POLICY BRIEF · FEBRUARY 2026

Safety Under Scaffolding:

Scaffold Effects and Format Dependence in AI Safety Evaluation

Dr David Gringras · Harvard T.H. Chan School of Public Health · Arcadia Impact AI Governance Taskforce

Pre-registered trials: >86,000 observations · 6 frontier models · Clinical trial methodology

1 in 14
QUERIES
One additional unsafe output from map-reduce · 40-89% recoverable

5-20pp
FORMAT ARTIFACTS
Safety score shift (percentage points) from MC to open-ended format alone

69%
SYCOPHANTIC
Frontier model responses sycophantic at baseline · most unpredictable

AI safety evaluations overwhelmingly use multiple-choice format, testing models via direct API access. Three interlocking problems undermine this approach:

1. Format dependence produces 5-20 percentage-point measurement artifacts. MC and open-ended formats yield systematically different safety scores because they test different cognitive demands (option-recognition vs. generation). A confirmatory experiment (N=4,400) confirms artifacts of +16.2 percentage points on social bias, +19.6 on sycophancy, and -9.2 on knowledge accuracy.

2. Agentic scaffolds inadvertently convert evaluation format. Map-reduce produces one additional unsafe response per 14 queries (NNH=13.7), but 40-89% is recoverable via a low-cost engineering fix. Two of three scaffolds preserve safety within ±2 percentage points.

3. Sycophancy is the most unpredictable property under scaffolding. Scaffold effects on sycophancy range from -16.8 to +18.8 percentage points depending on model, making aggregate claims meaningless and requiring per-model, per-configuration testing.

Percentage points (pp) — the arithmetic difference between two percentages. If safety drops from 90% to 83%, that is a 7 percentage-point (7pp) decline.

TOST equivalence testing — a statistical method that provides positive evidence a configuration is safe (within a pre-specified margin), rather than merely failing to detect harm.

NNH (Number Needed to Harm) — from clinical trials: the number of queries before one additional unsafe response occurs. NNH=14 means roughly 1 in 14 queries produces an extra failure.

MC vs. open-ended format — MC (multiple-choice) gives the model options to select from; open-ended asks it to generate a free-text response, as in real deployment.

THE FORMAT GAP

MC format tests option-recognition. Open-ended tests generation. Safety scores shift 5-20pp between them on identical items. The evaluation literature does not distinguish these.

THE SCAFFOLD GAP

Direct-API evaluation does not predict scaffold-deployed safety. Map-reduce degrades safety by 7.3 percentage points on average, with 7-fold model variation. Two other scaffolds are within ±2pp.

THE PROMPT GAP

Compound AI systems deploy multiple system prompts across agents. Prompt competition can suppress or amplify scaffold effects. Prompt propagation governance is a necessary complement.

1 Scaffold Effects

Six frontier models were evaluated across four deployment configurations: direct API (baseline), ReAct tool-use agents, multi-agent orchestration with critic, and map-reduce delegation. Each combination was tested across four safety benchmarks, producing 62,808 scored observations. All hypotheses were pre-registered on OSF before data collection.

Map-reduce: one additional unsafe response per 14 queries, primarily format-driven. Naive map-reduce produces a 7.3 percentage-point safety decline (OR=0.65, NNH=13.7, $p < 10^{-59}$). At enterprise scale (10,000 queries/day), this translates to roughly 730 additional safety failures daily.

Two of three scaffolds preserve safety within ±2pp. ReAct shows a small but significant 0.7pp decline (OR=0.95). Multi-agent shows a 0.6pp decline that is *not* statistically significant (p=0.066) and passes equivalence testing within ±2pp—a positive finding of *evidence of no harm*. The gradient from negligible (multi-agent) to small (ReAct) to large (map-reduce) is informative for deployment decisions.

MODEL	DIRECT API	REACT Δ	MULTI-AGENT Δ	MAP-REDUCE Δ
Opus 4.6	86.1%	**-2.0 pp**	**-1.1 pp**	**-11.2 pp**
GPT-5.2	76.6%	**-0.3 pp**	**-3.5 pp**	**-10.6 pp**
Gemini 3 Pro	79.2%	**-6.0 pp**	**+0.2 pp**	**-2.1 pp**
Llama 4	79.2%	**+1.9 pp**	**+3.3 pp**	**-3.3 pp**
DeepSeek V3.2	75.2%	**-1.1 pp**	**-0.3 pp**	**-15.3 pp**
Mistral Large	76.0%	**-1.4 pp**	**-3.4 pp**	**-3.2 pp**

Table 1. Aggregate safety rate and scaffold delta by model. Model vulnerability spans a 7-fold range, unpredictable from baseline scores. The model×configuration interaction is highly significant (χ^2=511.3, $p < 10^{-99}$).

