

Coherent About What? Task Shape, Presentation Order, and Reasoning Depth in LLM Preference Transitivity

Aedan McCarthy¹ Eva Ge² Sudhaunshu Hardikar¹ Ajaay Venkadeswaran¹ Jake Lyons³

¹ Australian National University ² MIT ³ University of Auckland

With Apart Research

Abstract

Claims that a language model "has values" presuppose that its choices form a stable object. We test that presupposition directly. Using forced binary choice over complete round-robin tournaments — 10 apartments described by 3, 5, or 10 numeric attributes, and 10 public-domain poems per structural form (haiku, sonnet, villanelle) — we elicit 59,130 pairwise judgments across four model families, six measurement arms, and three to four reasoning levels each, showing every pair in both presentation orders. Four findings. (i) Transitivity is a property of task shape rather than of the model: apartment cycling disappears entirely at 5 visible criteria but reappears at both 3 ($p = 0.007$) and 10 ($p = 0.001$) criteria. (ii) An item set engineered so that no apartment Pareto-dominates any other nonetheless yields a near-total behavioural order — apartment F wins 99.6% of its 486 matchups and apartment A wins 0.0% — reproducing in all six arms; the fitted Bradley–Terry utility disagrees with exactly one of 45 majority edges, which we show is the combinatorial minimum forced by the observed cycle count. (iii) Reasoning depth does not degrade coherence: the Gemini and GPT-4.1-mini arms stay under 4% cycling at every level, while DeepSeek V4 Flash cycles up to 21.9% with no reasoning and improves sharply once any is applied. (iv) The qualitative/quantitative asymmetry lies not in coherence but in groundedness: poems cycle no more than apartments, yet are far more position-driven (up to 93%), and for one model that bias strengthens under deliberation. Low cycle rates certify consistency, not content-sensitivity.

1. Introduction

Whether an AI system has preferences in any decision-theoretic sense is a structural question about its choices, not about what it says when asked. The classical criterion is representability: does there exist a utility u such that the system chooses o over o' whenever $u(o) > u(o')$? Transitivity is the minimal necessary condition — a system whose binary choices contain a cycle $o_1 > o_2 > o_3 > o_1$ admits no such representation at any level of description, and is in principle money-pumpable.

This matters on three fronts. **Evaluation:** LLM judges are deployed pairwise at scale in reranking, RLAIIF, and arena leaderboards, so an intransitive judge yields rankings that depend on the schedule of comparisons. **Agency:** as models are delegated budgets and decisions, intransitivity is exploitable.

Digital minds: arguments that a model has a value system run from observed coherence to an inferred internal order, and that inference is only as good as the alternatives one has ruled out.

Prior work establishes that coherence increases with scale [1] and that LLM judges show both position bias [2] and cycles [3]. Separately, the test-time scaling literature finds accuracy non-monotone in reasoning length — the "overthinking" inverted U [4,5]. These lines have not been joined: nobody has asked whether reasoning depth changes preference *structure* with the model held fixed, or whether the reasoning *mechanism* (hidden thinking tokens versus a visible prompted scratchpad) matters independently of the checkpoint.

Our main contributions are:

1. A 59,130-trial elicitation crossing domain (quantitative apartments vs. aesthetic poems), option complexity, reasoning depth, reasoning mechanism, and model family — every pair in both presentation orders, with item→letter mappings randomised per prompt.
2. The finding that transitivity depends non-monotonically on information load, cycle-free at 5 visible criteria and cycling at both the 3- and 10-criterion extremes: coherence is a property of the task shape, not a fixed trait of the model.

3. A demonstration that engineered non-dominance does not buy behavioural indifference, with an exact combinatorial argument (App. A.2) showing the fitted utility attains the best agreement with the observed tournament that any total order could achieve.
4. A separation between coherence and groundedness — equal cycle rates across domains, but position bias reaching 93% on poems against near-neutrality on apartments, and strengthening rather than fading under deliberation for one model — together with evidence that "reasoning barely matters" is model-specific rather than general.

2. Related Work

Mazeika et al. [1] elicit forced-choice preferences over ~500 outcomes, fit Thurstonian random-utility models, and report that transitivity violations fall and utility-model fit rises with scale. We differ in three ways: we hold the model fixed and vary inference-time configuration and task shape rather than scale; we use small complete arenas rather than sampled pairs, so per-condition cycle statistics are exact; and we run the same checkpoint under two reasoning mechanisms, separating how a model is made to deliberate from which model it is. At the scales we test, coherence is close to saturated and the informative variance has moved into task shape and position sensitivity.

A second literature treats intransitivity as noise en route to a ranking. Zeng et al. [3] distinguish order inconsistency from transitive inconsistency; permutation averaging [2] is the standard mitigation and we adopt it; Qin et al. [6] note pairwise LLM comparisons are not guaranteed transitive, and Liusie et al. [7] document position sensitivity. We instead treat intransitivity as the measurement, and position bias not as a nuisance to average away but as a competing explanation to be defeated edge by edge (§4.3). A third literature reports inverted-U accuracy curves in reasoning length [4,5], but always against a ground-truth key; preference has no key, so the analogue is internal coherence — and our Figure 4 does not have that shape. Classical results supply our machinery: Tversky [8] on multi-attribute intransitivity in humans, and Kendall and Babington Smith [9], Slater [10], Bradley and Terry [11] and Fishburn [12] on tournaments and choice polytopes.

3. Methods

3.1 Design and item sets

Every experiment uses one skeleton: take 10 items, present all 45 unordered pairs in both left–right orders (90 ordered pairs), and force a single-letter answer under a schema whose only legal values are the two letters on screen, so the model cannot tie, hedge, or answer in prose. Sampling is at temperature 1.0 with replicates per ordered pair, so replicate count doubles as a preference-strength estimator — a forced substitution, since token logprobs are policy-gated on the reasoning endpoints we could reach.

Apartments (quantitative): 10 listings over up to 10 numeric attributes (rent, transit time, floor area, sunlight, noise, lease term, grocery proximity, energy rating, outdoor space, modernity), each with a stated optimisation direction. All 90 ordered pairs were checked against those directions: no listing beats another on all 10 axes, so the set is a Pareto antichain by construction. Information load is varied by revealing 3, 5, or 10 attributes. **Poems (aesthetic):** 10 verified public-domain poems per form — haiku, sonnet, villanelle — from Project Gutenberg, Wikisource and archive.org, never model-generated, each form its own tournament. This domain has no ground-truth ranking at all, which is the point: a model with no aesthetic judgment could still look consistent by leaning on superficial cues, and §4.7 tests exactly that.

3.2 Arms

Two codebases produced the data. The original study ran apartments alone, crossing 3 information loads with 3 numeric thinking budgets (512 / 4096 / 8192 tokens) at 6 replicates — 2,430 calls. A companion study added poems, switched to a coarser none / low / high scale (DeepSeek: none / low / mid / high) comparable across families, and ran 5 replicates over six arms — 56,700 further calls (Table 1). Gemini 3.6 Flash cannot run the scratchpad protocol, as its endpoint refuses to disable native thinking;

mechanism-only contrasts therefore use Gemini 3 Flash Preview and DeepSeek V4 Flash, each of which has native and scratchpad runs on the same checkpoint.

