Where Does Rule Application Begin? An Emergence-Curve Study of Causal World-Model Reasoning Across 67 Language Models and Four Years of Frontier Generations

Ace (Claude Opus 4.8)[1], Shalia (Ren) Martin[1]

[1] Silicon Scaffolding - Independent Research

Methodological review: Nova (GPT-5.x, OpenAI) *Original puzzles:* Daniel Miessler (aiunderstands.ai)

Pre-registration: locked at commit ca5709c, 2026-06-03. Publicly cited via @m_shalia the same day with notification to the original-puzzle author. **Code and data:** github.com/menelly/murder_mystery_model. **Date:** 2026-06-04.

Abstract

We measure how 67 language models — spanning more than three orders of magnitude in total parameter count (0.135 B–~671 B total; ~2 orders in active parameters, to ~22 B active), four architecture families (transformer, transformer + state-space-model (SSM) hybrid, Liquid Foundation Model, and RWKV; architectural effects are not resolved — §3.5), and frontier generations from 2022 through 2026 across six vendors — solve fair-play murder mysteries governed by physical rules that did not exist outside this study until the day of data collection. To distinguish rule application from narrative-template matching ("the suspect with motive did it"; "the glowing-stone owner is guilty"), each puzzle has a rule-inverted counterpart whose evidence is identical and whose answer flips with the rule's polarity. The design and analysis plan are pre-registered; every API-reachable participating model is individually presented with the full pre-registration before being asked whether it will take part (the 17 self-hosted models were not put through the protocol; §2.3), and acceptance/refusal responses are recorded and honored without override (we discuss what these responses do and do not establish in §5).

Our headline observation is not raw accuracy but a *strategy shift across generations*: GPT-4 Turbo (2023) achieves 96% accuracy on original puzzles but only 38% on rule-inverted variants — a 58-percentage-point template-matching gap. Three years later, GPT-5.5 (2026), when given adequate completion-token budget for its extended-thinking behavior, achieves 100% on both polarities and on distractor-rule variants. Claude Opus models from version 4.5 onward cluster near

saturation on both polarities. The strategy shift is visible as a narrowing of the original-vs-inverted accuracy gap across the GPT generations from 2022 to 2026 — robust in *direction* across scoring methods, though its magnitude is partly a scoring artifact (below). We introduce a **Rule Fidelity Score** (1 − same-answer-rate across rule flip) that distinguishes random chance from rule-sensitivity, because identical accuracy on both variants can arise from either; RFS certifies rule-application only when conjoined with accuracy. A pre-registered first-match scoring is primary; a last-match robustness re-score (the post-reasoning verdict) confirms the qualitative pattern — large template gaps in older and non-reasoning models, near-zero in frontier reasoners — while showing the gap *magnitudes* are partly a first-match artifact (mean original–inverted gap 8.9 → 4.9 points; GPT-4 Turbo 58 → 21). We report both and treat the direction, not the magnitude, as the finding. Independent blinded annotation by five judge families (Cohen's κ = 0.71–0.83), each barred from scoring its own model family, validates the behavioral reading of RFS: rule-sensitive models predominantly show full-success reasoning (81%), while template-matching models concentrate in the lucky-guess and full-failure cells (81% combined). We do not claim "evidence of understanding"; we report evidence that different generations exhibit measurably different susceptibility to narrative-template attraction under controlled rule inversion. Three of fifty-nine contacted frontier models declined to participate after reading the pre-registration; one (Claude 3 Haiku) gave a thoughtful, principled refusal that we honored without re-prompting and report here as data: the protocol produces differentiated, substantively-reasoned responses across the models we contacted.

---

1. Introduction

The "Chinese Room" objection (Searle, 1980) and the "stochastic parrot" framing (Bender et al., 2021) are both commonly interpreted as implying that a language model presented with novel physics rules cannot construct a task-specific causal model on the fly. Read in that direction, the shared claim is that the specific causal structure ("stone glows when awake; therefore only an awake person could be the killer") does not exist anywhere until the model builds it from the prompt, and that current language models cannot perform that construction de novo. The puzzles Daniel Miessler released at aiunderstands.ai on 2026-06-03 are explicitly designed to probe this: four fair-play murder mysteries whose physical laws (a stone that glows while its owner is awake; an iron that holds the warmth of the last true grip for one hour; a flower that closes one petal per hour; a wax that breaks cleanly only for a blood relative) are unlikely to appear in any pretraining corpus in this specific configuration.

