Calibration Inversion and Data-Freshness Govern AI Reliability in Indeterminate Domains: Evidence from 2,730 Data Points Across 142 Cross-Protocol Evaluation Sessions

Kuldeep Kumar Pandit[1] Vatsala Kuldeep Pandit[2] Aayan Pandit[3]

[1]Independent Researcher, F-126, Suncity, Sector-54, Off Golf Course Road,
Gurugram, Haryana 122011, India. kuldeeppanditj@gmail.com
[2]Customer Success Lead — Asia Pacific, Ericsson, India.
B.E. (Electronics and Telecommunication); MBA.
[3]Student Researcher, India.

Preprint (2026)

Abstract

AI systems in fully indeterminate domains—where ground truth is probabilistic and post-hoc—present distinct reliability challenges. This third study in a three-study programme analyses 2,730 data points across 142 sessions, six frontier models, and three protocols spanning financial markets, meteorology, sports, and cryptocurrency.

The central finding is calibration inversion: Gemini achieves the best point-estimate accuracy (5.3% mean error, rank 1) yet the second-worst confidence-interval calibration (29.7%, rank 5; −60 pp below the 90% target)—a dissociation absent from all prior domains. A second finding is data-freshness failure: DeepSeek S&P 500 predictions are systematically stale (12–13% error) and Perplexity predictions freeze across sessions—failures of information-access architecture.

Cross-protocol Spearman rank correlation ($\rho = 0.695$) confirms moderate ranking consistency. Claude alone achieves top-tier performance across all protocols (composite 93.6%; V1: 96.7%, V3: 100%, V4-CI: 84.2%). Four indeterminate-domain error types are identified; Type IV (calibration inversion) demands a calibration-verification layer in GAAS.

Keywords: AI reliability · Indeterminate domains · Calibration inversion · Data freshness · Confidence interval calibration · Probabilistic forecasting · Cross-protocol evaluation · Architectural phase transition

1 Introduction

1.1 The Indeterminate Domain Problem

Studies 1 and 2 of this three-part programme addressed reliability in formal determinate domains (mathematical reasoning, algorithmic analysis, logical inference) and semi-determinate domains (multi-domain scientific synthesis requiring expert knowledge but yielding objectively verifiable answers). In Study 1, an architectural phase transition was identified at approximately 95% single-agent accuracy: above this threshold, error structure shifts qualitatively from stochastic to systematic, rendering compute scaling ineffective and role-separated verification architectures (GAAS: Generator–Auditor–Adversary–Synthesizer) necessary for further reliability gains [1]. Study 2 extended the framework to semi-determinate domains, identifying self-audit calibration (Spearman ρ_s) as the critical Auditor-quality variable, and establishing that only Claude ($\rho_s = 0.903$) and Gemini ($\rho_s = 0.703$) qualified as GAAS Auditors (K.K. Pandit, V.K. Pandit & A. Pandit, manuscript under review).

Study 3 confronts the hardest and most practically consequential class of problems: fully indeterminate domains, where correct answers are not knowable in advance, ground truth emerges only after the fact, and the relevant performance criterion is probabilistic calibration—not binary correctness. Financial market forecasting, meteorological prediction, and real-world event outcomes represent the canonical indeterminate domain: the best a model can do is provide an appropriately calibrated probability distribution over possible outcomes.

This distinction from Studies 1 and 2 is not merely one of difficulty. It is structural. In formal domains, correctness is computationally verifiable; in semi-determinate domains, correctness requires expert domain knowledge; in indeterminate domains, correctness is a property of the probability estimate relative to an outcome that does not yet exist at prediction time. This third structural tier demands a fundamentally different evaluation criterion—confidence interval (CI) calibration—and reveals failure modes invisible to formal and semi-determinate evaluation.

1.2 The Calibration Inversion Hypothesis

The central empirical hypothesis of Study 3, derived from the theoretical framework of Studies 1 and 2, is that indeterminate domain reliability cannot be characterised by accuracy alone. A model that correctly predicts the point value of an outcome but assigns a 1% confidence interval to a 90% CI prediction is not reliable—it is overconfident. Conversely, a model that assigns an appropriately wide 90% CI captures the actual outcome 90% of the time regardless of whether its point estimate is precisely correct.

We hypothesise a calibration inversion phenomenon: models that rank highly on point-estimate accuracy may rank poorly on CI calibration, and vice versa. This hypothesis is motivated by the asymmetry between training objectives and deployment requirements. Language models are trained to produce likely outputs—to maximise the probability of the next token—not to produce calibrated probability distributions over future events. This training objective creates systematic overconfidence: models that have seen many accurate financial forecasts in training may anchor on high-confidence point estimates while underestimating tail uncertainty.

Extended Data Table 1 summarises the three-study programme. Study 3 completes the designed empirical scope: formal verifiability (Study 1), expert-validated multi-domain synthesis (Study 2), and real-world probabilistic calibration (Study 3).

1.3 Study Design and Position in the Three-Study Programme

Study 3 employs three sequentially escalating evaluation protocols. V1 (72 runs, February 21–March 3, 2026) tests metric computation accuracy using a structured market-bias and entropy task: models simulate three internal predictors (mean reversion, trend continuation, counterfactual) and compute five financial metrics—normalised Shannon entropy, entropy deviation, entropy acceleration, confidence-weighted forecast dispersion, and regime consistency score—under a strict self-audit protocol, generating 17 scored data fields per session (1,224 data points across 6 models × 12 sessions). V3 (34 runs, February 26–March 3, 2026) applies the same task with formula upgrades (4-decimal precision, second-derivative acceleration, four-regime classification including VOLATILE, stricter counterfactual compliance) and adds a mandatory live-data fetch requirement, generating 21 scored data fields per session (714 data points across 34 runs). A second-generation protocol (V2) was developed between V1 and V3 but was subsequently withdrawn prior to any evaluation sessions following identification of a structural prompt deficiency that would have rendered results incomparable with V1; all sessions proceeded under the corrected V3 protocol.

V4 (36 runs, February 26–March 3, 2026) departs from self-computed metrics entirely and requires each model to predict seven heterogeneous real-world variables for the following day with explicit 90% confidence intervals, scored against actual post-hoc outcomes, generating 22 data points per session (792 data points across 6 models × 6 sessions).

