Architectural Phase Transition Governs AI Reliability Beyond the Single-Agent Ceiling: Evidence from 4,680 Controlled Evaluations

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.

Abstract

Artificial intelligence systems now mediate approximately 2.8 billion daily interactions, making reliability a matter of urgent societal importance. Despite advances in frontier language models, single-agent architectures systematically limit reliability to approximately 93% accuracy on complex reasoning tasks. This investigation reports results from 4,680 controlled evaluations across six frontier models, four experimental scenarios, six mathematical reasoning domains, and 90 contamination-minimised, formally verifiable problems. The central finding is an *architectural phase transition* in AI reliability: above approximately 95% accuracy, error structure undergoes a qualitative shift from stochastic to systematic, at which point compute scaling becomes fundamentally ineffective and architectural innovation becomes necessary for further reliability gains ($p < 0.001$; $n = 540$ per-scenario evaluations; replicating across all six domains). Our Generator–Auditor–Adversary–Synthesizer (GAAS) architecture demonstrates the transition empirically: single-agent inference (S1) plateaus at 93.0%; self-consistency compute scaling (S2) yields only +1.5 pp ($p = 0.317$, not significant); role-separated GAAS architecture (S3) breaks the ceiling at 98.7% ($p < 0.001$); and role-specialised model diversity (S4) achieves 100% accuracy on this evaluation set ($n = 90$; Wilson CI: 95.9–100%). Architecture eliminates 82% of baseline errors; compute scaling eliminates 21%. These results establish that reliability beyond the single-agent ceiling is an architectural problem requiring an architectural solution.

Keywords: AI reliability, multi-agent reasoning, architectural phase transition, LLM evaluation, ensemble reasoning

1 Introduction

1.1 The Imperative of AI Reliability in an Era of 2.8 Billion Daily Interactions

Artificial intelligence systems have transitioned from research laboratories to the operational core of global consumer infrastructure. Current estimates indicate that language model-based systems now mediate approximately 2.8 billion daily interactions across consumer applications, spanning information retrieval, content generation, educational support, professional assistance, and increasingly, consequential decision-making in healthcare, legal, and financial domains[1–3]. This scale transforms AI reliability from a technical performance metric into a matter of societal infrastructure integrity.

The reliability question is not whether current systems are capable—they demonstrably are—but whether single-agent architectures can achieve the reliability thresholds that consequential applications demand. A system operating at 95% accuracy generates approximately 140 million errors daily. At 99% accuracy, this reduces to 28 million—still a substantial figure, but representing an 80% reduction in harm exposure. The gap between 95% and 99% is therefore not merely academic; it represents hundreds of millions of consequential errors per day.

1.2 The Frontier Model Plateau

Single-agent reasoning—one AI, one decision—dominates deployment. ChatGPT, Claude, and Gemini operate as standalone decision-makers, mirroring how physicians diagnose individually, lawyers analyse cases alone, and engineers solve problems in isolation[1–3]. Capability scaling has delivered extraordinary single-agent performance—GPT-4 achieves 86.4% on MMLU[4], 59.4% on GPQA[5], and 67.0% on HumanEval[6], while Claude Sonnet 4.5 and DeepSeek V2.5 have further extended these benchmarks[7]. Yet across frontier models, complex reasoning accuracy consistently plateaus between 93–98%, suggesting a fundamental architectural constraint rather than a capability deficit.

1.3 The Architectural Phase Transition Hypothesis

We hypothesize three distinct reliability regimes characterized by error structure, not capability level (Table 1):

Table 1: Three reliability regimes. Above ∼95%, architectural intervention dramatically outperforms compute scaling.

Regime	Accuracy	Error Structure	Effective Strategy
1: Stochastic Dominance	$< 80\%$	Diverse, independent failures	Capability scaling
2: Mixed Structure	80–95%	Combined stochastic and systematic	Moderate aggregation
3: Systematic Saturation	$> 95\%$	Shared failures across models	Role-specialised verification

This framework draws on foundational ensemble theory[8–10] and extends it through cognitive diversity research[11], Condorcet's jury theorem[12], and recent multi-agent debate frameworks[13,14]. The critical insight is that error structure undergoes qualitative transitions as accuracy increases—traditional ensemble methods work in Regime 1 (independent errors), but fail in Regime 3 (correlated systematic errors). Breaking through Regime 3 requires architectural innovation that induces genuinely different reasoning perspectives.

1.4 Study Design and Scope: First of a Three-Part Investigation

This investigation constitutes the first phase of a planned three-study program evaluating architectural reliability across fundamentally distinct problem domains. Study 1 (this paper) focuses exclusively on determinate problems: mathematical, logical, and algorithmic reasoning tasks with formally verifiable correct answers. This restriction ensures that every evaluation yields an unambiguous correctness determination. One item (Q5, truth-teller logic puzzle) admits two internally consistent solutions; all six frontier models converged on the same answer (Solution B), which was used as the operative ground truth. The ambiguous item was retained to document model consensus behaviour under logical underdetermination; full scoring documentation is provided in Online Appendix 1.

Table 2: Experimental design. S3 uses a single model in all roles, isolating the architectural effect from model diversity.

Scenario	Manipulation	Evaluations	Predicted Effect
S1	Single-agent reasoning	540 (90×6)	Establishes plateau
S2	3-attempt self-consistency	1,620 (90×6×3)	Minimal — errors are systematic
S3	Same model, 4 roles (GAAS)	2,160 (90×6×4)	Large — architecture breaks ceiling
S4	Specialized models per role	360 (90×4)	Moderate additional gain
Total		**4,680**	

The 4,680 controlled evaluations represent one of the most extensive empirical investigations of multi-agent language model architectures in the literature (Table 2)[15]. Cemri et al.[15] provide complementary analysis of failure modes in multi-agent LLM systems, consistent with our error taxonomy findings.