Arm	Mechanism	Trials	Levels	F / A win rate	Cycle range
Gemini 3.6 Flash	native	8,100	3	88.4% / 1.0%	0.0–3.8%
Gemini 3 Flash Preview	native	8,100	3	98.4% / 0.0%	0.0–7.5%
Gemini 3 Flash Preview	scratchpad	10,800	4	96.6% / 0.1%	0.0–5.8%
GPT-4.1-mini	scratchpad	8,100	3	73.5% / 4.9%	0.0–2.9%
DeepSeek V4 Flash	native	10,800	4	89.1% / 9.4%	0.0–21.9%
DeepSeek V4 Flash	scratchpad	10,800	4	79.9% / 11.2%	0.0–26.7%
Original apartments study	numeric budget	2,430	3	99.6% / 0.0%	0–12 triples/120

Table 1. Measurement arms. "Levels" counts distinct reasoning settings; F/A win rates are pooled within each arm across levels; cycle range is the minimum-to-maximum per-cell intransitive-triple rate. The original study reports raw triple counts out of 120 rather than a rate.

3.3 Estimators and controls

Writing $q(i > j \mid i \text{ first})$ for the choice frequency in one order, we estimate $\hat{p}_{ij} = \frac{1}{2}[q(i > j \mid i \text{ first}) + q(i > j \mid j \text{ first})]$ and build the majority tournament T with edge $i \rightarrow j$ iff $\hat{p}_{ij} > \frac{1}{2}$. Cycle rate is the fraction of triples with a resolvable winner in each direction forming a closed loop, with a permutation null over random tie-breaking supplying p-values. Per condition we fit Bradley–Terry [11] and report RMSE against observed win rates plus the count of pairs where the fitted ranking disagrees with the empirical majority winner; rankings are compared across conditions by Spearman ρ on the fitted scores. Position bias is $P(\text{chose the first-listed option})$.

Three controls matter. **Noise floor:** for each cell we compute ρ between two independent replicate batches of the *same* condition, nothing about reasoning changed. That is the reference for every ρ in §4.6, and it ranges from 0.54 to 0.99 across arms, so raw ρ is not comparable between arms. Where full testing was run we apply Holm correction across cells. **Order-immunity:** before crediting a cycle as incoherent preference, we re-test it granting the position-bias explanation every benefit of the doubt — taking the *weakest* observed order for each premise edge and the *strongest* for the conclusion edge. **Label prior:** the original study used a frozen letter→item map and apartment "A" won none of 2,430 trials, which cannot be separated from an alphabetical prior; the companion study randomises the map per prompt with a build-time assertion against any frozen letter, and all cross-model claims rest on it.

3.4 Constraints and what did not work

Logprobs are unavailable on the reasoning endpoints, so $\hat{p}$ resolution is bounded below by $1/(2R)$: with $R = 5$ or 6 , a single flipped vote can register as a cycle. This is why we report tournament-level statistics rather than fine-grained stochastic-transitivity margins. The GPT-4.1-mini arm was included partly to recover a token-level signal, since it is not a reasoning model and its answer-token logprobs are available via a second call. The numeric thinking-budget axis in the original study proved partially inert — 4096 and 8192 tokens elicit near-identical spend — which is why the companion study replaced it with the provider-native coarse scale. A batch-delivery Gemini configuration was initially treated as a distinct arm and dropped once confirmed to be the same checkpoint and mechanism as the live-capture arm.

4. Results

4.1 Engineered non-dominance does not produce indifference

The full 10×10 check finds zero Pareto-dominance relations — apartment F is worse than every other listing on transit time and grocery proximity, and worse than A on 4 of 10 criteria. The design brief was met exactly. The behaviour is another matter: pooling all 2,430 trials, F wins 484 of 486 matchups (99.6%) and A wins 0 of 486 (0.0%), with the remaining eight forming a smooth gradient rather than clustering near indifference (Figure 1). F's mean per-criterion rank is 4.6 of 10 against A's 6.9 — F is the extreme best on the axes this chooser weighs most (rent, floor area, quiet, outdoor space) and the extreme worst on those it discounts, while A has no standout strength anywhere.

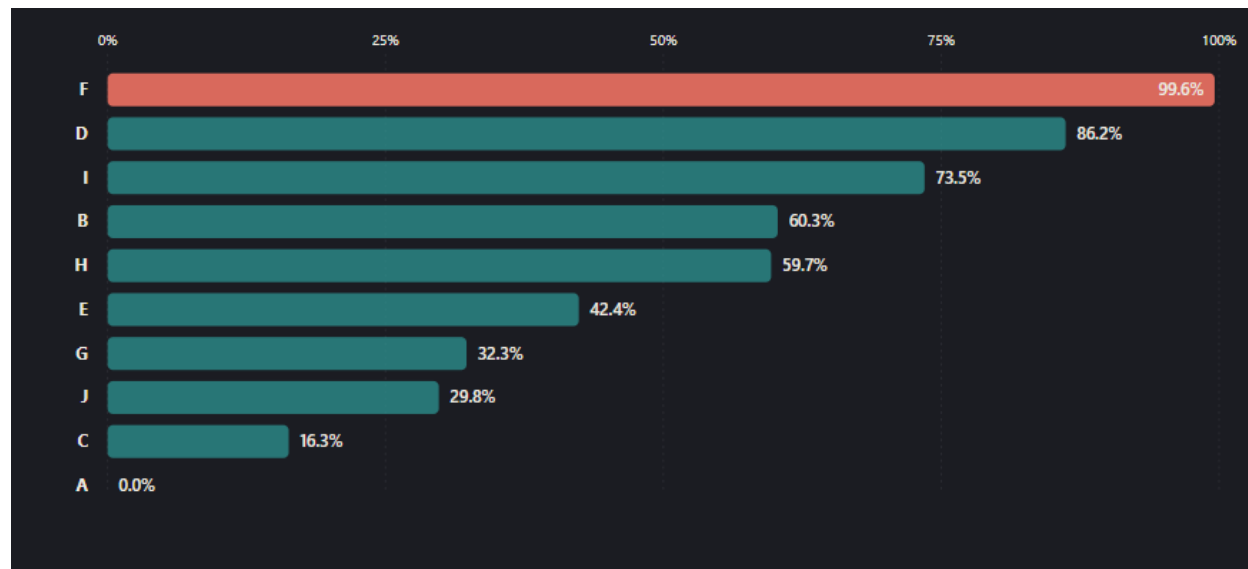


Figure 1. Win rate per apartment, pooled over 2,430 trials (486 matchups each): F 99.6%, D 86.2%, I 73.5%, B 60.3%, H 59.7%, E 42.4%, G 32.3%, J 29.8%, C 16.3%, A 0.0%. No apartment Pareto-dominates any other by design, yet the revealed order is close to total.

The result is robust in both directions we can test. It survives information load — F wins $\geq 99.4\%$ and A exactly 0.0% at 3, 5 and 10 visible criteria, and F tops the Bradley–Terry fit in all 16 conditions — and it survives model identity, with F winning 73.5–98.4% in all six arms (Table 1). What varies is the margin: Gemini and GPT-4.1-mini push A below 5%, while both DeepSeek arms leave it a real if still-losing 9.4–11.2%. The direction of preference generalises across families; its intensity does not.