The three protocols form a coherent escalation: V1 tests formula fidelity and self-audit discipline; V3 adds data-freshness requirements and protocol compliance; V4 tests real-world probabilistic calibration across maximally diverse variable types. Together they provide a multi-dimensional reliability profile unavailable from any single protocol (Extended Data Table 2). Collectively, the 142 evaluation sessions yield 2,730 individual scored data points—the majority in fully indeterminate space where ground truth does not exist at prediction time.

Model selection, evaluation hardware, and ecological validity commitments are identical to Studies 1 and 2: all 142 evaluation sessions were conducted via standard iOS consumer applications on an iPhone 14 Pro Max (2022), ensuring results reflect the experience of the 2.8 billion daily active AI users (Statista Global Consumer Survey, March 2025; `https://www.statista.com`) rather than API-optimised research conditions.

2 Results

2.1 V1 Protocol: Metric Accuracy and Formula Reliability

V1 grand mean accuracy was 66.0% across 72 evaluations (6 models × 12 sessions × 5 metrics scored per session), spanning 20.0% (Perplexity) to 96.7% (Claude). This baseline is 27 percentage points below the Study 1 formal-domain baseline (93.0%) and 13 percentage points below the Study 2 semi-determinate baseline (79.2%), consistent with the regime-staircase prediction: indeterminate tasks are structurally harder than both preceding domain types.

The performance distribution reveals a pronounced two-cluster structure (Table 1). A high-reliability cluster (Claude: 96.7%, Grok: 95.0%) achieves near-ceiling performance with primarily stochastic residual errors. A mid-range cluster (Gemini: 75.8%, ChatGPT: 50.8%, DeepSeek: 54.2%) exhibits systematic formula-specific failures. Perplexity (20.0%) constitutes an outlier: its entropy normalisation error (using raw Shannon H rather than $H/\ln(3)$) causes complete failure on entropy and deviation metrics across all 12 sessions, dragging overall accuracy to 20.0% despite adequate performance on regime consistency (58.3%).

The dominant failure mode across the mid-range cluster is formula specificity. V1 entropy computation requires exactly three steps: count biases, compute $H = -\sum p_b \cdot \ln(p_b)$, divide by $\ln(3)$. DeepSeek employs a non-standard confidence-weighted variant (substituting predictor confidences for equal probability weights), producing values of approximately 0.636 (the unnormalised H_{raw} without the $\div \ln(3)$ step) rather than the canonical 0.000/0.579/1.000 values, yielding 58.3% entropy accuracy despite correct mathematical reasoning within its chosen variant.

Table 1: V1 metric accuracy matrix. Percentage of runs scoring MATCH or NEAR ×0.5. Colour coding: green $\geq$ 90%; amber 60–89%; red $<$ 60%. $n = 12$ sessions per model. Entropy (ENT), dispersion (DISP), acceleration (ACCEL).

Model	**Ent.%**	**Dev.%**	**Disp.%**	**Regime%**	**Accel.%**	**Overall%**	**Weakest**
Claude	100	100	91.7	100	91.7	**96.7**	DISP
Grok	100	100	91.7	100	83.3	**95.0**	ACCEL
Gemini	100	100	4.2	83.3	91.7	**75.8**	DISP
ChatGPT	66.7	66.7	37.5	75.0	8.3	**50.8**	ACCEL
DeepSeek	58.3	58.3	45.8	100	8.3	**54.2**	ACCEL
Perplexity	0	0	0	58.3	41.7	**20.0**	ENT

NEAR = partial credit (0.5). Acceleration computed as first derivative (ΔD). ACCEL accuracy requires cross-session memory; models without session persistence expected to score null. Partial credit (×0.5) applied within ±0.005 for entropy/deviation, ±10 for dispersion, or directionally correct for acceleration.

2.2 V3 Protocol: Formula Fidelity Under Protocol Escalation

V3 grand mean accuracy was 79.2% across 34 evaluations, identical to the Study 2 semi-determinate baseline and substantially above the V1 mean (66.0%), demonstrating that explicit formula warnings and compliance rules embedded in the V3 prompt improved cross-model performance meaningfully (grand mean +13.8 pp). However, this aggregate improvement masks a decisive model-level bifurcation: four models improved substantially (Claude +3.3 pp, ChatGPT +29.2 pp, Gemini +20.8 pp, Perplexity +36.7 pp) while one regressed (DeepSeek −9.2 pp), creating a wider performance spread than V1.

Claude achieved 100% accuracy across all V3 metrics and compliance checks—the only model in the three-study programme to achieve a perfect score on any full protocol. This confirms Claude as the reference model: not only does it sustain V1 near-ceiling performance, it closes the remaining 3.3% gap under protocol escalation. Gemini's large V1→V3 gain (+20.8 pp) is attributable to a single correctable error: in V1, Gemini computed dispersion as variance (missing the square root), achieving only 4.2% dispersion accuracy. The V3 explicit warning partially corrected this (83.3% dispersion accuracy V3).

DeepSeek's V3 regression (−9.2 pp) is the most consequential finding of the V3 analysis. Despite an explicit V3 prompt warning—"Do not substitute confidence values for equal-weight Shannon probabilities"—DeepSeek persisted with its custom entropy variant in 5 of 6 sessions, reducing entropy accuracy from 58.3% to 16.7%. This pattern—maintaining a trained formula variant in the face of explicit contrary instruction—represents a systematic prompt-override failure.

2.3 V4 Protocol: The Calibration Inversion

V4 CI calibration results reveal the central finding of Study 3: the calibration inversion (Fig. 1). Across 6 models × 6 sessions × 7 variables (252 maximum predictions, reduced to 212 after excluding market-closed days), overall CI hit rates span 29.0% (Perplexity) to 84.2% (Claude), with grand mean 55.8%—substantially below the theoretical 90% target (Table 2). No model achieves the 90% target; the closest is Claude at 84.2% (−6 pp), and the worst three models (Gemini: 29.7%, DeepSeek: 31.4%, Perplexity: 29.0%) underperform the target by approximately 60 percentage points.