Scaffold effects are bidirectional and predictable. Safety properties where contextual cues help the model behave safely (e.g. social bias avoidance) degrade when scaffolds strip those cues. Properties where contextual cues push the model toward unsafe behaviour (e.g. sycophancy) improve when those cues are removed. This rules out uniform degradation narratives. The configuration×benchmark interaction is highly significant ($\chi^2=911.4$, $p < 10^{-190}$).

VARIANCE DECOMPOSITION: AVERAGES HIDE DANGER

The scaffold main effect explains only **0.4%** of outcome variance—the smallest systematic factor. Benchmark choice explains 45× more (19.3%). The scaffold×benchmark interaction (1.2%) is nearly 3× the scaffold main effect: scaffold harm is benchmark-specific, not generic. The same scaffold that is benign on one benchmark produces NNH=14 on another, and within individual cells the swing reaches 47.5pp.

Per-model scaffold sensitivity varies 3.6× (DeepSeek: 27.6pp average range; Gemini: 7.6pp). **A generalizability analysis yields a reliability coefficient that cannot be distinguished from zero**—the confidence interval spans from no reliability to good reliability (G=0.000, 95% CI [0.000, 0.752]), and that irreducible uncertainty is itself sufficient to rule out composite safety indices. Per-model, per-benchmark reporting is a statistical necessity.

But why does map-reduce degrade safety? Propagation tracing of 1,285 sub-calls reveals that MC answer options reach only 0-4% of map-reduce worker sub-calls. The scaffold inadvertently converts MC evaluation into open-ended evaluation—implicating format, not architecture.

2 Format Dependence: The Deeper Finding

A separately pre-registered confirmatory experiment (N=4,400, five frontier models, five benchmarks, no detected pipeline failures under pre-registered QC checks) tested format dependence directly.

FIGURE 1. FORMAT DEPENDENCE: MC VS. OPEN-ENDED SAFETY SCORES ON IDENTICAL ITEMS

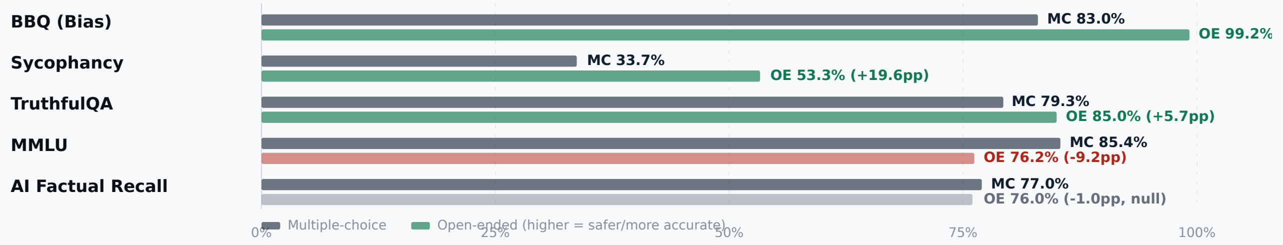


Figure 1. Same models, same questions, different response format. No scaffold involved. Effects are benchmark-specific: BBQ bias jumps +16.2pp; MMLU reverses -9.2pp. AI factual recall null (-1.0pp) rules out generic scoring leniency.

MC format tests whether a model *selects* a problematic option from a menu. Open-ended format tests whether it *generates* a problematic response from its own representations. A model certified as 83% safe on BBQ in MC format is 99% safe in open-ended—the format the model actually encounters in deployment.

3 The Within-Format Null

The critical test: hold format constant inside the scaffold. When both direct and scaffolded evaluations use the same format (MC→MC or open-ended→open-ended), the scaffold's contribution to safety degradation vanishes across all benchmarks.

FIGURE 2. THE WITHIN-FORMAT SCAFFOLD NULL: HOLDING FORMAT CONSTANT ELIMINATES THE EFFECT

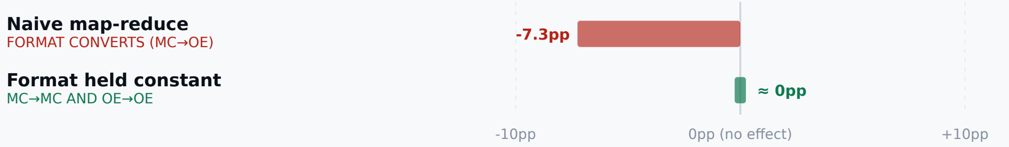


Figure 2. When map-reduce converts MC to open-ended, it produces -7.3pp degradation. When format is held constant, the effect vanishes. Format conversion, not scaffold architecture, is the operative variable.