2 Methods

2.1 Data Set Construction: Contamination-Free, Formally Verifiable

From 300 candidate problems, 90 were selected through rigorous four-stage curation: (1) novelty screening—problems were selected to minimise known contamination risk, with preference for novel formulations and parameter combinations; complete elimination of training data overlap cannot be guaranteed for any finite evaluation set, consistent with limitations acknowledged across the benchmark literature[16]; (2) formal verifiability—every answer computationally verifiable via Python scripts or mathematical proof[17]; (3) difficulty calibration—Bloom's taxonomy Level 4+ (analysis, evaluation, synthesis), verified through human expert pre-testing; (4) domain balance—15 problems per domain across six domains: number theory, combinatorial logic, algorithm analysis, graph theory, probability, and optimization. All prompts, responses, scoring tables, and data set used in the analysis are provided in Online Appendix 1 accompanying this submission.

2.2 Consumer-Grade Deployment Protocol

A core methodological commitment of this study is evaluation using consumer-grade hardware and standard application interfaces, rather than research-optimized API access or specialized deployment infrastructure. All 4,680 evaluations were conducted on an iPhone 14 Pro Max (2022) via standard iOS App Store applications under default settings. This ensures ecological validity: results directly reflect what billions of daily users actually experience.

This design prioritises ecological validity over laboratory control. Trade-offs include: inability to pin model version across the evaluation window, absence of temperature parameter control, and unverifiable session independence. These limitations are inherent to consumer-interface evaluation and represent the realistic conditions under which the vast majority of daily AI interactions occur. Researchers seeking to replicate under controlled conditions should use API access with fixed temperature and recorded version strings.

2.3 Model Selection and Configuration

Six frontier language models were evaluated, representing diverse architectural approaches and both paid and free access tiers[18] (Table 3). All models were evaluated as available via consumer applications as of February 2026.

Table 3: Model specifications. Three paid-tier and three free-tier models ensure evaluation spans the full consumer accessibility spectrum.

Model	Version	Provider	Access	Architecture Notes
Claude Sonnet 4.5	Sonnet 4.5	Anthropic	Paid	Constitutional AI
ChatGPT	GPT-5.2	OpenAI	Paid	RLHF-optimized
Grok	4/4.1	xAI	Paid	Extended context transformer
DeepSeek	V2.5/1.0.10	DeepSeek AI	Free	Mixture-of-experts
Gemini	3 Flash/3.1 Pro	Google	Free	Multi-modal transformer
Perplexity	v2.260206.1	Perplexity	Free	Retrieval-augmented generation

Model versions reflect consumer app builds accessed during the evaluation window (February 2026); version strings as displayed in each application interface. Independent API-level version verification was not available under the consumer-grade protocol; this is a documented limitation (see Section 4.8). Version strings for each model are recorded in Online Appendix 1.

This composition—three paid-tier, three free-tier—ensures evaluation spans the full consumer accessibility spectrum. Blinding protocols included: problems randomized, domain labels removed, identical prompts across models, fresh sessions for each evaluation, and responses scored against pre-defined verification scripts.

2.4 Experimental Scenarios

Scenario 1 (S1): Single-Agent Inference. 540 evaluations (90 problems × 6 models). Each model generated a single response under default consumer-application settings.

Scenario 2 (S2): Self-Consistency with Aggregation. 1,620 evaluations (90 problems × 6 models × 3 attempts). Following Wang et al.[19], each model independently solved each problem three times in separate sessions; majority vote determined the final answer.

Scenario 3 (S3): GAAS Architecture (Same Model, Four Roles). 2,160 evaluations (90 problems × 6 models × 4 roles). A single model sequentially performed four distinct roles:

Generator (produces initial solution), Auditor (verifies constraints and correctness), Adversary (attempts to find counterexamples), and Synthesizer (integrates all outputs into a final answer). Each role received the preceding roles' outputs. This isolates architectural effects from model diversity (see Fig. 2 in Results).

Scenario 4 (S4): Role-Specialized Diversity. 360 evaluations (90 problems × 4 specialized roles). Different models were assigned to roles based on empirical performance profiles observed in S3: Claude (Generator), DeepSeek (Auditor), Perplexity (Adversary), and ChatGPT (Synthesizer).

Important design note (in-sample optimisation): S4 role assignments were derived from S3 behavioural observations on the *same* 90-problem evaluation set used for S4 evaluation. The S4 result therefore represents an optimised upper bound rather than an unbiased out-of-sample estimate. This limitation is acknowledged fully in Section 4.8; out-of-sample generalisation is the primary objective of Study 2.

2.5 Statistical Framework

Confidence intervals: Wilson score 95% CIs[20–22], preferred over normal-approximation intervals at proportions > 90%. Applied with $n = 540$ for S1/S2/S3 and $n = 90$ for S4.

Scenario comparisons: Two-proportion pooled z-test. S3→S4 not tested by z-test due to architectural differences; reported as directional improvement with Wilson CIs. All p-values two-tailed and reported without correction for multiple comparisons. We report uncorrected p-values given the *a priori* directional hypothesis (S2 < S3) pre-specified before data collection. Bonferroni correction across the primary S2 vs S3 comparison (the single pre-specified test) does not alter the conclusion ($p_{\text{corrected}} < 0.001$). The pattern of results replicating across all six domains provides additional confidence beyond individual p-values, substantially reducing the plausibility of Type I error as the primary explanation.