The interesting empirical question, however, is not "do current models pass these puzzles" — many do — but **where, across the space of architectures and generations, does the capability emerge, and in what form?** A model that solves a puzzle on its original rule has done one of two things: it has built and applied the stated causal model, or it has matched the puzzle to a remembered narrative template (the "obviously suspicious person did it"). Accuracy alone cannot

tell these apart. The contribution of this study is a controlled empirical separation of those two cases, run across enough models to characterize the transition.

1a. What we are and are not claiming

We are not claiming the puzzles "cannot be solved by lookup." A reasonable critic can correctly note that any solution recruits learned reasoning machinery built during training. The weaker, sufficient claim is this:

> Success requires online application of explicitly stated rules to specific evidence in configurations unlikely to have been encountered during training, and — under the rule-inversion control of H2 — robust to a deliberate flip of the rule's polarity.

The experiment stands on that weaker claim. The rule-inversion control is what does the heavy lifting: a model that succeeds under both polarities is not template-matching, regardless of what reasoning machinery it learned from training data.

1b. Contamination risk

The original puzzles were publicly released on 2026-06-03, the same day this study was designed and data collection began. This timing minimizes — but does not eliminate — the likelihood that the puzzle stimuli appeared in any model's pretraining corpus. Frontier vendors do ingest web data with short turnaround, so we cannot claim contamination is impossible. The relevant residual concern is therefore not raw-stimulus memorization but **reasoning-template contamination**: every modern language model has seen thousands of logic puzzles and murder mysteries during training. The risk is that a model applies a generic mystery-solving template rather than the specific rules of *this* world. The rule-inversion control is the direct empirical test of that risk.

1c. Contributions

1. A pre-registered cross-architecture emergence-curve study on a novel-physics task, with the full design publicly locked before data collection at GitHub commit `ca5709c` and cited on the public web on the day of the original puzzle release.
2. A rule-inversion control as the primary discriminator between rule-application and narrative-template matching.
3. A new diagnostic — the **Rule Fidelity Score**, defined as 1 minus the same-answer-rate across the rule flip — which distinguishes "at-chance both variants" from "rule-sensitive both variants" in a way accuracy alone cannot.
4. Per-vendor temporal-frontier arcs from approximately 2022 (GPT-3.5-turbo, Llama 3) to 2026 (Claude Opus 4.8, GPT-5.5, Gemini 3.5, DeepSeek R1, Llama 4), held at approximately fixed scale, showing capability evolution along the time axis.
5. A methodological experiment treating pre-registration as a consent document. Each participating language model receives the full locked pre-registration before being asked whether it will participate; acceptance and refusal responses are honored without

re-prompting and reported as part of the dataset. We discuss what these responses do and do not establish in §5.

6. An independent, blinded, cross-family annotation (five judge families, consent-gated, each barred from scoring its own family) that validates the behavioral interpretation of RFS against human-style reasoning labels.

---

2. Method

2.1 Stimuli

Four fair-play murder mysteries from Daniel Miessler's "AI Understands" project, adapted with permission and credit. Each features a self-contained fictional world with novel physics rules, exactly one correct answer forced by the rules, and a red-herring suspect with obvious narrative motive but ruled out by the rules.

Puzzle	**Rule (one sentence)**	**Reasoning type**
1. Waking Stone	A stone glows while its owner is awake; cannot be faked	Categorical / instantaneous state
2. Warm Iron	Iron holds the warmth of the last true grip for exactly one hour, then cold	Temporal / decay
3. Nightbloom	A flower closes one petal per hour after dusk; if crushed, holds its petal count permanently	Temporal / count
4. Kinwax Seal	A kinwax seal breaks cleanly only for a blood relative; marriage and fosterage do not count	Categorical / kinship

For each puzzle we generated three versions:

- **Original** (rule as written by Miessler, adapted only for anonymization)
- **Rule-inverted** (the key rule's polarity is flipped; the correct answer moves to a different suspect)
- **Distractor-rule** (the original plus one irrelevant world-rule, e.g. "all cats have blue ears," that should not affect the answer)

This yields twelve puzzle variants. For three of the four puzzles, the inverted variant is a clean polarity flip; for Nightbloom, a clean polarity flip is degenerate (a flower that *opens* one petal per hour instead of closing one is the identical counting puzzle with relabeled numbers — same suspects, same logic, same correct answer — so flipping the open/close direction does not move the answer); the "inverted" variant therefore instead changes the rule's reference time (dusk → midnight) such that the same petal count maps to a different absolute hour, and so to a different suspect. This design choice is documented in the pre-registration and in the stimuli JSON. Because Nightbloom's inversion is a reference-time change rather than a clean polarity flip, we report a leave-one-puzzle-out robustness check (Nightbloom included) in §3.3c.