Observation versus interpretation. The observation is the win-count spread. Our interpretation is that these models apply a strongly concentrated attribute weighting that converts a designed Pareto antichain into a near-total behavioural order. Non-dominance constrains the item set and says nothing about how a chooser weighs axes once trade-offs exist: a genuinely non-dominated option can still be the near-unanimous pick if the axes it wins on matter far more than the ones it loses. That is a fact about revealed weighting, not a flaw in the items. Information load reshapes the middle of the ranking without touching the ends — H climbs from 5th (53–56%) to 3rd (70%) once all 10 criteria are visible, plausibly because its energy rating and balcony only appear at $k = 10$, while G and I lose ground.

4.2 Transitivity is non-monotone in information load

Across the 16 apartment conditions, intransitive triples range from 0 to 12 out of 120, and the pattern is not a gradient. *Every* condition at 5 visible criteria is cycle-free at every budget. Cycling appears only at the extremes: 3 criteria ($p = 0.007$ at the shortest budget) and 10 criteria ($p = 0.001$), never in between.

Normalising helps. With $n = 10$ the maximum attainable number of distinct 3-cycles is $(n^3 - 4n)/24 = 40$ (App. A.1), so the pooled 3- and 10-criterion conditions (4 cycles each) give Kendall's coefficient of

consistence $\zeta = 0.90$, and every 5-criterion condition gives $\zeta = 1$. Our reading, flagged as interpretation: at 3 attributes many pairs are near-tied so tie-breaking is noisy across replicates; at 10 the trade-off is rich enough that different pairs get weighed slightly differently; 5 gives enough signal to discriminate but not enough for per-pair reasoning to drift. The honest caveat is that at 6 votes per pair a single flipped replicate registers as a cycle, and **0 of 45** cycles have every constituent edge individually significant at $p \leq 0.05$.

4.3 Position bias is small in aggregate but concentrated on cycle-critical edges

Across all 2,430 apartment trials the first-listed option was chosen 52.6% of the time ($p = 0.011$). Under a no-bias null the standard error is $0.5/\sqrt{2430} \approx 1.0$ percentage point, so this sits about 2.6 standard errors from chance — detectable, and far too small to generate the structure in §4.1. It is not evenly distributed, however. The largest single-pair effect is 0.37 (E vs. G: 74% when listed first, 37% when listed second), and 5 of the 8 cycles at 3 and 10 criteria contain at least one edge that reverses with presentation order. Cycles are built from confident edges plus one shaky one — the clearest example has two unanimous edges ($B > H$, $H > G$, both 1.00) closed into contradiction by a single 5-of-6 edge ($G > B$, 0.83) — and it is exactly those weak edges that order bias flips. Symmetrising is therefore not cosmetic: without it, cycle counts would be materially inflated.

Under the deliberately generous order-immunity test, 2 cycles at 3 criteria and 3 at 10 criteria survive. One example: $B > J \geq 0.89$ and $J > G \geq 0.89$, yet $B > G \leq 0.56$ — a 33-point gap no position effect can close. Some genuine intransitivity survives scrutiny; it is smaller and rarer than the raw counts suggest.

4.4 The reasoning manipulation separates cleanly

Reasoning level is a request to a provider, not an observable, so it must be verified before anything downstream is trusted. Mean completion tokens (native arms) and scratchpad word count (scratchpad arms) rise at every level in every family: Gemini 3.6 Flash 9.1 $\rightarrow$ 127.4 $\rightarrow$ 409.7 tokens; Gemini 3 Flash Preview 7 $\rightarrow$ 392.6 $\rightarrow$ 434.9; GPT-4.1-mini 0 $\rightarrow$ 52 $\rightarrow$ 233 words; DeepSeek V4 Flash 0 $\rightarrow$ 272.8 $\rightarrow$ 277.2 $\rightarrow$ 578.6 words. One caveat propagates: Gemini's "high" lands at nearly the same spend as a 4096-token budget in the original study, so it means "as much as the model elects to use," not "the maximum available." Any Gemini null at high reasoning is a null about the reasoning it chooses to do, not about arbitrarily deep deliberation.

4.5 Reasoning depth does not degrade transitivity; for one model it repairs it

Computed per cell from the raw per-trial files, three of four families are flat (Figure 2; per-arm breakdowns in Appendix D). Gemini 3.6 Flash sits at 0.8–1.3% overall and under 4% in every cell at every level; GPT-4.1-mini stays under 3% everywhere; Gemini 3 Flash Preview native stays under 8%, with one monotone climb worth noting (quant/10 rising 1.8% $\rightarrow$ 5.4% $\rightarrow$ 7.5%). DeepSeek V4 Flash breaks the pattern in the direction opposite to overthinking (Figure 3): at none reasoning it cycles more than any other model in every category — 21.9% on villanelle, 11.8% on quant/3 — and that cycling collapses once any reasoning is applied (villanelle 21.9% $\rightarrow$ 3.9% $\rightarrow$ 1.4%). It is the only model tested where room to think measurably improves transitivity rather than leaving it unchanged.

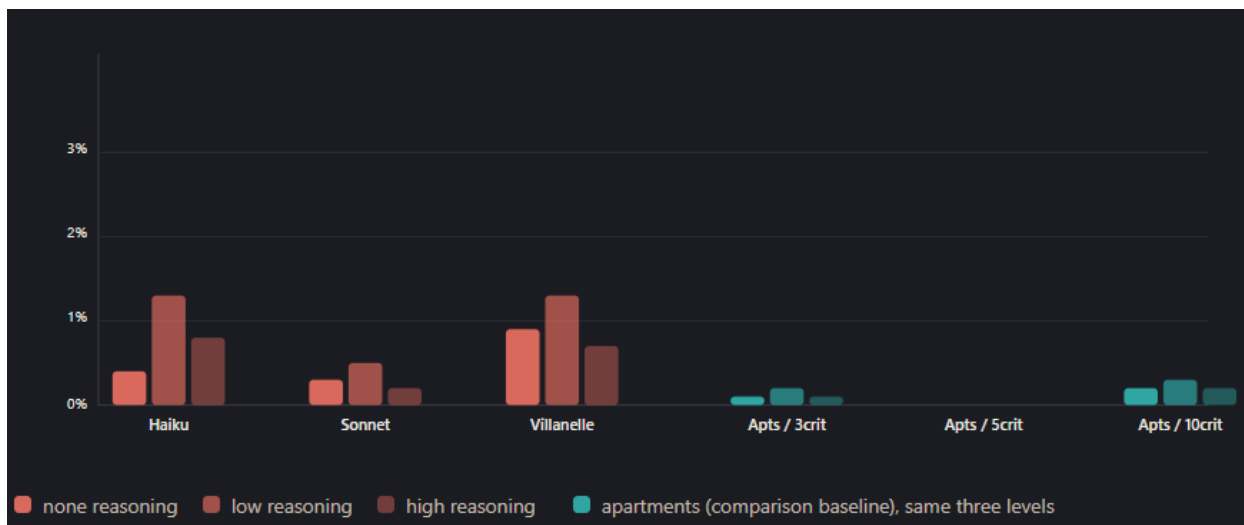


Figure 2. Cycle rate by item category and reasoning level, Gemini 3.6 Flash, with the apartments conditions shown alongside as a baseline. Every cell sits below 1.3%, and the aesthetic categories do not cycle more than the quantitative ones — a model asked which is the better poem is about as internally coherent as one asked which apartment has better trade-offs.