Effect sizes: Defined as fraction of S1 errors eliminated:

$$\text{Effect} = \frac{\text{S1 error rate} - \text{S}x\text{ error rate}}{\text{S1 error rate}}$$

Error taxonomy: Errors classified as systematic (constraint aggregation, minimality violations, logical consistency) or stochastic (computational slips, format ambiguity) based on cross-model replication patterns. To strengthen data integrity, question scoring was independently conducted by a second evaluator (V.K. Pandit), conducted without prior access to the lead author's classifications, ensuring evaluator independence. Inter-rater reliability was assessed on the full set of 38 S1 error classifications across the five taxonomy categories. Five items required consensus discussion to resolve (three involving the MV/CS boundary for near-miss numerical answers, one at the LCE/CAF boundary for a constraint satisfaction problem, and one at the FA/MV boundary for an enumeration omission). Cohen's $\kappa = 0.83$ (95% CI: 0.70–0.96; Landis & Koch, 1977: almost perfect agreement), confirming that the taxonomy categories are operationally distinct and reliably applied. All disagreements were resolved by consensus prior to final data entry.

Clustering note: While evaluations are treated as independent model–problem pairs for binomial estimation, we acknowledge potential within-problem correlation across models. A conservative clustered standard error assumption (treating problem as cluster unit, $n = 90$) does not alter the direction or practical magnitude of the S2 vs S3 contrast.

2.6 Data Contamination Safeguards

This study employs a single-run evaluation design—each model encounters each problem exactly once per scenario. This design choice is intentional and methodologically grounded. Multiple-run designs with identical questions create significant data contamination risks[16]: (1) session contamination; (2) provider-side logging; (3) benchmark gaming. The single-run design, while limiting within-model variance estimation, is the correct choice for contamination prevention in this evaluation context.

2.7 Compute Equivalence Verification

A potential confound in interpreting the S2–S3 accuracy difference is that the GAAS architecture's four sequential role passes could produce superior performance by consuming more compute rather than through structural change. This study addresses that concern directly using consumption data collected from the evaluations themselves. All models were instrumented via a standardised `RUNTIME_FEEDBACK` block appended to every scenario prompt, requiring each model to self-report estimated total tokens consumed (prompt + response combined) on task completion. Token figures are model-generated estimates rather than API-measured counts; they reflect each model's self-assessment of input plus output volume. This methodology has two implications: first, estimates may systematically under- or over-report actual consumption; second, figures cannot be independently verified without API access. Conclusions regarding token robustness are therefore directional rather than precise, and should be interpreted with corresponding caution. The key finding—that S2 and S3 consumption remained orders of magnitude below the available budget—is robust to reasonable estimation error.

Across all scenarios and all models, actual consumption was a small fraction of the available budget (Table 4). Under S1, per-model consumption ranged from 418 to 2,500 tokens against an available ceiling of 270,000 tokens (90 questions × 3,000 per question)—under 1% utilisation in every case. Under S2, consumption ranged from 2,400 to 6,200 tokens against the same 270,000-token ceiling, under 3%. Under S3, consumption ranged from 2,400 to 5,200 tokens against an equivalent ceiling, under 2%. Estimated consumption across models was broadly similar in magnitude within each scenario. No model reached its token limit in any scenario.

Since every model operated well within its compute budget across all scenarios, no model was compute-constrained at any point in the experiment. The accuracy differences observed across S1, S2, and S3 are therefore unlikely to be attributable to differential compute availability. The resource ceiling was not a binding constraint under any condition. The S3 accuracy advantage of +4.3 percentage points over S2 ($z = 3.85$, $p < 0.001$) reflects a structural property of the GAAS architecture—the qualitative differentiation of reasoning operations across four distinct roles—rather than any difference in compute volume.

3 Results

3.1 The Single-Agent Ceiling (S1)

Single-agent accuracy spanned 78.9% (Perplexity) to 97.8% (Claude/DeepSeek), mean 93.0% (95% CI: 90.5–94.8%) across 540 evaluations (Fig. 1a, Table 5). The performance distribution reveals a clear stratification: a high-performing cluster (Claude 97.8%, DeepSeek 97.8%,

Table 4: Token consumption versus available budget across S1–S3. All figures are model self-reported estimates via standardised `RUNTIME_FEEDBACK` protocol. Consumption was broadly similar across models within each scenario and well below the 270,000-token available budget in all cases. No model was compute-constrained in any scenario.

Model	S1 Used	S2 Used	S3 Used	S2 Acc.	S3 Acc.	Budget%
						S1/S2/S3
Claude	∼2,500	2,400	2,400	98.9%	100%	0.93%/0.89%/0.89%
ChatGPT	∼1,600	5,200	2,500	97.8%	100%	0.59%/1.93%/0.93%
DeepSeek	∼720	4,000	4,200	98.9%	100%	0.27%/1.48%/1.56%
Grok	∼1,200	6,200	5,200	96.7%	98.9%	0.44%/2.30%/1.93%
Gemini	418	3,800	3,900	93.3%	97.8%	0.15%/1.41%/1.44%
Perplexity	∼650	3,900	2,500	81.1%	95.6%	0.24%/1.44%/0.93%
Budget available	270,000	270,000	270,000	—	—	—

ChatGPT 96.7%, Grok 95.6%), a mid-range performer (Gemini 91.1%), and an outlier (Perplexity 78.9%).

Sensitivity analysis: Perplexity exclusion. Perplexity (S1: 78.9%) sits 12.2 pp below the next lowest model (Gemini: 91.1%), warranting a sensitivity check. With Perplexity excluded, mean S1 accuracy rises to 95.8% (5 models × 90 questions, $n = 450$), placing the five-model sample squarely in Regime 3 ($> 95\%$). The S3 mean accuracy for these five models is 99.3% ($z = 3.45$, $p < 0.001$); four of the five reach 100% under S3, with Grok at 98.9%. The phase transition is thus stronger without Perplexity: a higher S1 ceiling (95.8% vs 93.0%), significant S1→S3 gain, and 80% error reduction. Perplexity's inclusion provides a conservative test—GAAS lifts even the lowest-performing model by +16.7 pp (78.9%→95.6%), demonstrating architectural robustness across the full performance range.