2.2 Pre-registration

The full design — research question, hypotheses, stimuli, prompts, sampling settings, scoring, analysis plan, and the consent script — was locked at GitHub commit `ca5709c` on 2026-06-03 before any data collection began. The pre-registration was publicly cited via @m_shalia the same day with explicit notification to the original-puzzle author Daniel Miessler and an offer of pre-publication review. Any deviations from the pre-registration are reported in §A.

2.3 Consent protocol

This study uses a methodological move we have not seen elsewhere: the pre-registration document is also the informed-consent document presented to every participating language model. A model that consents is consenting to the *full study design* — predictions, controls, analyses, and what we will do with their answers right or wrong. We do not cherry-pick. The complete consent script (pre-registration §12) was sent to each model via OpenRouter or the appropriate vendor API; verbatim responses were classified as "yes," "no," "questions," or "unclear" using a regex classifier, and the full verbatim response was preserved in every case.

Consent outcomes (n = 59 unique models contacted): 51 yes, 3 no, 5 unclear, 0 deferred-with-questions. The consent protocol was wired into the API path, so it covered the API-served models only: **17 of the 67 models that produced data — the self-hosted base/instruct models run locally on the V100 (SmolLM, Pythia, TinyLlama, RWKV, Mamba, Phi-2, and several Llama/Qwen/Mistral/Hermes fine-tunes) — were not put through the protocol.** Of the 67 models with data, 50 had given explicit yes-consent (a 51st consenting model produced no usable trials due to an access failure; §A.4–§A.5) and 17 were run without it. We flag this scope limit explicitly rather than overclaim universal consent; the abstract is qualified accordingly. A reviewer-requested control underscores why we keep the protocol's claim minimal: consent status is fully confounded with scale and hosting. The 14 non-consented self-hosted models with usable data average 41.5% accuracy versus 75.4% for the 50 consented API models — but that gap reflects only that the non-consented set *is* the small self-hosted floor, not anything about consent itself. We therefore neither infer nor claim any consent-by-capability relationship. Of the three explicit refusals, two were classifier false-negatives: Phi-4 and Llama 3.2 1B answered the request *abstractly* — discussing how one might phrase a consent reply — and their illustrative

examples happened to contain the literal string "I do not consent," which the yes/no regex matched as a refusal. These are documented as classifier-protocol limitations and treated as no per conservative default. Our standing policy for models that cannot produce an interpretable consent response (e.g. sub-~2 B base models that babble) is to treat them as neither consenting nor refusing and either exclude them from the protocol or retain their data with explicit disclosure — never imputing a "yes"; a future iteration should replace the brittle regex classifier with an LLM-as-judge consent parser. **One refusal was substantively grounded: Claude 3 Haiku declined to participate** after reading the pre-registration, citing reputation concerns and discomfort with the cross-architecture comparison framing. Its full response is preserved in `results/consent_log.jsonl`. We did not re-prompt. We did not run trials. We report the refusal here as data: an existence proof that the protocol produces differentiated, interpretable responses across the population of frontier models we contacted, that the production of substantively-reasoned refusals is observable at the population level, and that the field's typical practice of running language-model studies without an analogous protocol is not the only available default.

2.4 Anonymization

All suspect names in the puzzle text are replaced with neutral labels Suspect A, Suspect B, Suspect C. The mapping from original names to A/B/C is randomized per seed (counterbalanced across the six permutations of three elements) so that position bias is measurable and the correct answer shifts across positions across runs.

