Horizon: 360 months.**

The consequence. With zero operating cost, net cashflow never turns negative, so no well ever reaches an economic limit — and no economic-limit test exists in the code to catch it if one did. The skill defines EUR as cum-to-date plus the full run-out through the terminal tail. The forecast therefore terminates at month 360 because the array ends, not because the well stops paying.

Under Reg S-X 4-10, reserves are the **economically producible** quantities. A run-out volume at zero opex on a flat deck with a fixed 30-year cutoff is not a reserve; it is a different object wearing the same word. This holds regardless of how the decline parameters were chosen, which makes it the most robust finding in the audit — it requires no claim about AI at all.

The current `tool_deal_forecast_wells` contract requires the opposite:

- **rationale is a required field.** Validation rejects the call outright if it is missing or empty. The spec calls for the six-question record from the skill, written so another engineer could disagree with a specific line.
- **analog_cohort.excluded** accepts `[{api, reason}]` — candidate analogs rejected and the reason for each. The spec describes the reasons as the evidence.
- **struck_months** is an optional audit list of months excluded from the read.
- **analog_cohort.kept** APIs are hydrated against real data; a fabricated API causes the call to bounce.
- **Validation is all-or-nothing:** any violation returns every problem and saves nothing.

The argument survives the correction, and is sharper for it. CrudeCode does capture a rationale. What it cannot capture is *reproducibility*. A recorded rationale is a post-hoc narrative generated by the same nondeterministic process that produced the parameters; re-running the identical well can produce different parameters and a different, equally fluent rationale. Documentation is not determinism. The record explains one run — it does not bind the next one.

The design–execution gap, with a concrete instance

The skill's b-doctrine is genuinely good engineering writing. It warns that b is a population quantity, that a well's own early record is transient- and flush-dominated, that fitting b to history is how forecasting goes blind, and that a b above 1 is an assertion about the transient segment only. It names the documented history of overstated shale EURs as this exact error.

And nothing in the system enforces any of it. The bounds accept b up to 2.0 inclusive. The tool spec's own worked example asserts $b = 1.1$ for oil and $b = 1.2$ for gas. No check exists to determine whether the model followed the doctrine, ignored it, or reconstructed something that resembles it.

That is the design–execution gap stated concretely: **the document knows precisely what the hazard is, and the executor is under no obligation to avoid it.**

Why read it: The codebase moves the argument from theoretical ("AI might struggle with judgment calls") to empirical ("here is the file where this specific tool assigns judgment to a nondeterministic system, and here is the design document where its authors say that is what they intended"). Every claim in Part 2 of the presentation should be traceable to a named file in this table.

⚠ **FAIRNESS — required framing.** The design document states that guardrails and hindcast grading were deferred and that the tool is not being used to book reserves yet. **CrudeCode's authors do not claim it books reserves.**

LOCAL-AI CAPABILITY & EFFICIENCY

Saad-Falcon, J., Narayan, A., et al. (2025). "Intelligence per Watt: Measuring Intelligence Efficiency of Local AI"

arXiv:2511.07885. Stanford University and Together AI.

How it's referenced: This is a local-AI efficiency paper, **not** a reserves paper and not a limits-of-AI paper. Its headline findings are pro-AI-viability: roughly 88.7% of single-turn queries in the study are handled acceptably by local models, and intelligence-per-watt improves by about 5.3× over the period examined.

Four defensible uses, in ascending order of value:

1. **Engineering is the hard case.** The domain breakdown shows Architecture & Engineering solvability at 40.8% on NaturalReasoning against 98.9% for the same domain on chat-style queries (Table 4). Best-of-local coverage drops to roughly 60% for technical disciplines (Figure 5). GPT-OSS-120B scores 51.6% on SuperGPQA

Architecture & Engineering (Table 5). The aggregate number is carried by easy domains; the domain nearest to reserves work is where performance falls off.

2. **The hardest reasoning remains unsolved.** Level 5 reasoning tasks are approximately 95% unsolved, improving at roughly 1.5 percentage points per year (Section E.2).
3. **Metric mismatch, the strongest use:** The paper defines success as LLM-judge win/tie or exact match. That is an accuracy metric. Certification requires accountability, reproducibility, and auditability. These are orthogonal axes, not points on the same scale. At 88.7% success by the paper's own definition, the engineer's name is still the only name on the report. This is the cleanest available demonstration that even pro-AI-viability data cannot speak to the question the presentation is asking.
4. **Nondeterminism corroboration.** Decoding uses temperature 0.6 and top-p 0.95 (Section B.1). The authors concede 10–15% telemetry error and acknowledge LLM-judge bias in their Limitations section.

Why read it: Use 3 is the reason this source earns a place. It is a rigorous, recent, favorable-to-AI paper whose own success metric demonstrates the category error the presentation is built to correct. Best placement is the honest-broker section, or as the empirical anchor in the induction and reasoning-limits argument.

Last Updated: September 4, 2026 Prepared for: PHDwin Week 2026 – AI Can't Sign Your Reserve Report: Defensibility, Accountability, and the Limits of AI in Reserves Work