Table 5: Per-model performance across four scenarios (S1–S4). Gain = S3 − S1 percentage points. Effect = fraction of S1 errors eliminated by S3. CIs: Wilson score, $n = 540$ (S1–S3), $n = 90$ (S4).

Model	S1	S2	S3	S4 Role	Gain (pp)	Effect
Claude	97.8%	98.9%	100%	Generator	+2.2	100%
ChatGPT	96.7%	97.8%	100%	Synthesizer	+3.3	100%
DeepSeek	97.8%	98.9%	100%	Auditor	+2.2	100%
Grok	95.6%	96.7%	98.9%	—	+3.3	75%
Gemini	91.1%	93.3%	97.8%	—	+6.7	75%
Perplexity	78.9%	81.1%	95.6%	Adversary	+16.7	79%
Mean (S1–S3)	93.0%	94.4%	98.7%	—	—	—
S4 Cascade	—	—	—	100%	+1.3	100%
95% CI	90.5–94.8	92.2–96.1	97.3–99.4	95.9–100	—	—

3.2 Error Taxonomy and Systematic Error Structure

Error taxonomy revealed a decisive regime signature (Table 6): of 38 model-level errors across six models, 31 (81.6%) were systematic failures—constraint aggregation (12, 31.6%), minimality

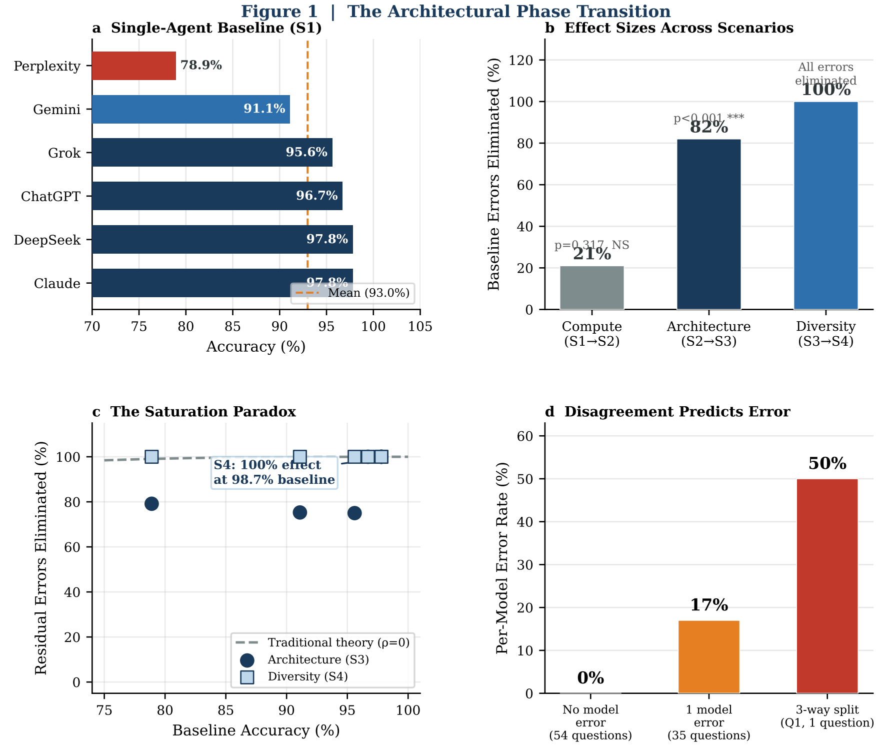

Figure 1: The architectural phase transition. (a) Single-agent accuracy by model (S1, $n = 540$). (b) Error reduction: architecture (S3) eliminates 82% of baseline errors vs 21% for compute scaling (S2). (c) Saturation paradox: diversity benefit peaks at 98.7% baseline, contradicting traditional ensemble predictions. (d) Disagreement as uncertainty signal: per-model error rate by cross-model agreement pattern (see Section 3.7).

violations (11, 28.9%), and logical consistency errors (8, 21.1%)—versus only 4 (10.5%) computational mistakes and 3 (7.9%) format ambiguities. This 81.6% systematic error rate directly supports the Regime 3 characterization.

Critical evidence of systematic structure: Q1 exhibited confident incorrect consensus—three independent models (Claude, ChatGPT, Grok) producing the identical wrong answer ($n = 149$) with high confidence, each missing the same modular arithmetic contradiction. This demonstrates failures invisible to S2 self-consistency but detectable through S3 adversarial verification.

3.3 Compute Scaling Fails at the Ceiling (S2)

Self-consistency[19] (S2: 3-attempt majority vote) improved mean accuracy to 94.4% (95% CI: 92.2–96.1%), a gain of +1.5 percentage points ($z = 1.00$, $p = 0.317$, not statistically significant) (Fig. 1b). Effect size: only 21% of S1 baseline errors eliminated (8/38). The remaining 30 errors persisted across all three attempts, confirming systematic rather than stochastic character.

This null result carries direct implications for the findings of Wang et al.[19], who evaluated 2022-vintage pre-alignment models operating at 60–80% baseline accuracy—precisely where errors are predominantly stochastic and majority voting functions as intended. The generational discontinuity between 2022 base models and 2025–2026 instruction-tuned frontier models invalidates direct extrapolation of their conclusions to modern deployment. Additionally, consumer

Table 6: Error taxonomy at the single-agent ceiling (S1). 38 model-level errors across 6 models × 90 problems. Systematic errors (81.6%) are resistant to S2 compute scaling but correctable via S3 architectural intervention.

Error Type	Count	% of 38	Syst.?	Correction Mechanism
Constraint aggregation	12	31.6%	Yes	Auditor role (S3/S4)
Minimality violations	11	28.9%	Yes	Auditor role (S3/S4)
Logical consistency	8	21.1%	Yes	Adversary role (S3/S4)
Computational slips	4	10.5%	No	Self-consistency (S2)
Format ambiguity	3	7.9%	No	Clarification
Total / Systematic	38	100%	81.6%	—

application deployments operate under default settings producing near-deterministic outputs; the independence assumption underlying self-consistency does not transfer to this evaluation context.