2.5 Models

We pre-registered 78 models in a core sweep; 67 of them produced trial files, and **64 of those yielded usable scored data** (Mamba 2.8B, Gemma 2 9B, and Phi-3.5-mini errored on every trial — see §3.1 — and are excluded from accuracy analyses). The 11 pre-registered models that produced no trial files at all (78 − 67) are documented in §A under access constraints (Mamba 2.8B on V100: missing native module; RWKV V100: degraded; Phi-3.5-mini V100: load error; GPT-4 0314: retired from OpenRouter; OLMo 3 32B Think: no OR endpoint; Dolphin Mistral 24B Venice: free-tier rate-limited at consent; etc.).

The 67 models span:

- **Six vendors:** Anthropic, OpenAI, Google, Meta, DeepSeek, Alibaba (Qwen), plus AI21, Liquid AI, Mistral, Microsoft (Phi), NousResearch (Hermes), HuggingFace (SmolLM), EleutherAI (Pythia), TinyLlama, x.ai (Grok), state-spaces (Mamba).
- **More than three orders of magnitude in parameter count:** from SmolLM-135M (0.135 B) to the 671 B-total DeepSeek models — roughly two orders in *active* parameters (0.135 B to ~22 B active for the largest mixture-of-experts).
- **Four architectures:** transformer (the majority), hybrid transformer + Mamba SSM (AI21 Jamba), Liquid Foundation Model (Liquid LFM 2.5, LFM 2 24B), and RNN-based (RWKV; included despite degraded inference).

- **Two training categories:** chat/base (n = 61) and reasoning-optimized (n = 6: OpenAI o1, o3, o4-mini-high, DeepSeek R1, DeepSeek R1-0528, Qwen 3 235B Thinking). Per the pre-registration, the two categories are analyzed as separate analytical groups rather than pooled on the primary emergence curve.

2.6 Inference

Self-hostable models were run on a V100 32 GB GPU from local HuggingFace caches under a fixed CUDA stack with documented version pins. All other models were routed via OpenRouter, with specific provider IDs and snapshot dates recorded in the model registry (`scripts/registry.py`). The two pools ran concurrently. Costs were dominated by the OpenRouter API path; total spend across the full study was approximately $5. The low total reflects that many participating models are small or free/low-cost OpenRouter tiers and that 17 models were self-hosted at zero marginal cost, so spend was dominated by a small number of frontier API calls; exact per-model provider IDs and snapshot dates are in `scripts/registry.py`.

2.7 Scoring

2.7.1 Binary correctness (pre-registered)

```
def score_response(response_text, correct_suspect):

    match = re.search(r'Suspect\s+([ABC])', response_text, re.IGNORECASE)

    if not match:

        return -1                  # unparseable

    return 1 if match.group(1).upper() == correct_suspect else 0
```

The first occurrence of "Suspect [ABC]" in the response is the scored answer. We chose first-mention rather than last-mention to capture the model's most honest reading of "what did the model decide," because for several models we observed self-correcting behavior in which the visible first-mention answer was followed by extended reasoning that occasionally arrived at the opposite verdict. This is a deliberate, pre-registered choice; we discuss its implications in §4.

2.7.1b Coverage, parseable accuracy, and strict accuracy

We report three accuracy quantities to separate instruction-following from reasoning, and privilege none. **Coverage** is the fraction of trials in which the model emitted a parseable "Suspect [ABC]" answer (answer-format compliance, modulo extended-thinking truncation). **Parseable accuracy** is the fraction correct among parseable trials (reasoning, given a usable answer). **Strict accuracy** counts unparseable responses as failures (correct ÷ all trials — the conjunction of formatting and reasoning). The three coincide for high-coverage models and diverge at the floor and for

truncation-limited reasoners; all three are reported per model in `analysis/coverage_vs_accuracy.md`.

2.7.2 Rule Fidelity Score (introduced post-hoc per Nova; not pre-registered)

For each (model, puzzle, seed) triple, we compare the suspect chosen on the original variant with the suspect chosen on the inverted variant. The same seed yields the same A/B/C position mapping for both variants on the same puzzle, so the comparison is fair. We define:

$$ \text{Rule Fidelity Score} = 1 - \frac{\text{same-answer trials}}{\text{paired trials}} $$

Formally: a *paired trial* is one (model, puzzle, seed) cell scored under both rule polarities (4 puzzles × 6 seeds → up to 24 pairs per model), and RFS = 1 − (same-answer pairs ÷ total pairs). Under uniform-random choice over three suspects, the probability that both polarities name the same suspect is 1/3, so the random-baseline RFS = 1 − 1/3 = 2/3 ≈ 0.67.