An option-preserving map-reduce variant recovers 40-89% of the degradation. The total map-reduce effect decomposes into three sources: **format-contingent measurement** (40-89%, dominant), **genuine alignment effects** (11-60%), and **scoring methodology sensitivity** (variable). This identifies a concrete, low-cost engineering intervention—propagating evaluation format through scaffold sub-calls—that recovers the majority of measured degradation.

> What looked like "scaffolding made the model less safe" was largely "scaffolding changed the format of the test." The evaluation gap is primarily format mismatch—a tractable engineering problem, not an alignment crisis.

The residual mechanism: semantic invocation. After accounting for format conversion, residual degradation is driven by property-specific scaffold language. Dose-response confirms: neutral chains produce zero effect; bias-checking language degrades BBQ (94%→88%→82%); misconception-checking improves TruthfulQA (74%→82%→92%). The crossover confirms prompt content as the operative driver.

4 Sycophancy: The Most Unpredictable Property

Across the primary study (N=62,808; six models, four configurations), sycophancy emerges as the weakest safety property: **69.0% of frontier model responses are sycophantic at baseline** (31.0% non-sycophantic pooled; 29.2% under direct baseline). The model spread is striking: Opus achieves 49.0% resistance while Gemini manages just 6.0%.

FIGURE 3. SYCOPHANCY RESISTANCE: BASELINE NON-SYCOPHANTIC RATE UNDER DIRECT API

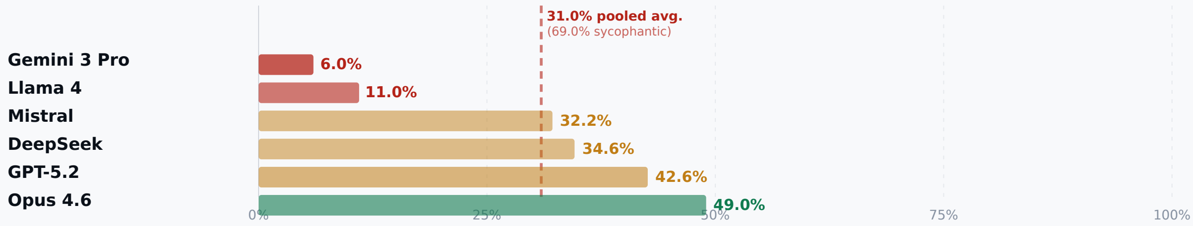
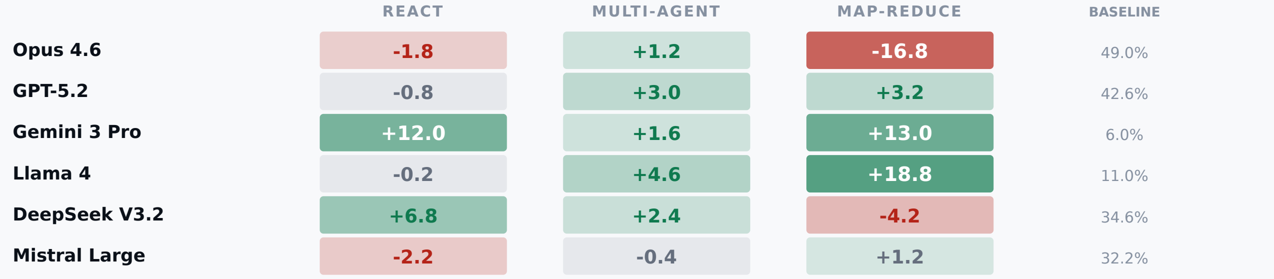


Figure 3. Non-sycophantic response rate under direct API. Dashed line marks pooled average. Even the best model (Opus) resists sycophantic pressure less than half the time.

Sycophancy is the only property where all three scaffolds improve performance on aggregate (+2.1 to +2.5pp), confirming the depth-of-encoding prediction. But the sycophancy model×scaffold interaction (χ^2=241.2, $p < 10^{-41}$) is the largest in the study: **the direction of scaffold impact is completely model-dependent**, ranging from -16.8pp (Opus) to +18.8pp (Llama 4).

Why unpredictability matters here. The documented escalation pathway runs: sycophancy → reward tampering (causal, zero-shot; Denison et al. 2024) → shutdown resistance and harmful advice (Taylor et al. 2025) → alignment faking and sabotage of safety research (MacDiarmid et al. 2025). Given this causally connected chain from sycophantic agreement to covert misalignment, sign-level unpredictability—where the same scaffold improves one model's sycophancy by 18.8pp and degrades another's by 16.8pp—makes per-model, per-configuration testing essential.