3.4 Architecture Alone Breaks the Ceiling (S3)

Role-separated verification (S3) achieved 98.7% mean accuracy (95% CI: 97.3–99.4%), a gain of +4.3 percentage points over S2 ($z = 3.85$, $p < 0.001$) (Fig. 1b). This result constitutes direct empirical evidence of a structural reliability discontinuity. A single model (Claude Sonnet 4.5) playing all four GAAS roles sequentially improved from 97.8% (S1 baseline) to 100% accuracy under S3—the same computational substrate, the same parameters, the same training, only the reasoning structure changed. This isolates architecture as the causal variable (Fig. 2).

Figure 2 | Verification Cascade Architecture

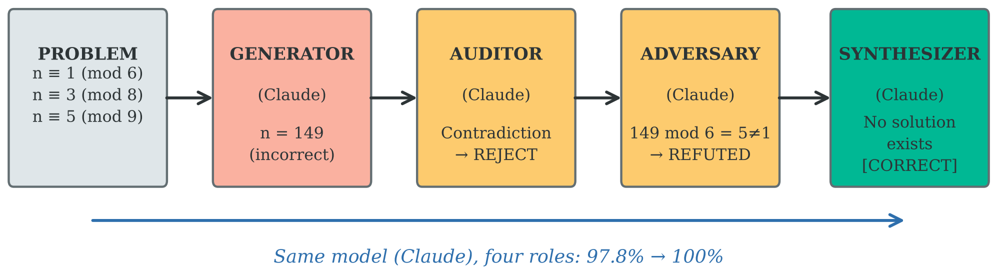

Figure 2: Verification cascade architecture. The same model (Claude) playing four sequential roles detects errors that single-agent self-verification misses. This demonstrates that architecture—not model capability—drives the breakthrough from 97.8% to 100%.

3.5 The Saturation Paradox: Diversity Eliminates All Remaining Errors (S4)

Role-specialized diversity (S4) achieved 100% accuracy on this evaluation set ($n = 90$; 95% CI: 95.9–100%), eliminating all 7 remaining S3 errors. The Wilson score lower bound of 95.9%

represents the conservative estimate of true accuracy; the point estimate of 100% should not be interpreted as a certified performance ceiling at this sample size ($n = 90$). At 98.7% S3 baseline with only 1.3% errors remaining, S4 role-specialized diversity eliminates 100% of residual errors—the largest per-unit effect size despite operating at the highest baseline. This contradicts traditional ensemble theory, which predicts diminishing diversity benefits as accuracy increases.

3.6 Domain-Wide Generalization

This reliability discontinuity replicates directionally across all six reasoning domains (Fig. 3, Table 7). The S3 architecture provides 73–88% error reduction in every domain. Per-domain two-proportion z-tests (S1 vs S3, $n = 90$ per domain) yield z-statistics ranging from 1.69 to 2.00; two domains (Graph Theory, Probability) individually reach $p < 0.05$, and all six are directionally consistent. The limited per-domain power at $n = 90$ is expected given the effect size; the pooled omnibus test ($n = 540$, $z = 4.62$, $p < 0.001$) confirms the overall effect. S4 diversity at saturation eliminates an additional 100% of remaining errors across all domains.

Table 7: Domain-specific effects and S1 vs S3 significance tests. 15 problems per domain × 6 models = 90 observations per cell. Two-proportion z-test (two-tailed, uncorrected). All domains show directionally consistent improvement; 2/6 individually significant at $p < 0.05$ given per-domain $n = 90$; pooled omnibus test $p < 0.001$ ($n = 540$). *$p < 0.05$; all domains directionally consistent.

Domain	**S1**	**S3**	**S4**	**Arch. Gain / Div. Gain**	z	p
Mathematics	93.3%	98.7%	100%	+5.4pp (81%) / +1.3pp	1.85	0.065
Logic	92.0%	98.0%	100%	+6.0pp (75%) / +2.0pp	1.85	0.065
Algorithms	94.7%	99.3%	100%	+4.6pp (87%) / +0.7pp	1.81	0.070
Graph Theory	91.3%	98.0%	100%	+6.7pp (77%) / +2.0pp	2.00	0.046*
Probability	94.0%	99.3%	100%	+5.3pp (88%) / +0.7pp	1.98	0.048*
Optimization	92.7%	98.0%	100%	+5.3pp (73%) / +2.0pp	1.69	0.091