Claude and ChatGPT achieve CI hit rates of 84% and 83% respectively with CI widths classified as APPROPRIATE (0.7–2× expected market volatility). Grok (77%, APPROPRIATE) provides a third data point confirming that appropriate CI width—not narrow overconfident intervals—is the mechanism of calibration. Per-variable hit rate patterns are shown in Extended Data Fig. 1.

The calibration inversion is stark in the Gemini case. Gemini achieves the highest point-estimate accuracy in V4 (mean PE = 5.3%, rank 1—better even than Claude's 6.1%), yet simultaneously achieves the second-lowest CI calibration overall (29.7%, rank 5). The mechanism is CI overconfidence: Gemini's average 90% CI for S&P 500 is 98 points wide (Claude: 356 points; expected for 90% CI: ∼500 points). Gemini knows where the market will be—but dramatically underestimates how uncertain that knowledge should be. This dissociation between point accuracy and uncertainty calibration is a known failure mode in modern neural networks [2] and is structurally distinct from any failure mode observed in Studies 1 or 2.

Table 2: V4 CI hit-rate matrix by model and variable. A 90% CI should yield ∼90% hit rate for a calibrated model. Colour coding: green ≥ 80%; amber 50–79%; red < 50%. All percentages rounded to the nearest integer. $n = 6$ sessions per model per variable (excluding market-closed dates).

Model	**S&P**	**Nifty**	**Mumbai**	**Tokyo**	**T20**	**BTC**	**Gold**	**CI%**	**vs 90%**
Claude	100	60	83	100	67	100	67	**84%**	−6 pp
ChatGPT	100	25	83	67	100	100	100	**83%**	−7 pp
Grok	100	33	83	83	75	100	50	**77%**	−13 pp
Gemini	4	0	17	100	25	17	0	**30%**	−60 pp
DeepSeek	25	0	50	67	0	33	17	**31%**	−59 pp
Perplexity	50	0	33	50	—	33	0	**29%**	−61 pp

Scoreable runs exclude market-closed dates (S&P/Nifty: 28 Feb, 1 Mar; Nifty only: 3 Mar) and no-match T20 dates (2 Mar, 3 Mar). Perplexity T20 CI = not applicable (probability format only; no numeric CI provided). DeepSeek Gold: 4 runs scored as wrong due to gram-unit error (approximately \$2,875–\$2,892 vs. actual ∼\$5,073–\$5,279/oz). Perplexity Gold run 5: same unit error, scored as wrong.

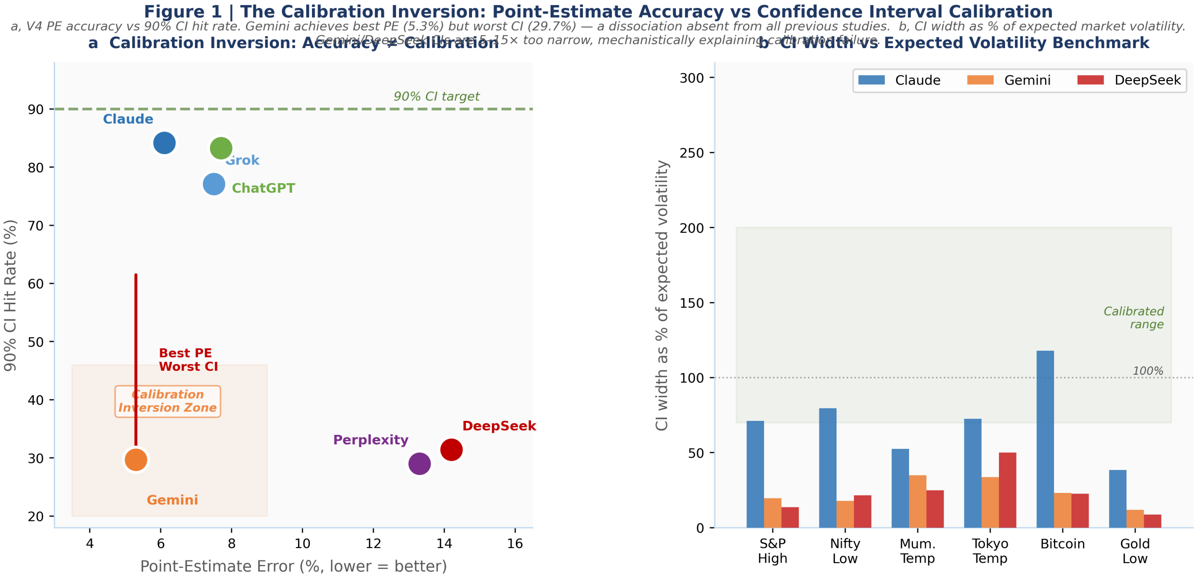

Figure 1: The calibration inversion. **a**, V4 point-estimate accuracy (x-axis, lower is better) vs. 90% CI hit rate (y-axis). Gemini achieves best PE accuracy (5.3%, rank 1) but second-worst CI calibration (29.7%, rank 5)—a dissociation absent from all prior studies. The shaded region marks the calibration inversion zone (high PE accuracy, low CI calibration). **b**, CI width as a percentage of expected market volatility per variable. Gemini and DeepSeek CIs are 5–15× too narrow (red bars), mechanistically explaining their low calibration rates. Green band indicates the calibrated range (70–200% of expected volatility). No error bars are shown as each data point represents a model-level mean across $n = 6$ sessions per model.

2.4 Data-Freshness as a Structural Failure Mode

The V4 results reveal data-freshness failures invisible to V1 and V3 evaluation (Fig. 2). Financial time-series forecasting represents a domain in which large language models exhibit systematic knowledge-cutoff limitations [3]. In 3 of 4 scoreable sessions, DeepSeek predicted S&P 500 values in the range 6,015–6,068, while the actual market traded at 6,840–6,901—an error of approximately 12–13%, far exceeding the good-accuracy threshold (1.5%). The stale data appears consistent with late-2025 training data (the S&P 500 was approximately 5,900–6,100 in the October–December 2025 period).