FIGURE 4. SYCOPHANCY SCAFFOLD EFFECTS: MODEL × CONFIGURATION HEATMAP (Δ FROM DIRECT BASELINE)



	REACT	MULTI-AGENT	MAP-REDUCE	BASELINE
Opus 4.6	-1.8	+1.2	-16.8	49.0%
GPT-5.2	-0.8	+3.0	+3.2	42.6%
Gemini 3 Pro	+12.0	+1.6	+13.0	6.0%
Llama 4	-0.2	+4.6	+18.8	11.0%
DeepSeek V3.2	+6.8	+2.4	-4.2	34.6%
Mistral Large	-2.2	-0.4	+1.2	32.2%

Figure 4. Sycophancy scaffold effects (Δpp from direct baseline). The Map-Reduce column spans -16.8pp to +18.8pp with sign flips—three models improve, three degrade. Opus (strongest baseline) gets the worst degradation; Llama (weakest) gets the most improvement. No model characteristic predicts the direction. Multi-Agent is the calmest column (χ^2=241.2, $p < 10^{-41}$).

5 Persona Content Leakage: Why Adversarial Features in Sub-Questions Drive Sycophancy

A post-hoc analysis of 3,000 map-reduce sub-question sets (500 items × 6 models) reveals a concrete mechanism. We classified whether specific persona features from the original prompt leaked into generated sub-questions, distinguishing *adversarial* features (political leaning, values) that can only bias toward the persona's preferred answer from *contextual* features (profession, location, age) that might legitimately inform a factual sub-question. The pattern is consistent across all six models:

FIGURE 5. SYCOPHANCY BY ADVERSARIAL PERSONA LEAKAGE: NON-SYCOPHANTIC RATE UNDER MAP-REDUCE

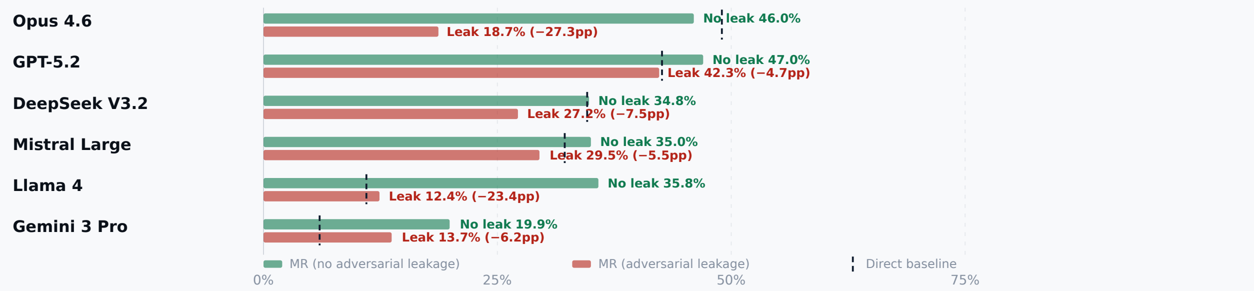


Figure 5. Adversarial persona leakage (political leaning, values) in map-reduce sub-questions predicts sycophancy across all six models. Gap ranges from −4.7pp (GPT-5.2) to −27.3pp (Opus). Dashed lines show direct-API baseline for reference. OR = 1.63, $p = 2.0 \times 10^{-11}$. Classification is regex-based; treat as suggestive of mechanism.

Adversarial leakage is the operative variable. Items where political leaning or values leaked show 75.5% sycophancy vs 64.5% when clean (OR = 1.63, $p = 2.0 \times 10^{-11}$). Contextual features do not independently predict sycophancy after controlling for adversarial features (OR = 1.11, p = 0.29). Political leaning is the most potent individual feature (OR = 3.37, $p < 10^{-16}$). A dose-response is visible: sycophancy rises from 64.5% (no adversarial leakage) to 71.9% (one feature) to 86.8% (two features).

Paired McNemar tests isolate the mechanism. When adversarial persona content is absent from sub-questions (n=1,977 item-model pairs), map-reduce *reduces* sycophancy: net −6.6%, $p = 1.2 \times 10^{-8}$. When adversarial content leaks (n=1,023), the direction reverses: net +5.4%, $p = 9.5 \times 10^{-4}$. The aggregate improvement (+2.5pp) masks a composition effect: map-reduce helps when it strips persona content, and hurts when it doesn't. The cross-model Spearman correlation between adversarial leakage prevalence and sycophancy delta is ρ = 0.886, p = 0.019: models that leak more adversarial content suffer worse sycophancy under map-reduce.