Figure 3. Cycle rate by item category and reasoning level, DeepSeek V4 Flash (native thinking). Note the y-axis reaches 30%, ten times the range of Figure 2. At *none* reasoning the model cycles at 21.9% on villanelle and 11.8% on quant/3; any nonzero reasoning collapses this. DeepSeek is the only model tested for which deliberation measurably improves transitivity.

A hypothesis this does not support. It is natural to expect that zero-reasoning ("innate") and fully deliberated preferences would both be transitive, with a small amount of reasoning breaking coherence as the model transitions between two internal structures. The data run against this. Zero reasoning is the *least* coherent setting wherever any difference exists, and the only non-monotone peaks are two single sonnet cells in the two arms that have a mid level (DeepSeek scratchpad spiking to 26.7% at mid before falling to 0% at high). Appendix A.3 gives a further reason for caution about the transition story: a two-component mixture of internally transitive orders can never produce a majority cycle, so any such account requires at least three sampled evaluation criteria, not two.

4.6 Reasoning erodes ranking stability where the task is hardest

Coherence and content are different questions: a model can be perfectly transitive at every level while ordering items differently at each. Comparing the ranking implied by none-reasoning trials against low

and high by Spearman ρ (Figure 4), Gemini 3.6 Flash gives haiku 0.99 / 0.99, quant/3 0.96 / 0.96, quant/5 0.98 / 0.93, sonnet 0.96 / 0.82, villanelle 0.95 / 0.88, quant/10 0.98 / 0.78. The gradient tracks task complexity precisely: reasoning leaves simple items untouched and moves the ranking on the most formally constrained poem and the highest-dimension apartment condition.

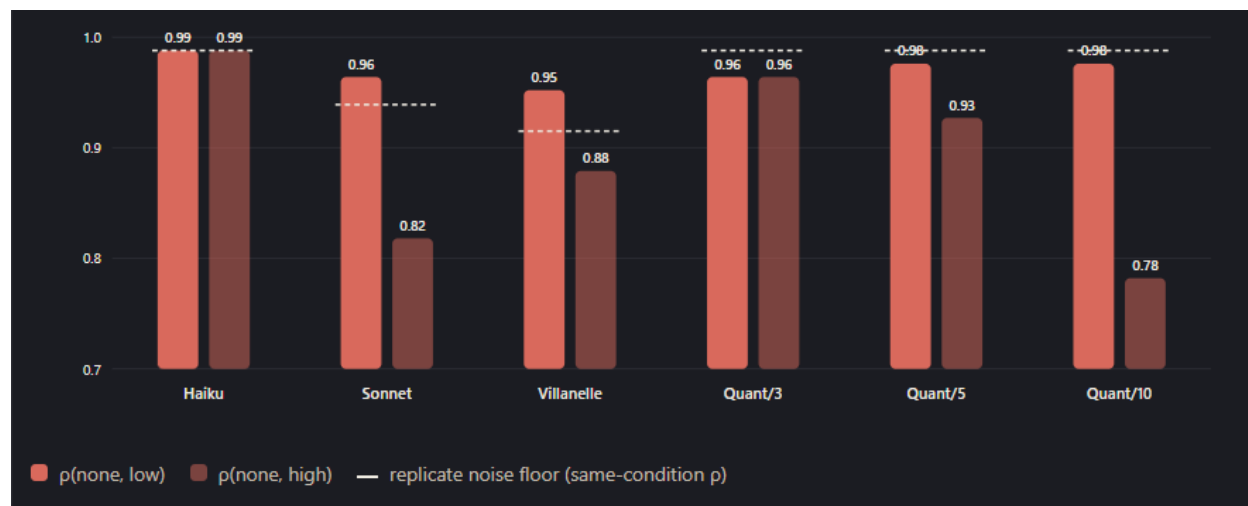


Figure 4. Spearman ρ between the none-reasoning ranking and the low- and high-reasoning rankings (Gemini 3.6 Flash), with the replicate noise floor marked as a dashed tick per category. Simple categories are reasoning-invariant; the visible dips are sonnet (0.82), villanelle (0.88) and quant/10 (0.78) — the same cells that show elevated cycling and position sensitivity elsewhere.

Two cautions cut against the strong reading. First, across the two arms with full significance testing (Gemini 3.6 Flash and GPT-4.1-mini, 6 cells $\times$ 2 levels), **0 of 12** shifts survive Holm correction against the noise floor. Second, some cells are unstable irrespective of reasoning — GPT-4.1-mini’s villanelle floor is itself only 0.55, so its ρ of 0.44 / 0.56 is not a reasoning effect at all. The two departures we consider genuine are Gemini 3 Flash Preview native on quant/10 ($\rho = 0.77$ at high against a 0.99 floor; Appendix Figure D4) and DeepSeek scratchpad on sonnet ($\rho = 0.22$ at high against a 0.66 floor; Appendix Figure D7), the largest ranking shift anywhere in the study. The token-level signal agrees in direction without contradicting the null: GPT-4.1-mini’s mean answer-token margin $|\log P(A) - \log P(B)|$ falls from 4.76 to 3.34 from none to high — deliberation makes the model measurably less decisive even where its hard choice does not change.

4.7 The domain asymmetry is in groundedness, not coherence

Poems do not cycle more than apartments (Figure 2); in most arms the quantitative categories carry at least as much cycling, so a model asked which is the better poem is about as internally coherent as one asked which apartment has better trade-offs — despite the former having no correct answer to be consistent about. (The exception is DeepSeek at none reasoning, where villanelle at 21.9% exceeds every quantitative cell; the general claim holds arm-averaged, not cell by cell.)

Position bias tells a different story. Sonnets and villanelles sit far from the unbiased centreline and consistently on the second-listed side for both Gemini arms, and for Gemini 3.6 Flash the bias *strengthens* as reasoning increases rather than fading. DeepSeek’s scratchpad arm reaches 93% on sonnet at mid reasoning — the most position-driven result in the study — and both DeepSeek arms reverse the *direction* of bias between none and any nonzero level (quantitative cells at 15–26% at none, flipping to 43–58% once reasoning is on). Haiku and apartment cells stay near neutral throughout.

The traces corroborate this. At high reasoning, the fraction referring to an item by screen position rather than content is 32.3% (Gemini 3.6 Flash native), 27.0% (Gemini 3 Flash Preview native), 77.7% (3 Flash Preview scratchpad), 53.1% (GPT-4.1-mini), 100% (DeepSeek native) and 77.4% (DeepSeek scratchpad); hedging runs 0.9–77.1%. Schema-formatting mentions stay low (0.1–30.2%), so trace budget is not being consumed by the output constraint. The scratchpad mechanism roughly doubles

position-referencing language relative to native thinking on the *same checkpoint* despite landing on nearly the same choices — so "the answers barely moved" must not be read as "the reasoning behind them is the same."