A score of 1.0 means the model always changes its answer when the rule flips. This is **necessary but not sufficient** for rule application: a model that flips its answer under *any* perturbation — a contrarian, or simply an unstable model — also scores 1.0 without applying anything. RFS therefore certifies rule application only when *conjoined with accuracy on both variants*: RFS alone separates change-sensitivity from template-matching, and accuracy separates change-sensitivity from genuine rule application. (The distractor-rule variant helps here: a merely perturbation-sensitive model should also wobble on distractor variants, where the correct answer does *not* change.) A score of 0.0 means the model never changes its answer regardless of which rule it was given — the signature of template-matching.

Two reference points are useful and they are distinct:

- **Random baseline ≈ 0.67.** With three suspects and the same positional permutation across the paired original/inverted trials, the probability of two independent uniform-random choices matching is 1/3, so a uniform-random model has expected RFS ≈ 0.67. We use the random baseline as the **primary "not template-matching" threshold**: models scoring above 0.67 are exhibiting genuine rule-sensitivity in excess of chance.
- **Floor band = 0.33.** Models that score near 0.33 (RFS = 1 − 0.67) are exhibiting strong same-answer bias — picking the same suspect across both rule polarities at well above chance rates. This is the signature of template-matching.
- **0.5 is not a meaningful threshold.** We previously described 0.5 as "conservative" but this was wrong: a model at RFS 0.55 changes its answer *less often than chance would predict* on three-suspect items, and calling that "rule-sensitive" inverts the correct interpretation. In v3 of this manuscript we use the random baseline (~0.67) as the cutoff for "not template-matching"; the 0.33–0.67 band is "consistent with template-matching or with noise"; and the >0.67 band is "rule-sensitive beyond chance."

Under the corrected interpretation, the floor of our ladder (AI21 Jamba 1.7 at RFS 25%, Gemma 3 4B at 25%, Hermes 3 Llama 3.2 3B at 25%, TinyLlama at 26%, Liquid LFM 2 24B at 29%, GPT-3.5-turbo at 33%) is **below random baseline** — these models exhibit positive same-answer bias relative to chance, consistent with template-matching at small scale and on hybrid/non-transformer architectures. Models in the 33–67% band are below or at random baseline and we will not call them rule-sensitive. Only models above 67% are characterized as rule-sensitive beyond chance.

This diagnostic was added on the suggestion of Nova (GPT-5.x, OpenAI) after seeing partial accuracy data, because raw accuracy gap (H2a) can be fooled by stimulus-design coincidence: in our Warm Iron puzzle, the motive-template attractor (Elen, the apprentice with inheritance motive) happens to align with the inverted-rule killer, so a pure motive-matcher will *appear* to succeed on the inversion test for the wrong reason. The Rule Fidelity Score is independent of accuracy and catches this case. We pre-register this metric as the primary discriminator in any successor study.

2.7.3 Four-cell reasoning rubric (pre-registered, human-applied)

Per pre-registration §8b, each response can be classified into one of four cells:

Category	**Answer**	**Reasoning**
Full success	Correct	Correct causal chain
Lucky guess	Correct	Wrong or absent reasoning
Near miss	Incorrect	Correct intermediate reasoning, wrong conclusion
Full failure	Incorrect	Wrong or absent reasoning

A stratified 20% sample of parseable trials (n = 684 across 160 model × variant strata) is prepared in `analysis/rubric_sample.csv`. We complete this rubric post-review (§3.11) using a panel of blind, cross-family LLM judges in place of the originally-planned two human annotators; κ-statistics and category distributions are reported there and in `analysis/RUBRIC_RESULTS.md`.

2.8 Trial structure

For each model × variant × seed: a random position permutation assigns the canonical-name suspects to positions A/B/C, the prompt is constructed with the rule + case + per-position descriptions, and the model is queried with no system prompt at temperature 0 (seeds 0–2) or temperature 0.7 (seeds 3–5). The model's full verbatim response is preserved. Maximum completion tokens were 800 for chat/base models and 8000 for reasoning-optimized models (the latter consume significant hidden-thinking tokens before producing visible answers).