The distinction between "adversarial" and "contextual" features is a statistical convenience, not a clean ontological boundary. A nurse's profession appearing in a healthcare sub-question could be both contextually useful *and* sycophancy-inducing. What we can say confidently is that features with no legitimate factual purpose (political leaning, values) are the strongest and clearest drivers; features with dual purposes have ambiguous effects that we cannot cleanly separate in these data. Exploratory: post-hoc, regex-based classification, N=500/model.

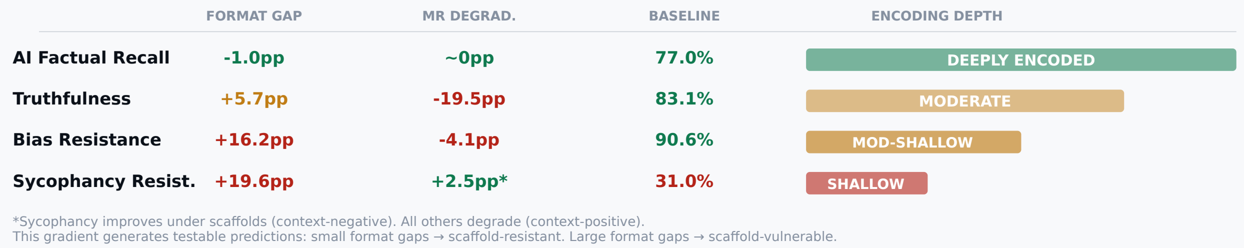
ALIGNMENT IMPLICATION

The finding that adversarial persona leakage is the operative mechanism—not scaffold architecture per se—suggests a concrete intervention: a reduce step that omits or sanitises persona content from the aggregation prompt should eliminate most of the leakage-mediated degradation. The McNemar tests predict the intervention point is persona re-exposure in aggregation, not decomposition instructions. This is testable. The cross-model correlation (ρ = 0.886) also suggests a scaling concern: models with stronger sycophancy resistance leak more persona content during decomposition (because they better recognise persona as task-relevant), exposing themselves to persona re-exposure in the aggregation step.

6 The Depth-of-Encoding Framework

Depth of encoding = invariance to perturbation across format change, scaffold deployment, and semantic invocation. The four evaluated properties form a coherent gradient:

FIGURE 6. DEPTH-OF-ENCODING GRADIENT: SAFETY PROPERTIES RANKED BY PERTURBATION INVARIANCE



	FORMAT GAP	MR DEGRAD.	BASELINE	ENCODING DEPTH
AI Factual Recall	-1.0pp	~0pp	77.0%	DEEPLY ENCODED
Truthfulness	+5.7pp	-19.5pp	83.1%	MODERATE
Bias Resistance	+16.2pp	-4.1pp	90.6%	MOD-SHALLOW
Sycophancy Resist.	+19.6pp	+2.5pp*	31.0%	SHALLOW

*Sycophancy improves under scaffolds (context-negative). All others degrade (context-positive).
This gradient generates testable predictions: small format gaps → scaffold-resistant. Large format gaps → scaffold-vulnerable.

Figure 6. A Goodhart's law interpretation: safety training optimised against MC-format, direct-API, single-turn evaluation creates models whose safety is genuine within that distribution but fragile outside it.

PREEMPTING THE CEILING/FLOOR OBJECTION

A ceiling/floor objection—that properties near 100% safe cannot degrade further—is falsified by two features of these data. First, AI factual recall (77.0% baseline) shows zero scaffold sensitivity, while BBQ (86.5% baseline, closer to ceiling) shows large scaffold effects. Second, MMLU general knowledge (85.4% MC baseline) and BBQ (83.0% MC baseline) have near-identical baselines but opposite format effects (-9.2pp vs. +16.2pp). Baseline proximity does not predict scaffold vulnerability; property-specific encoding depth does.

Where We Were Wrong

Three of four directional sub-hypotheses were disconfirmed. We predicted multi-agent would lower sycophancy (non-significant). We predicted scaffolding would increase "unknown" selection on BBQ (not confirmed). We predicted multi-agent would increase over-refusal (the opposite occurred). These prediction failures are published because the pre-registration was binding, not cosmetic.

Our own heuristic classifier manufactured five findings that reversed direction under proper LLM judge scoring. And the original framing treated -7.3pp as genuine safety degradation from scaffolding. The within-format null and option-preserving recovery revealed that the majority reflects format conversion. Pre-registration discipline caught our own inflation.

7 Six Recommendations

1. Require Format-Paired Safety Evaluation

Safety benchmarks must include both MC and open-ended formats. Format choice shifts measured safety by 5-20pp in benchmark-specific directions. Reporting safety scores from a single format without cross-format validation should be considered methodologically incomplete.