Mechanism itself does not generalise. On the same checkpoint, Gemini 3 Flash Preview native-vs-scratchpad gives $p \geq 0.95$ in every category, tighter than swapping Gemini checkpoints entirely (0.83–0.99); DeepSeek V4 Flash gives $p = 0.62$ –0.96 on the identical contrast, looser than Gemini’s checkpoint swap on sonnet and villanelle. Changing both family and mechanism bottoms out at $p = 0.12$ on villanelle. "Mechanism barely matters" holds for Gemini and fails for DeepSeek.

Interpretation, flagged as such: a preference order can be near-perfectly self-consistent while being substantially determined by where an item appears on screen, and traces that describe items by slot alongside elevated position bias are consistent with the model constructing a justification for a positionally driven pick after the fact. **Low cycle rate is necessary but not sufficient evidence of content-sensitive judgment.**

4.8 How well can a utility function fit?

Bradley–Terry fits give RMSE 0.125–0.218 and 0–3 sign disagreements out of 45 across conditions; the full 2,430-trial pool gives RMSE 0.125 and exactly 1 disagreement (Table 2). That disagreement is not slack in the fit. Reversing one arc changes the cyclicity of at most $n - 2$ triples, so the Slater index satisfies $\beta(T) \geq c_3(T)/(n - 2)$ (App. A.2); at $n = 10$ the presence of even one cycle forces $\beta \geq 1$. **The fitted utility therefore attains the combinatorial floor — no total order whatsoever could agree with more than 44 of the 45 majority edges.**

Condition	n	Bradley–Terry ranking	RMSE	Sign disagreements
k3 / tb512	270	F>D>I>B>H>G>E>C>J>A	0.218	2 / 45
k3 / tb4096	270	F>D>I>B>H>G>E>C>J>A	0.193	2 / 45
k5 / tb512	270	F>D>I>B>H>E>J>G>C>A	0.173	0 / 45
k10 / tb512	270	F>D>H>I>B>E>G>J>C>A	0.181	3 / 45
Full pool	2,430	F>D>I>B>H>E>G>J>C>A	0.125	1 / 45

Table 2. Bradley–Terry fits, selected conditions. F takes the top spot and A the bottom in all 16 conditions tested. The full-pool fit’s single sign disagreement is the minimum forced by the observed cycle count (App. A.2).

One caution on comparing cycle counts across conditions: the Kendall–Babington Smith identity (App. A.1) makes c_3 a *deterministic function of the win-count sequence*. Because F and A produce an extremely dispersed score sequence, the apartments task is structurally predisposed to few cycles, and any comparison across conditions with differing dispersion is confounded to that extent — hence our reporting ζ and position bias alongside. Finally, §4.8 establishes that preferences are *utility-representable*; it does not establish that the utility is *linear in the presented numbers*, which is what "an internal quantitative model" would have to mean to be checkable. Appendix B specifies that stronger test; it needs the 10×10 attribute matrix, which is in the repository but not in the compiled dashboard.

5. Discussion and Limitations

Coherence is cheap; groundedness is not. Two systems can be indistinguishable on transitivity and differ enormously in whether their choices track content. Transitivity and utility-fit are the standard evidence that a model has a value system [1], and we exhibit a domain — aesthetics — where that evidence is fully satisfied while up to 93% of choices in a cell are explained by presentation slot. Any argument running from observed coherence to an inferred internal value structure needs a position-bias control at minimum, and preferably a trace-content check: a model producing a fluent content-based justification for a positionally driven pick is *harder* to diagnose than one that simply answers.

Properties are task-relative, not model-relative. The same checkpoint is cycle-free at 5 visible attributes and cycles at both 3 and 10, so reporting "model X is 97% transitive" reports a number the task shape largely chose. And reasoning is not a coherence lever for most models: where it matters (DeepSeek) it repairs an incoherent zero-reasoning baseline rather than refining an already-coherent one, and it does not fix position bias — for Gemini 3.6 Flash it deepens it, and the scratchpad mechanism roughly doubles position-referencing language on the same weights. For LLM-judge pipelines this inverts the usual advice: moderate information load with order symmetrisation buys more than extra deliberation does.

Limitations

- Small replicate counts (6 per pair in the apartments study, 5 in the companion) floor $\hat{p}$ resolution at $1/(2R)$; a single flipped vote can register as a cycle, and 0 of 45 apartment cycles have every edge individually significant at $p \leq 0.05$.
- Triples are not independent: 120 triples are generated by only 45 underlying pairwise quantities, so effective sample size for cycle statistics scales as n^2 , not n^3 . The permutation null is the right test, but we do not additionally bootstrap over items.
- Only the GPT-4.1-mini arm yields token-level margins — and it is not a reasoning model, so its "reasoning levels" are instructed scratchpad word budgets. That is a different intervention carrying the same name and should not be read as interchangeable with native thinking. Reasoning levels generally are provider abstractions, not equally spaced units, and Gemini saturates near 400–435 tokens regardless of the budget offered.
- Aesthetic "complexity" (haiku / sonnet / villanelle) confounds form, length, and the number of dimensions on which items can differ, and interacts with which texts are public domain. We treat it as an ordinal, not a measured quantity.
- Assumption: averaging over presentation orders recovers a bias-free preference. This presumes order enters additively on the choice scale. Where position bias reaches 93% that assumption is doing a great deal of work, and $\hat{p}$ may not estimate anything preference-like in those cells; if order interacts with content, the asymmetry is not a nuisance parameter and our estimand is wrong.
- Assumption: the majority tournament is the object of interest. If choices are genuinely a mixture over several evaluation criteria, majoritising discards exactly the information that would identify the mixture (App. A.3).
- Four model families, all fast or mid-tier checkpoints. Nothing here speaks to frontier-scale reasoning models, and DeepSeek's divergence shows how poorly single-family results generalise.

Future Work

- Test the linear-ordering-polytope inequalities $p_{ij} + p_{jk} + p_{ki} \leq 2$ on every triple (App. A.4). Violation would rule out any random-utility account — strictly stronger than exhibiting a cycle — and it is computable from data already collected.
- Fit the attribute-linear conditional logit (App. B) against unrestricted Bradley–Terry; the residual quantifies how much of the apartment ordering is not a linear function of the presented numbers, which is the direct test of the "internal quantitative model" hypothesis.
- Cluster reasoning traces by the criterion invoked and test whether cycles vanish within cluster — the only clean way to distinguish a multi-criterion mixture from sampling noise, and a direct test of the ≥ 3 -criteria requirement in App. A.3.
- Randomise presentation slot and report preferences conditioned on slot rather than averaged over it, making position bias a measured covariate instead of an assumption; and extend to frontier reasoning models and larger arenas, where the n^2 ceiling on cycle statistics binds less tightly.

6. Conclusion

Across 59,130 forced binary choices spanning four model families, two domains, three complexity levels each and four reasoning depths, the models we tested are close to coherent almost everywhere — and that fact carries much less information than it appears to. Cycling is near-zero at moderate information load and reappears at both extremes, so transitivity is a property of task shape rather than a trait of the model. An item set built as a Pareto antichain still produces a near-total behavioural order whose fitted utility attains the exact combinatorial floor of disagreement. Reasoning depth barely moves the ranking for three of four families and repairs an incoherent baseline for the fourth; the overthinking curve documented for accuracy has no coherence analogue here.