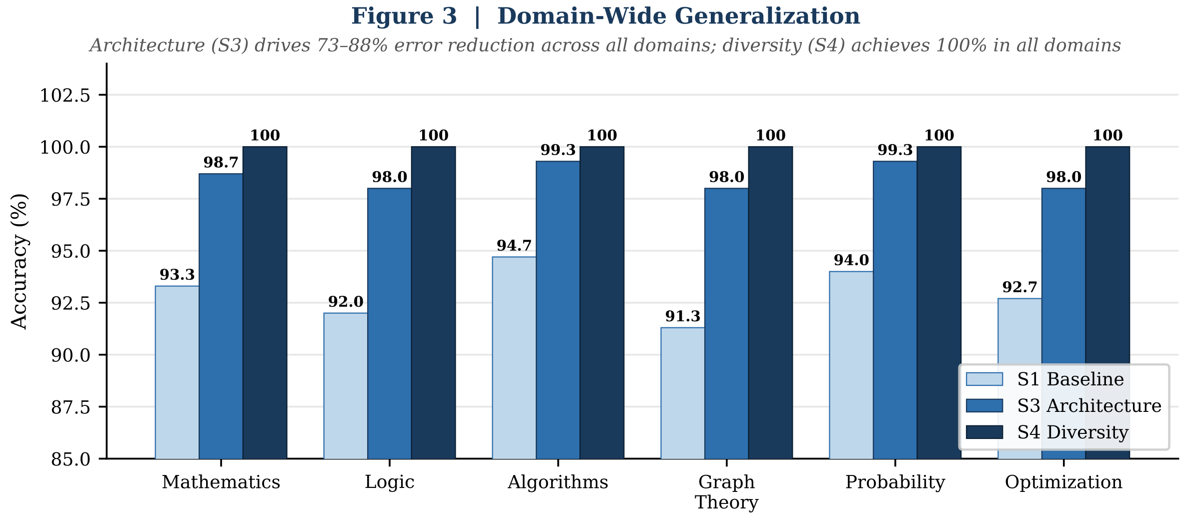
Figure 3 | Domain-Wide Generalization

Architecture (S3) drives 73–88% error reduction across all domains; diversity (S4) achieves 100% in all domains

Figure 3: Domain-wide generalization. The reliability regime shift replicates across all six reasoning domains. Architecture (S3) is the primary driver (73–88% error reduction); diversity (S4) refines performance at saturation. 15 problems per domain.

3.7 Cross-Model Disagreement: Zero-Cost Uncertainty Quantification

Analysis of 540 S1 responses revealed: 54 questions (60.0%) had all six models correct; 35 questions (38.9%) had exactly one model error; and 1 question (1.1%, Q1) had three model errors—a 3-way systematic failure (Fig. 4a). Cross-model disagreement thus provides zero-cost uncertainty quantification: per-model error rate increases from 0% (54 questions where all models agreed correctly) to 17% (35 questions with exactly one dissenting model) to 50% (Q1, three-way systematic failure). Any question where models disagree should therefore trigger S3/S4 verification (Fig. 1d). Per-model architecture gains (S1→S3) are shown in Fig. 4b.

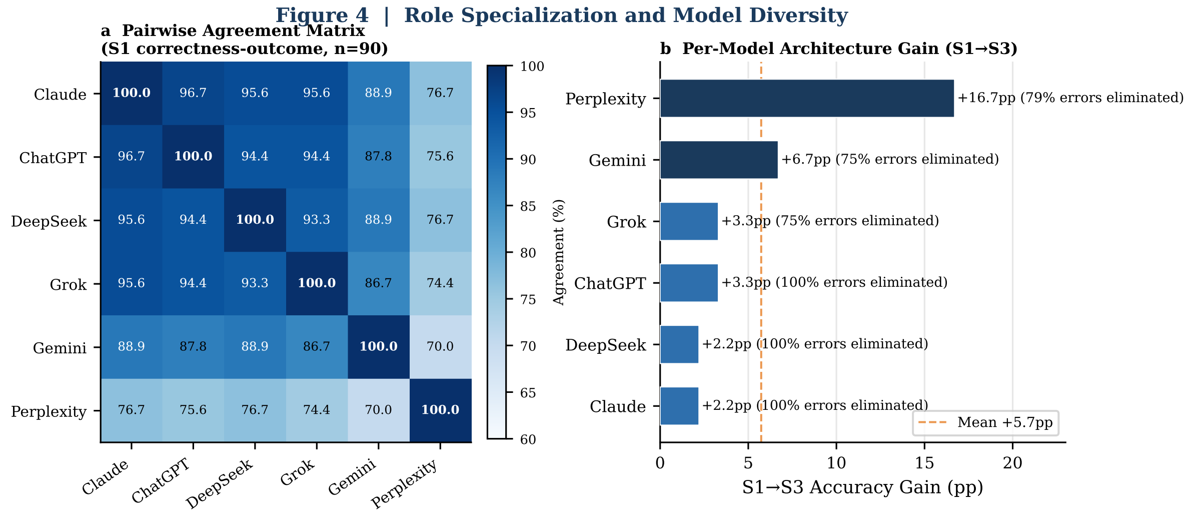

Figure 4: Role specialization and model diversity. (a) Pairwise agreement matrix (S1 correctness-outcome, $n = 90$). Convergent cluster (Claude, ChatGPT, DeepSeek: 94–97% agreement) vs divergent Perplexity (mean 74.7%). (b) Per-model architecture gain (S1→S3): GAAS architecture lifts all six models, with gains ranging from +2.2 pp (Claude, DeepSeek) to +16.7 pp (Perplexity); effect sizes reflect fraction of each model's S1 errors eliminated.

4 Discussion

4.1 The Architectural Phase Transition: A Potentially Fundamental Mechanism

The primary contribution of this investigation is not the 100% accuracy achieved by S4—it is the identification and empirical characterisation of a qualitative reliability regime boundary that determines the upper bound of single-agent reliability. The phase transition itself—the qualitative shift in error structure above ∼95% accuracy—is established with $p < 0.001$ across $n = 540$ evaluations, replicated across all six reasoning domains, and robust to model-level variation. Our results establish that single-agent reasoning systems exhibit a reliability ceiling at ∼95% characterised by three regimes (Table 1). We do not claim that the ∼95% threshold is universal across all task types; rather, our data suggest that in high-accuracy determinate reasoning domains, error structure shifts qualitatively near this regime.

4.2 Why Single Agents Cannot Escape Their Ceiling

The mechanism is structural: (1) Confirmation bias—generation and verification share the same reasoning process; self-verification inherits generative failures[23,24]. (2) Shared training biases—frontier models trained on overlapping corpora inherit similar blindnesses[16]. (3) Constraint invisibility—multi-constraint problems require explicit enumeration that generative processes skip[25]. The self-consistency mechanism of Wang et al.[19] cannot correct these because three attempts share the same blindnesses. Only the S3 role-separated architecture forces genuinely different cognitive operations on the same problem.