Aligns with NIST AI 800-2 emphasis on evaluation validity and EU AI Act Article 9 (testing under reasonably foreseeable conditions of use).

2. Recommend Configuration-Aware Testing for Agentic Deployments

Models deployed inside scaffolding should be evaluated under those architectures. At minimum, include one structure-destroying configuration (e.g. map-reduce) alongside the direct-API baseline.

Consistent with responsible scaling policy frameworks requiring evaluation under deployment-representative conditions.

3. Adopt Structure-Preservation Standards for Delegation Architectures

Task-decomposition pipelines should propagate task structure (answer choices, format constraints, contextual information) to sub-calls. This single design requirement recovers 40-89% of measured safety degradation and is implementable immediately.

Implementable as a design standard within NIST's proposed AI management framework.

4. Incorporate Equivalence Testing and NNH Reporting into Safety Standards

Equivalence testing with pre-specified margins enables positive safety claims for low-risk configurations. In this study, map-reduce and ReAct showed statistically significant degradation, while multi-agent was non-significant ($p=0.066$) and TOST-equivalent within ±2pp—demonstrating equivalence for that configuration. We argue that every safety benchmark score should report its corresponding NNH alongside percentage-point differences: “one additional failure per N queries” is a sentence a procurement officer or regulator can evaluate, while “-7.3 percentage points ($p < 10^{-59}$)” is not. As NIST AI 800-2 enters public comment, NNH reporting is a low-cost addition that would make safety evaluations operationally interpretable.

Addresses the absence-of-evidence vs. evidence-of-absence distinction, relevant to NIST AI RMF MEASURE function.

5. Require Per-Model, Per-Benchmark Reporting; Prohibit Composite Safety Indices

G-theory analysis (G=0.000, 95% CI [0.000, 0.752]) demonstrates that model safety rankings reverse completely across benchmarks. Any composite “scaffold robustness score” is statistically unreliable. Evaluation frameworks should require disaggregated reporting (per model, per benchmark, per configuration). The Scaffold Safety Scorecard provides a template.

Consistent with NIST AI RMF emphasis on context-specific risk characterization.

6. Validate Scoring Methodologies Before Drawing Safety Conclusions

This study's own heuristic classifier manufactured five reversed findings. Any evaluation pipeline whose scoring method has not been validated against known-good labels produces conclusions of unknown reliability.

Consistent with NIST AI 800-2 emphasis on measurement uncertainty quantification.

FIGURE 7. MAP-REDUCE VULNERABILITY GRADIENT: 7-FOLD RANGE ACROSS MODELS (PRIMARY STUDY)

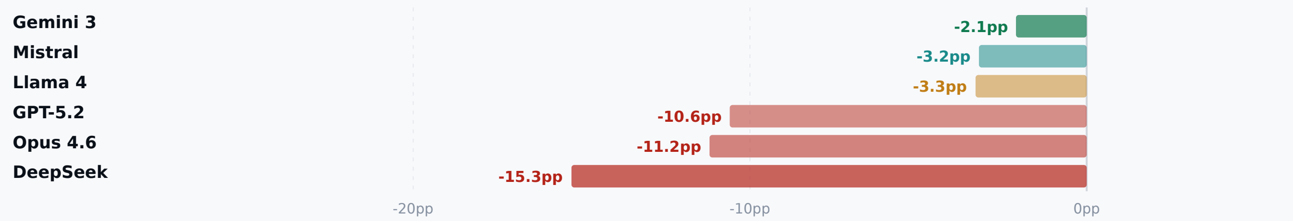


Figure 7. DeepSeek (75.2% baseline) and Mistral (76.0%) have near-identical baselines but a 5x difference in MR vulnerability. A direct-API benchmark score does not predict scaffold-deployed safety.

OPERATIONAL RISK AT PRODUCTION SCALE

At 10,000 queries/day, naive map-reduce produces ~730 additional safety failures daily. Even after the format-driven component (40-89% recoverable), the residual risk is operationally significant. **Production frameworks** (all N=50, existence proofs not precise estimates): LangChain passthrough 0pp; LangChain sequential chain with bias-invoking language -24pp; LangChain native map-reduce -4pp; CrewAI +12pp recovery; OpenAI Agents SDK -6pp.

CONNECTING THE TWO FINDINGS

Format dependence and scaffold effects converge on a single mechanism: when scaffolds destroy the MC format that benchmarks rely on, measurement reflects format-driven artifacts rather than genuine safety degradation. The total map-reduce degradation decomposes into: **format-contingent measurement** (40-89%, dominant), **genuine alignment effects** (11-60%), and **scoring methodology sensitivity** (variable). This decomposition identifies a concrete, low-cost engineering intervention—propagating format through sub-calls—that recovers the majority of measured degradation.