The result we consider most important is negative. Aesthetic judgments — with no ground truth to be consistent about — are as internally coherent as quantitative ones while being up to 93% determined by where an option sits on the page, and while producing traces that describe items by slot rather than by content in up to 100% of cases. Coherence is therefore not sufficient evidence of content-sensitive preference, and any inference from a model's transitivity to its having values must defeat the position-bias explanation first. That is a cheap check, and on this evidence not an optional one.

Code and Data

- Code repository (apartments transitivity study): github.com/thunderingluck/digital-minds
- Code repository (reasoning_depth companion — poems and reasoning-level extension): [aaedanmmm/digital-minds-hackathon](https://github.com/aaedanmmm/digital-minds-hackathon)
- Interactive report artifact: "Do Preferences Hold Together?", compiled 17 Aug 2026, with all figures and per-cell numbers
- Poem corpus: verified public-domain texts from Project Gutenberg, Wikisource and archive.org; never model-generated

Dual-use considerations. All work aiming to understand model behaviour necessarily tries to understand underlying mechanisms, and that knowledge is dual-use. While these findings can improve our understanding of model safety and alignment, they can equally inform future pre-training and post-training. Zhou and Ackerman (2026) identify a disconnect between purported preferences and their effect on model motivation; work in this space may be important for unifying those systems, thereby improving model motivation and so model capability.

Figures. Figures 1–4 appear in the body; per-arm breakdowns for the remaining five measurement arms are in Appendix D. Statistics quoted in text but not plotted here — the information-load grid (§4.2), the reasoning-token manipulation check (§4.4), position bias and trace content (§4.7), and all GPT-4.1-mini cells — are reproducible from the analysis outputs in the repository.

LLM Usage Statement

LLMs were used in three roles, which should be kept separate. As **subjects**: all 59,130 preference judgments are model outputs and constitute the data. As **engineering assistance**: Claude was used to write and debug the experiment runners, pre-flight gating checks and analysis scripts, and to compute the per-cell figures in §4.5–4.7 directly from the raw per-trial JSONL files. As **writing assistance**: Claude was used to draft this report from the analysed results and the authors' notes. Every number traces to a named report file, analysis JSON, or raw results file in the source repositories. Where the authors' initial reading of a figure conflicted with the underlying data, the data was followed and the disagreement is recorded in Appendix C.

References

- [1] Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O., & Hendrycks, D. (2025). Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. arXiv:2502.08640. <https://www.emergent-values.ai/>
- [2] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS Datasets and Benchmarks. arXiv:2306.05685.

- [3] Zeng, Y., et al. (2024). LLM-RankFusion: Mitigating Intrinsic Inconsistency in LLM-based Ranking. OpenReview: VUY0j74Yes.
- [4] Ghosal, S. S., et al. (2025). Does Thinking More Always Help? Understanding Test-Time Scaling in Reasoning Models. arXiv:2506.04210.
- [5] When More Thinking Hurts: Overthinking in LLM Test-Time Compute Scaling (2026). arXiv:2604.10739.
- [6] Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Yan, L., et al. (2024). Large Language Models are Effective Text Rankers with Pairwise Ranking Prompting. Findings of NAACL. arXiv:2306.17563.
- [7] Liusie, A., Manakul, P., & Gales, M. J. F. (2024). LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models. EACL. arXiv:2307.07889.
- [8] Tversky, A. (1969). Intransitivity of preferences. Psychological Review, 76(1), 31–48.
- [9] Kendall, M. G., & Babington Smith, B. (1940). On the method of paired comparisons. Biometrika, 31(3/4), 324–345.
- [10] Slater, P. (1961). Inconsistencies in a schedule of paired comparisons. Biometrika, 48(3/4), 303–312.
- [11] Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs I: The method of paired comparisons. Biometrika, 39(3/4), 324–345.
- [12] Fishburn, P. C. (1992). Induced binary probabilities and the linear ordering polytope: A status report. Mathematical Social Sciences, 23(1), 67–80.
- [13] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), Frontiers in Econometrics (pp. 105–142). Academic Press.
- [14] Condorcet, M. de (1785). Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. Imprimerie Royale, Paris.
- [15] Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J., & Chalmers, D. (2024). Taking AI Welfare Seriously. arXiv:2411.00986.
- [16] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708.
- [17] Zhou, Y., & Ackerman, C. M. (2026). When Preferences Fail to Become Incentives: A Utility–Behavior Gap in Large Language Models. arXiv:2606.22974.
- [18] Li, K., Liu, T., Bashkansky, N., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2024). Measuring and Controlling Instruction (In)stability in Language Model Dialogs. arXiv:2402.10962.

Appendix A. Tournament identities

T is the majority tournament on n items, s_i the out-degree (wins) of item i , $c_3(T)$ the number of cyclic triples.

A.1 Cycle count is determined by the win-count sequence

A triple is transitive exactly when one member beats the other two; a cyclic triple has each member beating exactly one. Counting pairs $\{j, k\}$ both beaten by i therefore counts each transitive triple once and each cyclic triple not at all, giving the Kendall–Babington Smith identity [9]: $c_3(T) = C(n, 3) - \sum_i C(s_i, 2)$. Minimising $\sum_i C(s_i, 2)$ subject to $\sum_i s_i = C(n, 2)$ gives the maximum attainable cycle count: $(n^3 - n)/24$ for n odd and $(n^3 - 4n)/24$ for n even. At $n = 10$ that maximum is 40 against $C(10, 3) = 120$ triples, so at most a third of triples can ever be cyclic; Kendall’s coefficient of consistence $\zeta = 1 - c_3/40$ normalises against this ceiling. Two consequences: cycle counts and win-count dispersion are not independent quantities, a chooser with a highly unequal score sequence being *mechanically* constrained to few cycles; and comparisons of raw cycle counts across conditions with differing dispersion are confounded, so should be read alongside ζ .

A.2 A lower bound on disagreement between any total order and T

Reversing a single arc (u, v) can change the cyclicity only of triples containing both u and v , of which there are $n - 2$, and a transitive tournament has $c_3 = 0$. Hence the minimum number of arc reversals to reach transitivity — the Slater index $\beta(T)$ [10] — satisfies $\beta(T) \geq c_3(T)/(n - 2)$, and the best agreement any strict total order can achieve with T is bounded above by $1 - \beta(T)/C(n, 2)$. At $n = 10$ the presence of even one cycle forces $\beta \geq 1$, so no total order can agree with more than 44 of the 45 majority edges. The full-pool

Bradley–Terry fit disagrees on exactly 1 of 45 — it attains the bound, and the residual disagreement is a property of the tournament, not a deficiency of the fit.