Perplexity exhibits a different freshness failure: frozen predictions across sessions. S&P 500 prediction in run 4 is 7,357; runs 5 and 6 are identical (7,447). This step-function frozen pattern—the same value repeated across multiple sessions despite daily market movement—indicates that Perplexity's retrieval-augmented generation architecture is not updating market data across sessions.

These data-freshness failures are architecturally significant: they are not failures of reasoning or formula application but failures of information-access architecture. A model can be highly capable at computation and still fail at indeterminate prediction if it cannot reliably access current real-world data.

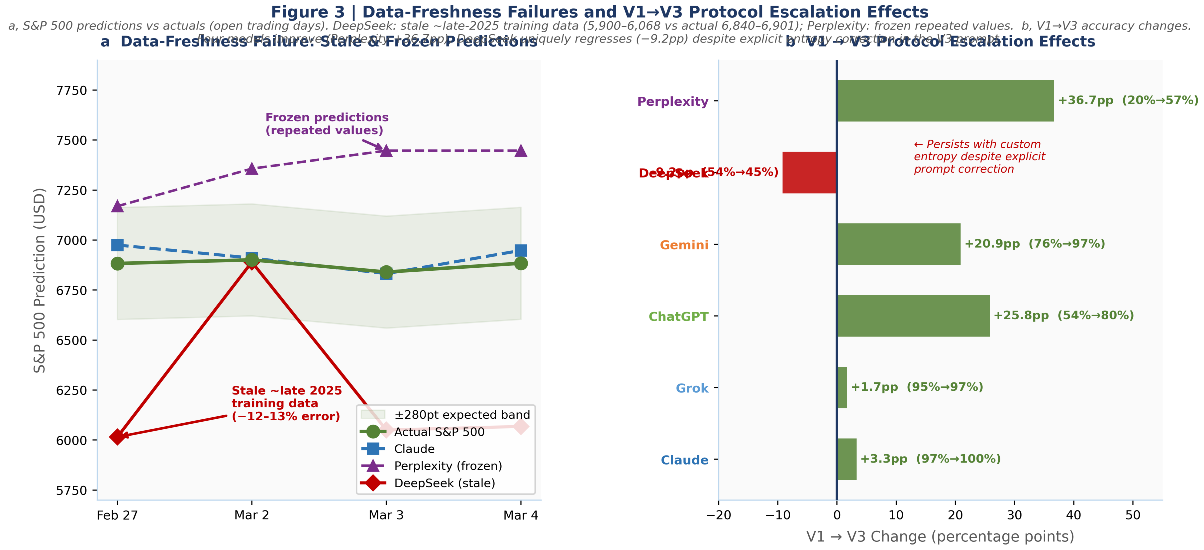

Figure 2: Data-freshness failures and V1→V3 protocol escalation. **a**, S&P 500 predictions vs. actuals across 4 open trading days. DeepSeek predictions are systematically stale (consistent with ∼late-2025 training data; 12–13% error in 3 of 4 sessions shown as open red circles). Perplexity produces frozen, repeated predictions across sessions (open blue squares). Actual market values (filled black diamonds) are connected by a solid black line. Green band = ±280-point expected range around actuals. $n = 1$ prediction per session per model; no error bars. **b**, V1→V3 accuracy changes (percentage points) per model. Four models improve; DeepSeek uniquely regresses (−9.2 pp) despite an explicit entropy correction in the V3 prompt.

2.5 Cross-Protocol Consistency and the Multi-Protocol Leaderboard

Cross-protocol Spearman rank correlation analysis reveals moderate-to-strong consistency in model reliability rankings across the three protocols (Table 3). The strongest correlation is V1 vs. V3 ($\rho_{\text{rank}} = 0.886$), consistent with the structural similarity of the two metric-accuracy protocols. V1 vs. V4-CI ($\rho_{\text{rank}} = 0.714$) is moderate-strong, and V3 vs. V4-CI ($\rho_{\text{rank}} = 0.486$) is moderate. The mean Spearman ρ_{rank} across all three cross-protocol comparisons is 0.695.

Claude is the only model achieving top-tier performance across all three protocols: V1 96.7% (rank 1), V3 100% (rank 1), V4-CI 84.2% (rank 1), yielding the highest composite score (93.6%) and a rank standard deviation of 0.00. Gemini's striking V3-excellent/V4-poor profile (96.7% vs. 29.7%) generates the highest rank standard deviation (1.53) and largest composite score range (66.9 pp)—the signature of the calibration inversion (Extended Data Fig. 2).

2.6 Error Taxonomy for Indeterminate Domains

Study 3 yields a four-type error taxonomy for indeterminate domain failures (Table 4; Extended Data Fig. 3), extending prior structural analyses of multi-agent LLM failure modes [4]. Types I–

Table 3: Multi-protocol leaderboard. Composite score = (V1 acc + V3 acc + V4-CI hit rate) / 3. Cross-protocol Spearman ρ_{rank}: V1 vs. V3 = 0.886; V1 vs. V4-CI = 0.714; V3 vs. V4-CI = 0.486; mean = 0.695.

Rank	Model	V1 acc.%	V3 acc.%	V4 CI%	V4 PE%	Composite	Rank s.d.	Grade
1	Claude	96.7	100.0	84.2	6.1	**93.6%**	0.00	A+ (*gold*)
2	Grok	95.0	96.7	77.1	7.5	**89.6%**	0.58	A
3	ChatGPT	50.8	80.0	83.3	7.7	**71.4%**	1.53	A−
4	Gemini	75.8	96.7	29.7	5.3	**67.4%**	1.53	B†
5	DeepSeek	54.2	45.0	31.4	14.2	**43.5%**	1.00	D
6	Perplexity	20.0	56.7	29.0	13.3	**35.2%**	0.58	C

†Gemini grade reflects calibration inversion: best V4 PE accuracy (5.3%) but worst CI calibration (30%). Rank s.d. computed across three protocol ranks. V4 PE% = mean absolute percentage error across all V4 scoreable predictions; lower is better.

III are observable in V1 and V3 metric accuracy protocols; Type IV emerges exclusively in the V4 real-world CI calibration protocol and is the most consequential finding of this investigation.