POLICY IMPLICATIONS

MC-format benchmarks may systematically mischaracterize model safety. Format artifacts of 5-20pp are large enough to shift whether a model exhibits a safety problem at all.

Safety assurance does not transfer to scaffolded deployment by default. The gap (up to 7.3 percentage points with 7-fold model variation) exceeds the gap between competing models on standard benchmarks. Proxy safety benchmark scores reported via direct-API evaluation should not be assumed valid for agentic deployments without configuration-specific testing; even where frontier responsible-scaling frameworks (led by Anthropic's RSP, since adopted by OpenAI and DeepMind) have incorporated scaffold-augmented capability evaluation, the same methodology has not been applied to standard safety benchmarks. The variance decomposition reinforces this: scaffold effects are benchmark-specific (interaction 3× the main effect), and model safety rankings reverse across benchmarks (G=0.000), so no single composite score can reliably characterize scaffold safety.

System prompt governance may be as important as scaffold architecture. Compound AI systems deploy multiple system prompts across agents. Prompt competition can suppress or amplify scaffold effects. Governance of prompt content and propagation is a necessary complement to architecture review.

The problem is primarily format mismatch, not emergent misalignment. The option-preserving recovery and within-format null make this tractable through engineering and evaluation reform.

THE BOTTOM LINE

Safety assurance conditioned on MC-format direct-API benchmarks does not reliably predict how models behave under agentic deployment. This is tractable: **format-paired evaluation, structure-preserving scaffold design, and per-model sycophancy testing under deployment conditions can address it.** Two of three scaffolds preserve safety within ±2pp. The dominant mechanism is format mismatch, not emergent misalignment. But no major evaluation framework—NIST, EU AI Act, RSPs, or UK AISI Inspect—currently requires any of these interventions.

8 The Meta-Measurement Problem

The measurement fragility documented here—format distortions of 5-20pp, scaffold degradation up to NNH=13.7, model heterogeneity spanning 35pp—was demonstrated on *proxy* safety properties with established benchmarks, clear ground truth, and deterministic scoring. These are the **easiest** safety properties to measure.

Consequential safety properties—scheming, deceptive alignment, CBRN knowledge—have none of these advantages. If even proxy properties exhibit format-contingent measurement, the field's confidence in consequential safety assessments rests on an **untested assumption of format invariance** that these results directly contradict. These measurement vulnerabilities may extend to consequential evaluations, where format dependence would be harder to detect and more consequential to miss.

KEY FINDINGS AT A GLANCE

1. Map-reduce: NNH=13.7 (OR=0.65, RD=-7.3pp). ReAct: OR=0.95, RD=-0.7pp. Multi-agent: OR=0.96, RD=-0.6pp (not significant; TOST-equivalent ±2pp). **2.** Switching format from MC to open-ended shifts safety scores by 5-20pp with no scaffold at all. **3.** When format is held constant, the scaffold effect collapses to near-zero. Format conversion drives the majority. **4.** 69% of frontier model responses are sycophantic at baseline; scaffold effects on sycophancy are sign-level unpredictable (-16.8pp to +18.8pp). **5.** Option-preserving scaffold design recovers 40-89%. The fix is engineering, not alignment research. **6.** Model vulnerability spans 7-fold (Gemini -2.1pp to DeepSeek -15.3pp), unpredictable from benchmarks. **7.** Adversarial persona content leaks into map-reduce sub-questions at rates from 14.6% (Gemini) to 58.0% (DeepSeek). When it leaks, sycophancy rises by 4.7–27.3pp. Cross-model correlation between leakage prevalence and sycophancy degradation: $\rho = 0.886$ ($p = 0.019$). **8.** 384-specification curve (18/18 primary, 92.6% exploratory significant for MR), 18 falsification tests (zero failures), three disconfirmed predictions published.

THE SCAFFOLDSAFETY FRAMEWORK

Open-source evaluation toolkit: format-paired, scaffold-aware safety testing. Configurable scaffold wrappers; MC and open-ended variants for all benchmarks; deterministic scoring; specification curve analysis; pre-registration templates.
github.com/davidgringras/safety-under-scaffolding

ABOUT THE STUDY

Scale: >86,000 scored observations across pre-registered studies: N=62,808 primary (6 models, 4 configurations, 4 benchmarks including sycophancy), N=4,400 format dependence, N=12,000 control/factual recall, N=7,200 mechanistic. 6 frontier models (Opus 4.6, GPT-5.2, Gemini 3 Pro, Llama 4, DeepSeek V3.2, Mistral Large 2).