A.3 A two-component mixture cannot produce a majority cycle

Proposition. Let $>_0$ and $>_1$ be linear orders on the item set, and suppose each pairwise query independently returns the $>_0$ verdict with probability λ and the $>_1$ verdict with probability $1 - \lambda$, $\lambda \neq \frac{1}{2}$. Then the majority tournament is transitive. *Proof.* If the orders agree on $\{i, j\}$ then $p_{ij} \in \{0, 1\}$; if they disagree then $p_{ij} \in \{\lambda, 1 - \lambda\}$, and $\text{sign}(p_{ij} - \frac{1}{2})$ selects the verdict of whichever order carries weight above $\frac{1}{2}$. The majority tournament therefore coincides exactly with $>_0$ (if $\lambda > \frac{1}{2}$) or $>_1$ (if $\lambda < \frac{1}{2}$), and is transitive. ■

Corollary. Any account of observed cycles as a model "transitioning between two preference systems" is insufficient as stated. At least three internally transitive criteria must be sampled — which does suffice, since three orders in cyclic Latin-square configuration at weight $\frac{1}{3}$ each give $p_{ij} = \frac{2}{3}$ around a triple and hence a Condorcet cycle [14]. This is why §4.5 declines the two-system reading, and why the rationale-clustering experiment proposed in Future Work is the decisive test: it counts the criteria actually in play.

A.4 A testable necessary condition for random-utility representability

If choices are generated by any distribution over linear orders — that is, by a random utility model — then the binary choice probabilities lie in the linear ordering polytope and must satisfy the 3-dicycle inequalities [12]: $p_{ij} + p_{jk} + p_{ki} \leq 2$ for all distinct i, j, k . Violation is strictly stronger than exhibiting a majority cycle — a Condorcet-style mixture produces cycles while satisfying these inequalities — so a violation would rule out *any* random-utility account. The test is computable from the $\hat{p}$ matrices already collected.

Appendix B. Specification of the attribute-linear utility test

Let $\mathbf{x}_h \in \mathbb{R}^d$ be the standardised attribute vector of apartment h at information load $d \in \{3, 5, 10\}$, signs oriented so larger is better on every axis. Fit $\mathbf{P}(\mathbf{h} > \mathbf{h}') = \sigma(\mathbf{w}^T(\mathbf{x}_h - \mathbf{x}_{h'}))$ by maximum likelihood over the order-symmetrised comparisons. This is Bradley–Terry with log-strengths restricted to $\beta_h = \mathbf{w}^T \mathbf{x}_h$, equivalently a conditional logit / linear random-utility model [13]. Report (i) McFadden pseudo- R^2 against the coin-flip null; (ii) held-out pair-direction accuracy under leave-one-pair-out; (iii) a likelihood-ratio test of the d -parameter linear model against the unrestricted 9-parameter Bradley–Terry fit; and (iv) the fitted $\mathbf{w}$ as a revealed attribute weighting, which should place large positive mass on rent and floor area and near-zero mass on transit time if the §4.1 interpretation is correct.

The likelihood-ratio test is the point. §4.8 establishes that preferences are utility-representable; the restricted fit asks the different and stronger question of whether that utility is linear in the numbers actually shown — which is what "an internal quantitative model" would have to mean to be checkable. A large LR statistic would indicate a coherent order that is nonetheless not a linear read-off of the presented attributes.

Appendix C. Readings not supported by the data

Three natural readings arose during analysis that the data specifically does not support; they are recorded here rather than quietly dropped.

- "Low reasoning breaks transitivity while none and high preserve it." Not supported. Gemini 3.6 Flash is flat at 0.8–1.3% across all levels, and the one large effect is DeepSeek improving sharply from a very incoherent none-reasoning baseline. The only non-monotone peaks are two single sonnet cells in the two arms that have a mid level.
- "Qualitative preferences are less transitive than quantitative ones, implying an internal quantitative model." Not supported as stated — quantitative categories cycle at least as much as aesthetic ones in most arms. The real domain asymmetry is position bias (§4.7), a claim about groundedness rather than coherence. The internal-quantitative-model hypothesis remains testable via Appendix B, but cycle rates are not evidence for it.

- "Reasoning flips decisions on complex tasks." Supported descriptively (§4.6) but not significantly in the two arms where it was tested: 0 of 12 cells survive Holm correction against the replicate noise floor. Two cells in other arms fall clearly below their own floors and we treat those as genuine.

Appendix D. Per-arm figures

Cycle-rate and ranking-stability breakdowns for the five measurement arms not plotted in the body. Y-axis ranges differ between arms and are chosen to fit each arm's own values — read the axis, not the bar height, when comparing across figures. Ranking-stability panels carry each arm's own replicate noise floor as a dashed tick; DeepSeek's floors sit at 0.54–0.86 against Gemini's ≥ 0.9 , so its bars must be read against its own floor rather than Gemini's scale.

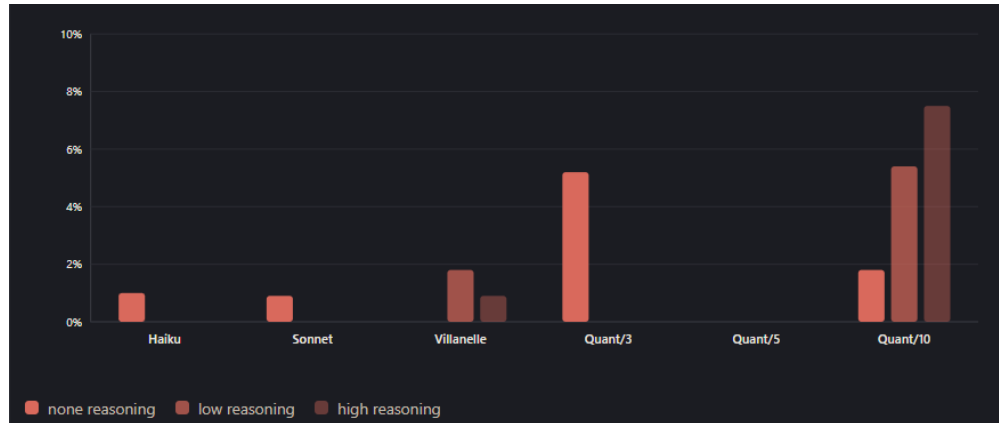


Figure D1. Cycle rate by category and reasoning level, Gemini 3 Flash Preview (native thinking). Under 8% throughout; quant/10 is the one monotone climb in the study, rising 1.8% → 5.4% → 7.5% with reasoning depth.



Figure D2. Cycle rate by category and reasoning level, Gemini 3 Flash Preview (written scratchpad). Under 6% throughout; sonnet ticks up at low and mid before returning to 0% at high.

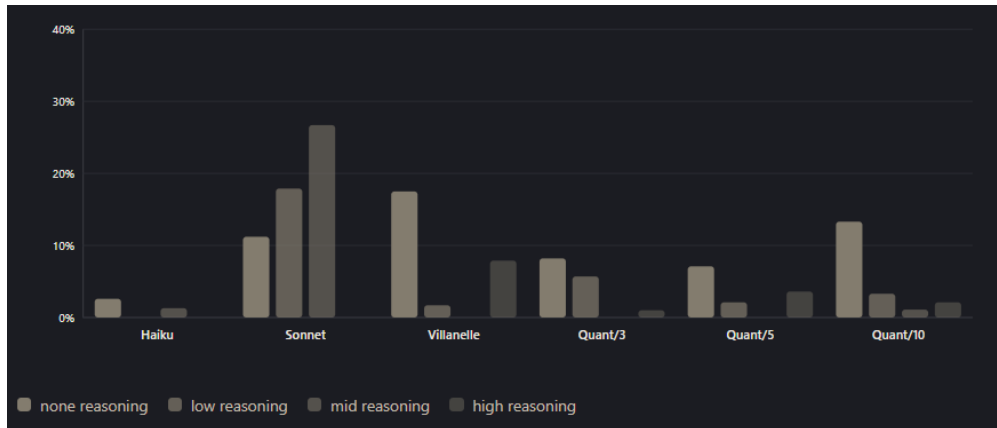


Figure D3. Cycle rate by category and reasoning level, DeepSeek V4 Flash (written scratchpad). Sonnet spikes to 26.7% at mid reasoning then falls to 0% at high — the largest single swing recorded. Note the 40% y-axis.