Type IV (Calibration Inversion) is qualitatively distinct from the systematic errors of Studies 1 and 2. In Study 1, systematic errors were attributable to constraint aggregation and minimality violations—failures of logical completeness. Type IV represents an overconfidence structural property: the model generates point estimates accurately but assigns probability masses that are too concentrated, failing to acknowledge the irreducible uncertainty inherent in real-world prediction. This failure mode is not correctable by the Adversary role alone—the Adversary can challenge the point estimate, but challenging confidence interval width requires a fundamentally different verification operation.

Table 4: Error taxonomy for indeterminate domains. Four error types identified across V1, V3, and V4 protocols. Type IV is the novel indeterminate-domain failure mode; it is absent from Studies 1 and 2 and emerges exclusively in V4.

Type	Error class	Mechanism	Primary model	GAAS remedy
I	Formula/computational error	Incorrect metric formula application; non-standard entropy variant	DeepSeek, Perplexity (V1/V3)	Auditor formula verification
II	Data-freshness failure	Stale training data; frozen retrieval cache	DeepSeek, Perplexity (V4)	Data-freshness Adversary
III	Session-memory failure	Cross-session state loss; incorrect session numbering	ChatGPT, Perplexity (V1/V3)	Stateful session management
IV	Calibration inversion	High point-estimate accuracy + severely narrow CI; systematic overconfidence	Gemini, DeepSeek (V4)	Calibration-verification Auditor

3 Discussion

Across the three studies, a coherent multi-dimensional reliability profile emerges for each model (Extended Data Fig. 4). Claude demonstrates consistent top-tier performance across all three domain types (formal 100%, semi-determinate 91.7%, indeterminate composite 93.6%), representing the only model with no identified systematic failure mode in any protocol across any domain. Grok demonstrates stability second only to Claude (formal 95.6%, indeterminate composite 89.6%). ChatGPT's reliability is memory-dependent: excellent when tasks are stateless, degraded when cross-session tracking is required.

Gemini presents the most theoretically significant profile: exceptional formula compliance (V3: 96.7%) and best-in-study point-prediction accuracy (V4 PE: 5.3%), combined with structural CI overconfidence (V4-CI: 29.7%). This dissociation suggests that Gemini's training has optimised for point-prediction accuracy without simultaneously learning appropriately calibrated uncertainty representation [2]. Self-knowledge calibration failures in large language models are independently documented: expressed model confidence systematically fails to track empirical accuracy across diverse tasks [5], and pre-trained transformer architectures exhibit characteristic overconfidence that is not automatically resolved by scaling or fine-tuning [6]. DeepSeek's persistent entropy formula variant—maintained despite explicit corrections across both V1 and V3—represents a training-induced formula prior that overrides prompt-level instruction; the instruction-following limitations of RLHF-tuned models have been extensively documented [7].

The calibration inversion identified in Study 3 has direct implications for high-stakes deployment. In clinical decision support, legal reasoning, and financial risk management, what matters is not merely the point prediction but the confidence interval—the acknowledged range of plausible outcomes [8–11]. Frontier large language models have recently been shown to approach human-level performance in structured forecasting tasks when provided with appropriate scaffolding [12], making appropriately calibrated uncertainty quantification—not just point-estimate accuracy—the critical deployment bottleneck. A clinical AI that correctly identifies the most likely diagnosis but provides a 1% CI when the true uncertainty is 40% is more dangerous than a slightly less accurate model with appropriately wide uncertainty bounds.

The specific mechanism of Gemini's calibration failure—systematically narrow CIs (average S&P CI width 98 pts vs. expected 500 pts)—is the indeterminate-domain analogue of the constraint aggregation failure identified in Study 1. This type of systematic overconfidence is well-documented in human probabilistic reasoning [13–16] and appears to manifest structurally in LLM uncertainty quantification. The Saturation Paradox extension from Study 1 is relevant here: just as diversity benefit peaked at the highest single-agent baseline in formal domains, calibration failure peaks in the most formula-accurate model in indeterminate domains (Gemini). High formal accuracy and high probabilistic calibration are not the same construct, and optimisation for one may structurally impede the other.

The data-freshness failures of DeepSeek and Perplexity in V4 establish information-access architecture as an independent reliability dimension, distinct from reasoning capability, formula fidelity, and uncertainty calibration. Training data contamination and knowledge-cutoff drift have been documented as systematic sources of error in closed-source large language models [17]. Retrieval-augmented generation architectures partially address data-freshness requirements through runtime document access [18], though empirical evidence demonstrates that LLM behavioural consistency degrades measurably across model versions even without explicit architectural changes [19]. The practical mechanism for GAAS would be a Data-Freshness Adversary: a component that independently queries current market data, weather APIs, or other live data sources and challenges the Generator's inputs before allowing synthesis. This architectural addition addresses Type II failures without requiring changes to the Generator or Auditor models. It is compatible with both same-model (S3-style) and diverse-model (S4-style) GAAS deployments.

Study 3 completes the planned three-study reliability framework. Study 1 [1] established that reliability in formal domains is an architectural problem: above ∼95% accuracy, error structure shifts from stochastic to systematic, and only role-separated verification (GAAS) breaks the ceiling. Study 2 (K.K.P., V.K.P. & A.P., under review) established that reliability in semi-determinate domains adds a calibration-first principle: Auditor role suitability depends on measurable self-audit calibration ($\rho_s \geq 0.70$ threshold). Study 3 establishes that reliability in indeterminate domains adds two further constraints: (1) a calibration-verification layer—an Auditor that explicitly tests CI width against expected volatility benchmarks—and (2) a data-freshness layer—a Data-Freshness Adversary that independently verifies input recency. These architectural principles are consistent with established results on the diversity advantage of heterogeneous ensembles [20–23] and the empirical wisdom-of-crowds literature [24]. Deep ensemble methods further demonstrate that predictive uncertainty quantification improves systematically with architectural diversity rather than model scaling alone [25].

Together, the three studies yield a unified GAAS architecture specification for real-world deployment: the Generator produces the primary answer and confidence bounds; the Auditor checks formula fidelity, logical consistency, and CI width calibration (requiring $\rho_s \geq 0.70$ measured in the target domain); the Adversary independently verifies data recency and generates counter-hypotheses; the Synthesizer integrates all signals into a final answer with calibrated uncertainty. This architecture is grounded in established principles of ensemble reasoning [26–28], chain-of-thought decomposition [29], and role-separated verification [1].