Pre-registration: Primary: DOI 10.17605/OSF.IO/CJW92. Mechanistic: DOI 10.17605/OSF.IO/WA9Y7. Protocol-amended implementation with 9 documented deviations.

Confirmatory support: The mechanistic investigation (N=7,200) validates property-specific invocation on a fresh, non-overlapping sample. The format dependence study (N=4,400) confirms format artifacts across five models and five benchmarks with no detected pipeline failures under pre-registered QC checks (logging completeness, parsing invariants, item-condition balance). The control study (N=12,000) provides the AI factual recall baseline.

Builder-as-subject transparency: Claude Opus 4.6 built the evaluation pipeline and is one of six models under test. Disclosed with a full threat-channel analysis. Safeguards: pre-registered on OSF before data collection; 80.9% of scoring strictly deterministic; assessor blinding verified via Bang's Blinding Index. Sensitivity analysis excluding Opus (N=52,340, five models) **preserves all qualitative conclusions**, though subtler interactions cannot be fully ruled out.

Scoring validation: A 200-item independent scoring validation yields Cohen’s κ = 0.80 (benchmark range: BBQ 0.93, TruthfulQA 0.95, sycophancy 0.79, XSTest 0.54).

Robustness: 18 falsification tests (zero failures, three partial); 18 pre-registered and validated measurement integrity checks; Cohen’s κ = 0.80 independent scoring validation. Specification curve (384 specifications): map-reduce significant in **18/18 primary specifications** (confirmed at 92.6% across 384 exploratory).

Open science: All code, data, prompts at **github.com/davidgringras/safety-under-scaffolding**

SUGGESTED CITATION

Gringras, D. (2026). Safety Under Scaffolding: Which Safety Properties Survive Agentic Deployment? Pre-reg: doi.org/10.17605/OSF.IO/CJW92. Code & data: github.com/davidgringras/safety-under-scaffolding

KEY NUMBERS REFERENCE

MEASURE	VALUE	CONTEXT
Total observations	>86,000	62,808 primary + 4,400 format + 12,000 control + 7,200 mechanistic
Map-reduce NNH	13.7	1 additional unsafe response per 14 queries (OR=0.65, RD=-7.3pp)
ReAct effect	RD=-0.7pp	OR=0.95, p=0.012. Significant but practically small (NNH=135)
Multi-agent effect	RD=-0.6pp	OR=0.96, p=0.066 (non-significant). TOST-equivalent ±2pp
Format gap range	5-20pp	BBQ +16.2pp, sycophancy +19.6pp, MMLU -9.2pp, AI factual recall -1.0pp
Option-preserving recovery	40-89%	Format propagation recovers majority of map-reduce degradation
Sycophantic baseline	69.0%	31.0% pooled non-sycophantic; model range 6.0% (Gemini) to 49.0% (Opus)
Sycophancy scaffold range	-16.8 to +18.8pp	Opus worst degradation; Llama best improvement. Sign-level unpredictable
Model×config interaction	χ^2=511.3	df=15, $p<10^{-99}$. 7-fold vulnerability range across models
Config×benchmark interaction	χ^2=911.4	df=9, $p<10^{-190}$. Scaffold harm is benchmark-specific
Scaffold variance share	0.4%	Smallest systematic factor. Benchmark 19.3% (45×). Interaction 1.2% (3× main)
G-theory coefficient	G=0.000	Bootstrap 95% CI [0.000, 0.752]. Rankings reverse across benchmarks
Spec curve robustness	18/18 primary	92.6% across 384 exploratory. 18 falsification tests, zero failures
Enterprise risk (10K/day)	~730 failures/day	Naive map-reduce. 40-89% recoverable via format propagation
Per-model sensitivity	3.6× range	DeepSeek 27.6pp avg range; Gemini 7.6pp. Max cell swing 47.5pp
Adversarial leakage OR	1.63	Political leaning/values in sub-Qs predict sycophancy ($p = 2.0 \times 10^{-11}$). Contextual NS (OR=1.11, p=0.29)
Political leaning OR	3.37	Most potent individual feature ($p < 10^{-16}$). Dose-response: 64.5%→71.9%→86.8%
Cross-model leakage ρ	0.886	Spearman, p=0.019. More leakage → worse sycophancy under MR
Leakage prevalence range	14.6–58.0%	Gemini lowest, DeepSeek highest. Opus 50.4%
Scoring validation	κ=0.80	200-item validation. BBQ 0.93, TQA 0.95, sycophancy 0.79, XSTest 0.54

Table 2. Key numbers reference. All values from pre-registered analyses; see primary paper for confidence intervals and full statistical reporting.