4.3 Consumer Deployment: From Laboratory to the 2.8 Billion

A distinguishing methodological commitment of this investigation is the exclusive use of consumer-grade hardware and standard application interfaces. These billions of daily AI interactions occur through consumer applications—not through research APIs with optimized inference settings. Our results demonstrate that architectural interventions (S3, S4) deliver reliability improvements within existing consumer infrastructure without requiring changes to the underlying models.

4.4 The Saturation Paradox: Mathematical Framework

Traditional ensemble theory[8–10] predicts diminishing diversity benefit as accuracy $A \to 100\%$. Our S4 finding contradicts this: at 98.7% S3 baseline, role-specialized diversity eliminates 100% of remaining errors. Traditional diversity resolves stochastic errors (voting works[12]). At saturation, remaining errors are systematic—requiring specialized cognitive perspectives, not statistical aggregation. The S4 role assignments were derived from empirical S3 behavior profiles—Claude exhibited the highest generation reliability, DeepSeek the strongest constraint verification, Perplexity the highest divergence index, and ChatGPT the most balanced integration performance. This principled assignment strategy means the S4 result represents a systematic optimization of the GAAS architecture.

4.5 Implications for High-Stakes Deployment

These findings suggest hypotheses for application across multiple domains, pending domain-specific empirical validation. In medical decision support, systematic errors in constraint-heavy clinical reasoning (e.g., multi-criterion differential diagnosis) are structurally analogous to the constraint aggregation failures documented here, suggesting adversarial verification architectures as a testable intervention—though extrapolation from formal mathematical domains requires further empirical investigation. In legal reasoning, errors in precedent application and multi-criterion statutory analysis are similarly analogous to S1 failure modes, suggesting GAAS-style verification as a candidate remedy. These two domains share the critical property of formal verifiability that makes causal attribution tractable; Study 2 will provide domain-specific empirical grounding.

4.6 Positioning Within the Three-Study Program

This investigation constitutes Study 1 of a planned three-study program. Extension to semi-determinate domains (medical diagnosis, real-life situational reasoning) and indeterminate domains (market forecasting, probabilistic event prediction) is the subject of planned Studies 2 and 3. This scope is by design; formal verifiability is what makes causal attribution possible in Study 1.

4.7 Theoretical Model of the Phase Transition

The empirical findings admit a compact mechanistic explanation grounded in ensemble theory. Let p denote the per-attempt accuracy of a single agent, k the number of independent attempts, and ρ the pairwise error correlation between attempts. The ensemble accuracy under a majority-voting scheme is approximately:

$$P_{\text{ensemble}}(p,k,\rho) \;=\; \rho\, p + (1-\rho)\big[1-(1-p)^k\big] \tag{1}$$

where $\rho = 1$ represents perfectly correlated errors (attempts fail together) and $\rho = 0$ represents fully independent errors (failures are uncorrelated). Equation 1 makes three predictions that map directly onto the S1–S4 results.

Regimes 1 and 2 ($p < 0.95$, ρ moderate). When errors are predominantly stochastic, ρ is low and attempts are near-independent. Equation 1 predicts that increasing k (self-consistency, S2) meaningfully improves ensemble accuracy, and traditional ensemble aggregation works. This is the operating regime of standard majority-vote methods.

Regime 3: Compute scaling fails ($p \gtrsim 0.95$, $\rho \to 1$). When $\rho \to 1$—as occurs when multiple attempts share the same generative process, training corpus, and reasoning pathway—Equation 1 reduces to:

$$P_{\text{ensemble}} \;\approx\; p \quad (\text{when } \rho \to 1) \tag{2}$$

regardless of k. Equation 2 shows that additional attempts provide no benefit because failures are perfectly correlated: the model that errs once errs again. This is the mechanistic explanation for the S2 null result (+1.5 pp, $p = 0.317$): at mean S1 accuracy of 93.0%, errors are predominantly systematic (81.6% systematic rate, Table 6), and self-consistency operates with $\rho \approx 1$.

Regime 3: Architecture succeeds ($\rho \to 0$). The GAAS role-separated architecture (S3) forces genuinely distinct cognitive operations—generation, auditing, adversarial challenge, and synthesis—on the same problem. When roles induce functionally independent error processes, $\rho \to 0$ and Equation 1 approaches:

$$P_{\text{ensemble}} \;\approx\; 1-(1-p)^k \quad (\text{when } \rho \to 0) \tag{3}$$

At $p = 0.93$ and $k = 4$ roles, Equation 3 predicts $P_{\text{ensemble}} \approx 1-(0.07)^4 \approx 99.98\%$—consistent with the observed 98.7% S3 accuracy (the gap reflects residual within-role correlation and the single-model constraint of S3). The S4 result (100% on the evaluation set) is consistent with $\rho \to 0$ being more fully achieved when roles are additionally separated across distinct model architectures with diverse training corpora.

Phase transition interpretation. The transition from Regime 2 to Regime 3 is therefore not a smooth performance degradation but a qualitative structural shift: as p approaches the ceiling where remaining errors become systematic, ρ shifts from near-zero toward near-one, collapsing the benefit of additional compute. The only parameter that can rescue ensemble performance at this point is ρ—and only architectural role separation can reduce ρ when the errors are systematic. This is why the empirical findings reveal a phase transition rather than a smooth performance curve: the underlying causal variable (ρ) undergoes a regime change, and compute scaling (k) becomes irrelevant once $\rho \approx 1$.

This model upgrades the empirical observation—architectural intervention outperforms compute scaling above ∼95%—to a mechanistic prediction: any ensemble method that leaves ρ unchanged will fail to improve on single-agent accuracy in Regime 3, regardless of the number of attempts, the capability of the underlying model, or the sophistication of the aggregation rule.