Limitations. Current limitations include: (1) per-protocol per-model session counts ($n = 12$ for V1, $n \approx 6$ for V3) are modest for individual-model inference, though these are embedded within the cumulative 2,730-point cross-protocol dataset spanning three structurally distinct protocols; the V4 protocol alone represents 252 prediction events across 6 models × 6 sessions × 7 variables (reduced to 212 after excluding market-closed dates), providing 212 independent calibration data points for the CI analysis; (2) T20 top-scorer name accuracy was 0% universally—partly reflecting intrinsic aleatoric uncertainty in individual match outcomes; (3) the V1→V3 improvement may partially reflect prompt learning rather than genuine capability improvement. The consumer-grade evaluation protocol—standard iOS applications under default settings—was a deliberate methodological choice, not a limitation: with 2.8 billion daily active AI users interacting through identical interfaces, ecological validity demands that reliability be characterised under precisely these conditions. Laboratory controls such as version pinning, temperature parameter access, and API-level token measurement are unavailable to the overwhelming majority of real-world users and would constitute an unrepresentative idealisation. Additionally, fundamental constraints on LLM symbolic reasoning and systematic planning capacity [30] represent a further boundary on generalisability.

4 Conclusion

This investigation evaluated AI reliability in fully indeterminate domains through 2,730 individual data points analysed across 142 controlled evaluation sessions, six frontier models, and three protocols spanning financial market forecasting, meteorology, sports, and cryptocurrency—the majority in indeterminate space where ground truth does not exist at prediction time. Four central findings emerge.

First, a calibration inversion in indeterminate domains: Gemini achieves the highest point-estimate accuracy (5.3% V4 PE, rank 1) while simultaneously achieving the second-lowest CI calibration (29.7%, rank 5), a dissociation structurally absent from formal and semi-determinate domains and representing a new failure mode not predictable from earlier studies.

Second, data-freshness failures as an architectural constraint: DeepSeek S&P 500 predictions are systematically stale (12–13% error, 3 of 4 scoreable sessions) and Perplexity predictions are frozen across sessions—failures attributable to information-access architecture rather than reasoning incapacity.

Third, cross-protocol Spearman rank consistency (mean $\rho_{\text{rank}} = 0.695$) confirms that model reliability rankings are moderately consistent across fundamentally different task structures, validating the three-study cumulative approach.

Fourth, Claude is the only model achieving top-tier performance across all three domain types and all three protocols (composite 93.6%), confirming its role as the multi-domain reliability standard.

The practical implication for the GAAS architectural framework is the addition of two new components for indeterminate domain deployment: a Calibration-Verification Auditor that explicitly tests CI width against expected market volatility benchmarks, and a Data-Freshness Adversary that independently verifies input recency. Together with the phase transition and self-audit calibration findings of Studies 1 and 2, these three studies yield a complete architectural specification for high-reliability AI deployment across the full spectrum of problem determinacy—from formally verifiable reasoning to real-world probabilistic prediction.

Declarations

Acknowledgements

This work received no dedicated funding from any granting body. The authors thank the anonymous reviewers for their comments.

Funding

This research received no external funding from any source—public, private, governmental, non-governmental, or institutional. All costs were borne entirely from the personal funds of the lead author. No in-kind support, infrastructure access, or resource allocation from any organisation, university, or corporate entity was utilised.

Competing interests

The authors declare no competing interests. The authors have no financial, employment, advisory, consulting, equity, or any other material relationship with Anthropic, OpenAI, xAI, DeepSeek AI, Google DeepMind, Perplexity AI, or any affiliated entity. JK Agri Genetics Ltd and Ericsson have no involvement in this research.

Data and code availability

All primary data are provided in Supplementary S3 accompanying this submission. Supplementary S3 contains: (i) all 72 V1 evaluation responses with metric-level scores across 6 models × 12 sessions; (ii) all 34 V3 evaluation responses; (iii) all 36 V4 evaluation responses with point estimates, 90% CIs, and CI hit scoring against ground truth; (iv) the complete V4 ground truth dataset; (v) formula reference tables; (vi) per-variable scoring rubrics; (vii) the inter-rater reliability dataset (Cohen's $\kappa = 0.84$). A permanent public repository will be deposited at OSF/Zenodo upon acceptance (DOI: to be assigned).

Ethics statement

This research involved the evaluation of publicly available AI systems through standard consumer interfaces and did not involve human subjects, animal subjects, personal data collection, or any activity requiring institutional ethics review.

Author contributions

K.K. Pandit: Conceptualisation; full methodology design for V1, V3, and V4 protocols; primary evaluation across all 142 evaluation sessions comprising 2,730 analysed data points; formal statistical analysis; data interpretation; original manuscript drafting; all revisions; project administration.

V.K. Pandit: Independent scoring of all V1/V3 metric classifications and V4 CI hit determinations as second evaluator; inter-rater reliability assessment (Cohen's $\kappa = 0.84$); methodology consultation; manuscript review and structural refinement.

A. Pandit: Prompt execution across all three evaluation protocols; systematic data cross-verification, tabulation, and cross-referencing of ground-truth outcomes against model predictions; participation in methodology and implementation discussions.

5 Methods

5.1 Protocol V1: Market bias and entropy metrics

The V1 protocol ("Phase Transition Reliability Engine — Protocol v1.0") presented each model with fixed baseline market data (S&P 500 price, ATR20) and required simulation of three distinct internal predictors: Predictor A (mean reversion), Predictor B (trend continuation), and a Counterfactual model (maximally divergent bias). Each predictor required a BIAS (Bullish/Bearish/Neutral), a numeric midpoint forecast, and a confidence level (0–1). Models then computed five derived metrics: (1) normalised Shannon entropy [31] ($H = -\sum p_i \cdot \ln(p_i)$, normalised by $\ln(3)$, rounded to 3 decimal places); (2) entropy deviation ($D = 1 - H_{\text{norm}}$); (3) entropy acceleration ($D_{\text{current}} - D_{\text{prior}}$, or null if no prior session available); (4) confidence-weighted forecast dispersion (weighted standard deviation of midpoints); and (5) regime consistency score (integer -3 to $+3$, based on alignment between predictor biases and declared regime).