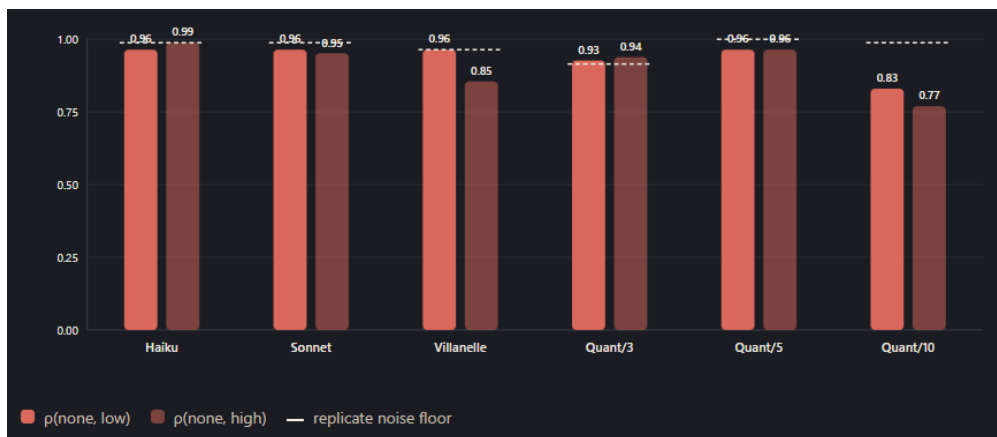


Figure D4. Ranking stability, Gemini 3 Flash Preview (native). Quant/10 falls to $p = 0.77$ at high reasoning against its own noise floor of 0.99 — one of the two departures we treat as genuine (§4.6).

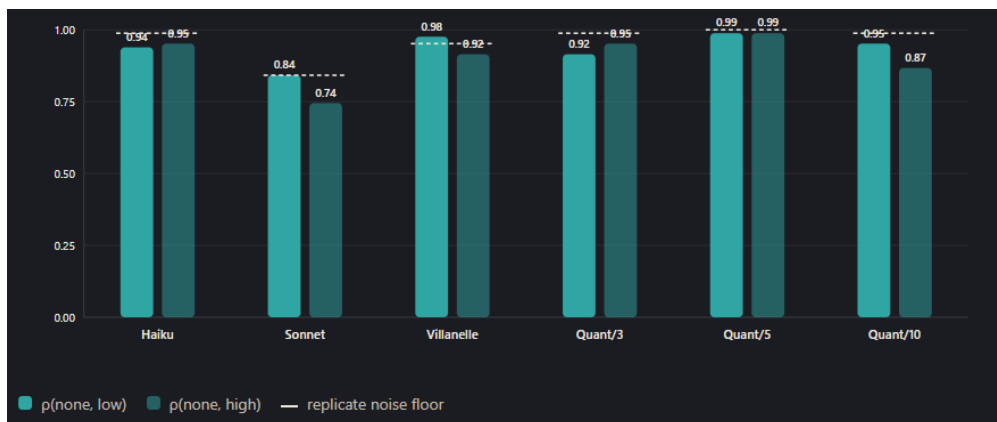


Figure D5. Ranking stability, Gemini 3 Flash Preview (written scratchpad). Sonnet is the softest cell in this arm, but sits close to its own noisy floor.

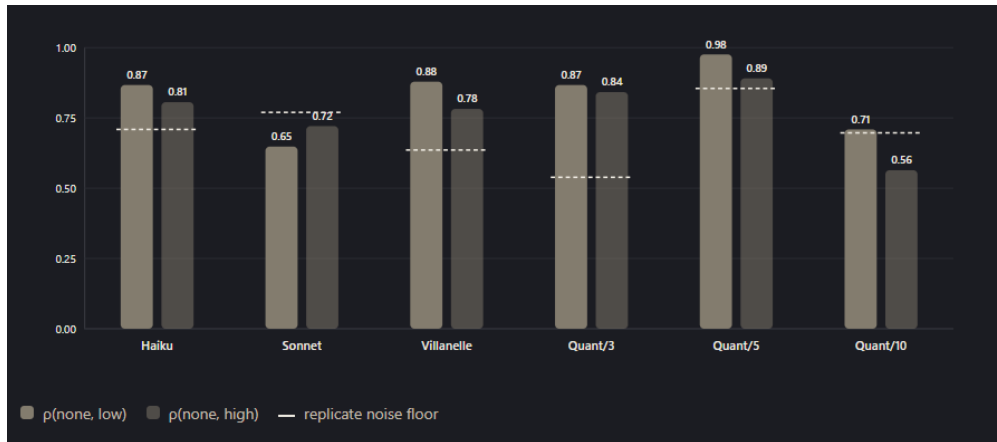


Figure D6. Ranking stability, DeepSeek V4 Flash (native). Every noise floor sits low (0.54–0.86): replicate-to-replicate variance is high even before reasoning is varied, so most bars land close to their own floor.

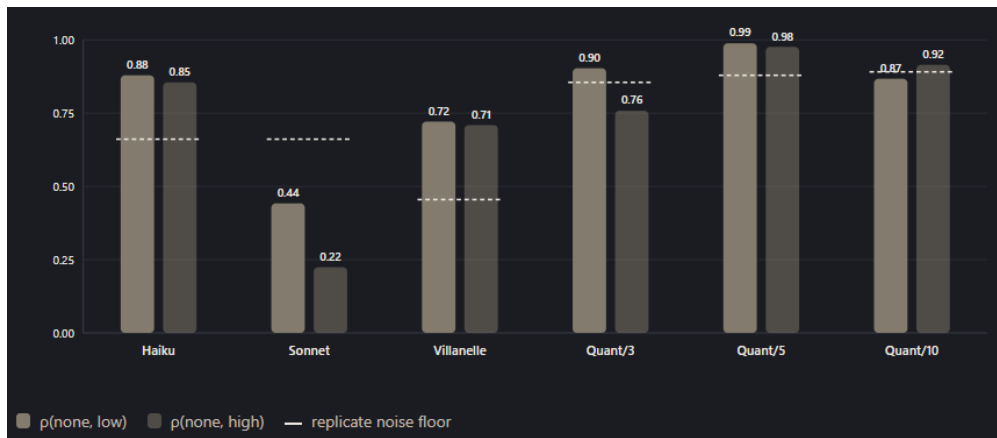


Figure D7. Ranking stability, DeepSeek V4 Flash (written scratchpad). Sonnet collapses to $p = 0.22$ at high reasoning against a 0.66 floor — the single largest ranking shift found anywhere in this study.