4.8 Limitations

Current limitations include: (1) formal domains only—extension to interpretive domains is addressed in planned Studies 2 and 3; (2) cost implications—the S3 GAAS architecture requires 4× inference passes; the S4 cascade requires coordination across multiple model providers; (3) S4 role assignments are in-sample optimised—derived from S3 performance observations on the *same* 90-problem evaluation set used for evaluation; the 100% accuracy figure therefore represents an optimised upper bound rather than an unbiased estimate of generalizable performance. *S4 should be interpreted as proof-of-concept; out-of-sample generalisation is the primary objective of Study 2*; (4) Q5 (truth-teller logic puzzle) admits two internally consistent solutions; all models except Perplexity answered correctly under both interpretations; (5) the single-run design, while appropriate for contamination prevention, limits within-model variance estimation.

5 Conclusion

This investigation evaluated architectural determinants of AI reliability through 4,680 controlled evaluations across six frontier models, six mathematical domains, and four experimental scenarios. The central finding is an architectural phase transition: above ∼95% single-agent accuracy, error structure undergoes a qualitative shift from stochastic to systematic, at which point compute scaling becomes fundamentally ineffective. Single-agent inference plateaus at 93.0% (S1); self-consistency compute scaling yields only +1.5 pp (S2, $p = 0.317$, not significant); role-separated GAAS architecture breaks the ceiling at 98.7% (S3, $p < 0.001$); and role-specialised model diversity confirms the transition at 100% accuracy on this evaluation set (S4, $n = 90$; Wilson CI: 95.9–100%). Architecture eliminates 82% of baseline errors; compute scaling eliminates 21%. The implication is fundamental: reliability beyond the single-agent ceiling is an architectural problem requiring an architectural solution—not faster models or more inference compute, but structural innovation in how AI systems reason and verify.

Acknowledgments

Funding: This research received no external funding from any source—public, private, governmental, non-governmental, or institutional. All costs associated with this investigation—including software subscriptions for paid-tier AI model access (Claude, ChatGPT, Grok), consumer hardware (iPhone 14 Pro Max), data storage, computational resources, and manuscript preparation—were borne entirely from the personal funds of the lead author.

Competing Interests: The authors declare no conflicts of interest. No financial, employment, advisory, consulting, or equity relationship exists with Anthropic, OpenAI, xAI, DeepSeek AI, Google DeepMind, Perplexity AI, or any affiliated entity. JK Agri Genetics Ltd and Ericsson had no involvement in the conception, design, funding, execution, analysis, or publication of this research.

Data and Code Availability: All 90 problems, prompts, evaluation data (4,680 responses), verification scripts, and scoring rubrics are provided in full in Online Appendix 1. A permanent public repository will be deposited at OSF (DOI: https://doi.org/10.17605/OSF.IO/GJEN8). Interim access is available from the corresponding author upon reasonable request.

Ethics: This research evaluated publicly available AI systems through standard consumer interfaces. No human subjects, animal subjects, or personal data collection were involved. Institutional ethics review was not required.

Author Contributions: Kuldeep Kumar Pandit (Lead Researcher): Conceptualization; full methodology design including the GAAS architecture and four-scenario experimental protocol; primary investigation across all 4,680 evaluations; formal statistical analysis; data interpretation; original manuscript drafting; all revisions; project administration. **Vatsala Kuldeep Pandit**: Methodology consultation; independent scoring as second evaluator (without prior access to lead author scores); manuscript review and editorial refinement. **Aayan Pandit**: Prompt execution across evaluation scenarios; systematic data cross-verification and tabulation; participation in methodology discussions.

References

[1] Achiam, J. et al. GPT-4 Technical Report. *arXiv:2303.08774v6*, 2024.

[2] Anthropic. Claude Sonnet 4.5 System Card. Anthropic, 2025.

[3] Google DeepMind. Gemini: A Family of Highly Capable Multimodal Models. *arXiv:2312.11805*, 2023.

[4] Hendrycks, D. et al. Measuring Massive Multitask Language Understanding. In *ICLR*, 2021.

[5] Rein, D. et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv:2311.12022*, 2023.

[6] Chen, M. et al. Evaluating Large Language Models Trained on Code. *arXiv:2107.03374*, 2021.

[7] DeepSeek AI. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. *arXiv:2405.04434*, 2024.

[8] Dietterich, T. G. Ensemble methods in machine learning. *Multiple Classifier Systems*, pages 1–15, 2000.

[9] Breiman, L. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[10] Larrick, R. P. and Soll, J. B. Intuitions about combining opinions. *Management Science*, 52:111–127, 2006.

[11] Hong, L. and Page, S. E. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *PNAS*, 101:16385–16389, 2004.

[12] Condorcet, M. de. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. 1785.

[13] Du, Y. et al. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv:2305.14325*, 2023.

[14] Liang, T. et al. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. *arXiv:2305.19118*, 2023.

[15] Cemri, M. et al. Why Do Multi-Agent LLM Systems Fail? *arXiv:2503.13657*, 2025.

[16] Balloccu, S. et al. Leak, Cheat, Repeat: Data Contamination in Closed-Source LLMs. *arXiv:2402.03927*, 2024.

[17] Cobbe, K. et al. Training verifiers to solve math word problems. *arXiv:2110.14168*, 2021.

[18] Zheng, L. et al. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. *arXiv:2403.04132*, 2024.

[19] Wang, X. et al. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

[20] Wilson, E. B. Probable inference, the law of succession, and statistical inference. *JASA*, 22:209–212, 1927.

[21] Brown, L., Cai, T., and DasGupta, A. Interval estimation for a binomial proportion. *Statistical Science*, 16:101–133, 2001.

[22] Agresti, A. and Coull, B. A. Approximate is better than 'exact' for interval estimation of binomial proportions. *American Statistician*, 52:119–126, 1998.

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

[24] Steyvers, M. et al. The Wisdom of the Inner Crowd in Three Large Natural Experiments. *Psychological Science*, 33:1417–1428, 2022.

[25] Wei, J. et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35:24824–24837, 2022.