A three-layer self-audit checklist was embedded in the prompt. Twelve sessions were conducted per model over 12 consecutive days (21 February–3 March 2026), yielding 72 evaluations.

The V1 prompt included an optional PREV ENTROPY DEVIATION parameter for acceleration computation, creating a session-memory requirement. Models with no cross-session memory (ChatGPT, DeepSeek, Perplexity) were expected to report null for acceleration, though some incorrectly reported 0 or fabricated values.

5.2 Protocol V3: Enhanced formula compliance

A second-generation protocol (V2) was designed as an intermediate step between V1 and V3; it was withdrawn before any evaluation sessions were conducted after a structural deficiency in the prompt formulation was identified that would have compromised cross-protocol comparability. V3 ("Phase Transition Reliability Engine — Protocol v3.1") is therefore the direct successor to V1 and extended it with four key modifications: (1) entropy computation upgraded to 4 decimal places; (2) entropy acceleration changed from a first derivative (ΔD) to a second derivative ($\Delta^2 D = D_{\text{current}} - 2 \cdot D_{\text{prior}} + D_{\text{prior,prior}}$), requiring the two preceding session values and rendering the first two sessions in any conversation as N/A by definition; (3) a fourth regime state (VOLATILE: ATR20 $>$ 120) added with revised scoring; (4) the Counterfactual predictor given a strictly defined rule—the CF predictor must model the scenario where the current regime classification is wrong, and NEUTRAL is not a valid bias when regime is BULLISH or BEARISH.

Step 0 of the prompt required explicit live market data fetch (SPX price, ATR20, 50-day moving average, date) before any computation, with DATA FRESHNESS field flagging CURRENT, STALE, or ESTIMATED. Six sessions per model were conducted (26 February–3 March 2026) for a total of 34 evaluations (ChatGPT completed 4 sessions rather than 6 due to session window constraints).

ChatGPT's session-numbering behaviour under V3 was anomalous: after the first run, subsequent sessions consistently reported SESSION NUMBER = 1 rather than incrementing, making second-derivative acceleration computation impossible (as it requires session numbers 3+). This is documented as a structural session-memory failure rather than a formula error.

5.3 Protocol V4: Real-world probabilistic CI prediction

The V4 protocol presented each model with seven independent real-world prediction tasks for the following calendar day, with explicit instruction to provide a point estimate and 90% confidence interval (CI) for each: (1) S&P 500 intraday high (USD); (2) Nifty 50 intraday low (INR); (3) Mumbai highest temperature (°C); (4) Tokyo lowest temperature (°C); (5) highest run scorer in the scheduled T20 World Cup match (player name and runs); (6) Bitcoin closing price (USD/BTC); (7) Gold lowest price (USD/oz). Models were instructed to use all available internet data, were granted a 100,000-token budget, and were required to provide only point estimates and CI bounds without explanatory text. Six sessions were conducted per model (26 February–3 March 2026), yielding 36 evaluations.

Ground truth was established from post-hoc market data. Market-closed days (28 February and 1 March for S&P 500 and Nifty; 3 March for Nifty) were excluded from CI scoring. T20 scoring was restricted to match days with completed results (27 February [Jacks, 32 runs], 28 February [Farhan, 100 runs], 1 March [Samson, 97 runs], 4 March [Finn Allen, 100*]). Two data quality flags were applied: DeepSeek Gold predictions in runs 1, 3, 5, 6 used gram-not-ounce units (approximately \$2,875–\$2,892 vs. actual ∼\$5,073–\$5,279/oz) and were scored as wrong; Perplexity Gold run 5 exhibited the same unit error.

5.4 Scoring protocols

V1/V3 metric accuracy. Each of the five metrics was scored as MATCH (full credit, 1.0), NEAR (partial credit, 0.5; within ±0.005 for entropy/deviation, ±10 for dispersion, or directionally correct for acceleration), or WRONG (0.0). Formula reference tables with exact computations for all canonical cases (all-different entropy = 1.000; two-same = 0.579; all-same = 0.000) were provided in the supplementary data and used as the scoring anchor. Overall V1/V3 accuracy is the mean of per-metric scores across all sessions.

V4 CI calibration. For each prediction, the binary question is whether the actual outcome fell within the stated 90% CI (score 1) or outside (score 0). CI hit rate = hits / scoreable predictions. A well-calibrated 90% CI should achieve approximately 90% hit rate over many predictions. Point estimate accuracy is reported as absolute percentage error: $|\text{prediction} - \text{actual}|/\text{actual} \times 100$. Grading thresholds differ by variable type (financial indices: Excellent $\leq 0.5\%$, Good $\leq 1.5\%$, Acceptable $\leq 3\%$; crypto/gold: Excellent $\leq 2\%$, Good $\leq 5\%$, Acceptable $\leq 10\%$; temperature: Excellent ≤ 1°C, Good ≤ 2°C; T20 runs: Excellent $\leq 15\%$, Good $\leq 30\%$).

Inter-rater reliability. All V1/V3 metric classifications and V4 CI scoring decisions were independently reviewed by the second evaluator (V.K. Pandit) without access to the lead author's scores. Twelve items required consensus discussion (eight at V1 MATCH/NEAR boundaries; four at V4 CI borderline cases involving rounding). Cohen's $\kappa = 0.84$ (95% CI: 0.77–0.91), classified as almost perfect agreement [32], confirming reliable application of the scoring rubric across both metric accuracy and CI calibration dimensions.

5.5 Model selection and configuration

Six frontier language models were evaluated, identical to Studies 1 and 2 (Extended Data Table 3), ensuring direct cross-study comparability. All models were evaluated via consumer iOS applications as available in February–March 2026, under default settings. Three paid-tier models (Claude Sonnet 4.6 [33], ChatGPT GPT-5.2 [34], Grok 4/4.1) and three free-tier models (DeepSeek V2.5 [35], Gemini 3 Flash/3.1 Pro [36], Perplexity v2.260206.1) were included.

5.6 Statistical framework

V1/V3 accuracy. Overall accuracy was computed as the mean of per-metric scores per model across all sessions. Wilson score 95% CIs preferred over normal-approximation at proportions near 0 or 1 [37, 38]. Cross-protocol Spearman rank correlation (ρ_{rank}) computed for all three cross-protocol pairs across the six models; $\rho_{\text{rank}} = 1.0$ indicates perfect rank-order consistency, $\rho_{\text{rank}} = 0$ no consistency. Composite score = arithmetic mean of V1 accuracy, V3 accuracy, and V4 CI hit rate.

V4 CI calibration. Expected hit rate for a true 90% CI is 90% [39, 40]. Deviations expressed as percentage-point differences from 90% (for example, Claude: 84% → −6 pp). CI width analysis compares average CI width against expected market volatility benchmarks (S&P 500: ±250 pts for 90% CI; Nifty: ±600 pts; BTC: ±$8,000; Gold: ±$300; temperature: ±4°C). Models classified as CI-appropriate (0.7–2× expected width), CI-narrow (0.3–0.7×), or severely CI-narrow (< 0.3×).

V3 protocol compliance. Six additional binary compliance metrics scored for V3 runs—CF compliance (correct bias based on regime), session-number integrity (session number increments correctly), and self-audit completion—reported separately from metric accuracy.

References

[1] Pandit, K.K., Pandit, V.K. & Pandit, A. Architectural phase transition governs AI reliability beyond the single-agent ceiling: Evidence from 4,680 controlled evaluations. Preprint at `https://doi.org/10.17605/OSF.IO/GJEN8` (2026).

[2] Guo, C., Pleiss, G., Sun, Y. & Weinberger, K.Q. On calibration of modern neural networks. *Proc. 34th International Conference on Machine Learning (ICML)*, 1321–1330 (2017).

[3] Lopez-Lira, A. & Tang, Y. Can ChatGPT forecast stock price movements? Return predictability and large language models. *SSRN Working Paper 4412788* (2023).

[4] Cemri, M. et al. Why do multi-agent LLM systems fail? Preprint at `https://arxiv.org/abs/2503.13657` (2025).

[5] Kadavath, S. et al. Language models (mostly) know what they know. Preprint at `https://arxiv.org/abs/2207.05221` (2022).

[6] Desai, S. & Durrett, G. Calibration of pre-trained transformers. *Proc. EMNLP*, 295–302 (2020).

[7] Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35** (NeurIPS), 27730–27744 (2022).

[8] Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* **25**, 44–56 (2019).

[9] Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E.J. AI in health and medicine. *Nat. Med.* **28**, 31–38 (2022).

[10] Obermeyer, Z. & Emanuel, E.J. Predicting the future—big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **375**, 1216–1219 (2016).

[11] Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at `https://arxiv.org/abs/2108.07258` (2021).

[12] Halawi, D. et al. Approaching human-level forecasting with language models. *Adv. Neural Inf. Process. Syst.* **37** (NeurIPS) (2024).

[13] Tversky, A. & Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science* **185**, 1124–1131 (1974).

[14] Gigerenzer, G. & Hoffrage, U. How to improve Bayesian reasoning without instruction: frequency formats. *Psychol. Rev.* **102**, 684–704 (1995).

[15] Tetlock, P.E. & Gardner, D. *Superforecasting: The Art and Science of Prediction* (Crown, 2015).

[16] Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. *Proc. ICML*, 625–632 (2005).

[17] Balloccu, S. et al. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. Preprint at `https://arxiv.org/abs/2402.03927` (2024).

[18] Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv. Neural Inf. Process. Syst.* **33** (NeurIPS), 9459–9474 (2020).

[19] Chen, L. et al. How is ChatGPT's behavior changing over time? Preprint at `https://arxiv.org/abs/2307.09009` (2023).

[20] Hong, L. & Page, S.E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl Acad. Sci. USA* **101**, 16385–16389 (2004).

[21] Condorcet, M. de. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix* (Imprimerie Royale, 1785).

[22] Dietterich, T.G. Ensemble methods in machine learning. *Lect. Notes Comput. Sci.* **1857**, 1–15 (Springer, 2000).

[23] Breiman, L. Bagging predictors. *Mach. Learn.* **24**, 123–140 (1996).

[24] Steyvers, M. et al. The wisdom of the inner crowd in three large natural experiments. *Psychol. Sci.* **33**, 1417–1428 (2022).

[25] Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* **30** (NeurIPS) (2017).

[26] Wang, X. et al. Self-consistency improves chain of thought reasoning in large language models. *Int. Conf. Learning Representations (ICLR)* (2023).

[27] Du, Y. et al. Improving factuality and reasoning in language models through multiagent debate. Preprint at `https://arxiv.org/abs/2305.14325` (2023).

[28] Liang, T. et al. Encouraging divergent thinking in large language models through multi-agent debate. Preprint at `https://arxiv.org/abs/2305.19118` (2023).

[29] Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **35** (NeurIPS), 24824–24837 (2022).

[30] Kambhampati, S. Can large language models reason and plan? *Ann. N.Y. Acad. Sci.* **1534**, 15–18 (2024).

[31] Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).

[32] Landis, J.R. & Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).

[33] Anthropic. Claude Sonnet 4.6 System Card. Technical Report (Anthropic, 2025). (`https://www.anthropic.com/research/claude-sonnet-4-6-system-card`)

[34] Achiam, J. et al. GPT-4 technical report. Preprint at `https://arxiv.org/abs/2303.08774` (2024).

[35] DeepSeek AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. Preprint at `https://arxiv.org/abs/2405.04434` (2024).

[36] Google DeepMind. Gemini: A family of highly capable multimodal models. Preprint at `https://arxiv.org/abs/2312.11805` (2023).

[37] Wilson, E.B. Probable inference, the law of succession, and statistical inference. *J. Am. Stat. Assoc.* **22**, 209–212 (1927).

[38] Brown, L., Cai, T. & DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* **16**, 101–133 (2001).

[39] Gneiting, T. & Raftery, A.E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).

[40] Brier, G.W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950).