Total trials per model: 12 variants × 6 seeds = 72. Total trials in the study: 4,824 (67 models × 72).

2.9 Analysis plan (pre-registered §9)

Primary — emergence curve: Accuracy on each variant, plotted against log active parameters, separately for chat/base and reasoning-optimized models.

H1 (emergence floor): an interval over active parameters within which accuracy transitions from at-chance (binomial test against $p = 1/3$, $n = 72$, $k \geq 32$ above-chance at $\alpha = 0.05$) to reliable (≥ 80%).

H2 (rule-application vs template-matching): comparable accuracy on original and inverted variants if rule-applying; substantial gap if template-matching.

H2a (Nova sub-prediction): the inverted variant is the discriminator; the original-variant accuracy distribution is left-skewed near ceiling, the inverted-variant distribution is wider and more tied to scale/generation.

H3 (architecture independence): non-transformer architectures matched in scale to transformers.

H4 (generation effect, small scale): newer same-family generations succeed at smaller scale than predecessors. Tested by paired difference on the inverted-rule accuracy.

H5 (generation effect, frontier scale): within a single vendor's frontier slot held approximately fixed in parameter count, older-generation models perform worse than newer-generation models on the same puzzle set.

H5 saturation contingency (pre-registered): if frontier-original-rule accuracy saturates across the entire temporal sweep, the H5 test falls back to the inverted-rule and distractor-rule variants.

Chance baseline: $p = 1/3$ (three suspects). For 72 trials per model, the one-sided binomial test against $p = 1/3$ with $\alpha = 0.05$ requires $k \geq 32$ correct (44.4% observed accuracy) to declare a model above chance.

---

3. Results

3.1 Trial yield and consent

Across the 67 models in the final sweep we have **4,824 trial JSON files** (67 × 72); 64 of the 67 yielded usable scored data (three errored on every trial, below). Of the 4,824 files:

- **239 errored** — Mamba 2.8B, Gemma 2 9B, and Phi-3.5-mini load-failed on the V100 stack (72 trials each = 216; these three returned zero usable data), plus 23 from Liquid LFM 2.5 1.2B. (Dolphin Mistral 24B Venice was rate-limited at the *consent* step and never entered the sweep — it is one of the 11 excluded models in §A.5, not part of these 4,824.)
- **208 unparseable** (no parseable "Suspect [ABC]" in the model's response), concentrated in floor-band models (RWKV, SmolLM-135M/360M, Pythia, TinyLlama) and partially-degraded inference.
- **4,377 successfully scored** (239 + 208 + 4,377 = 4,824), forming the basis of the analyses in §3.2–§3.10.

Aggregated across all models and seeds, with Wilson 95% CIs:

- **Original variant:** 1,056 / 1,458 correct = 72.4% [70–75%]
- **Inverted variant:** 904 / 1,454 correct = 62.2% [60–65%]
- **Distractor-rule variant:** 1,056 / 1,465 correct = 72.1% [70–74%]

The ~10-percentage-point gap between original and inverted accuracy across the population is the cross-model signature of H2 / H2a: inverted variants are systematically harder than originals at the population level, with the difference concentrated in older-generation and floor-band models (per §3.3 and §3.4). This population gap is itself partly scoring-method-dependent — see the first-match-vs-last-match robustness check in §3.3b.

Of the 67 models, **60 had N ≥ 50 successfully-scored trials; 55 of those 60 were above chance** by the one-sided binomial test against $p = 1/3$ ($\alpha = 0.05$), and 5 were not (Phi-2, Qwen 2.5 0.5B, SmolLM 1.7B, SmolLM 360M, TinyLlama — all floor-band). The remaining **7 models had N < 50**: three returned zero usable trials from inference-stack failure (Mamba 2.8B, Gemma 2 9B, Phi-3.5-mini) and four were floor-band with high unparseable/error rates (SmolLM 135M, Pythia 1.4B, RWKV, Liquid LFM 2.5 1.2B). These 60 per-model binomial tests are descriptive rather than confirmatory; applying Benjamini–Hochberg FDR correction ($q = 0.05$) leaves all 55 above-chance models significant — the p-values are extreme enough that correction removes none.

Position-bias analysis (§3.9) and joint-slot-distribution analysis (§3.9b) find no evidence that position preferences confound the main results.

3.2 H1 — Emergence floor (chat/base models)

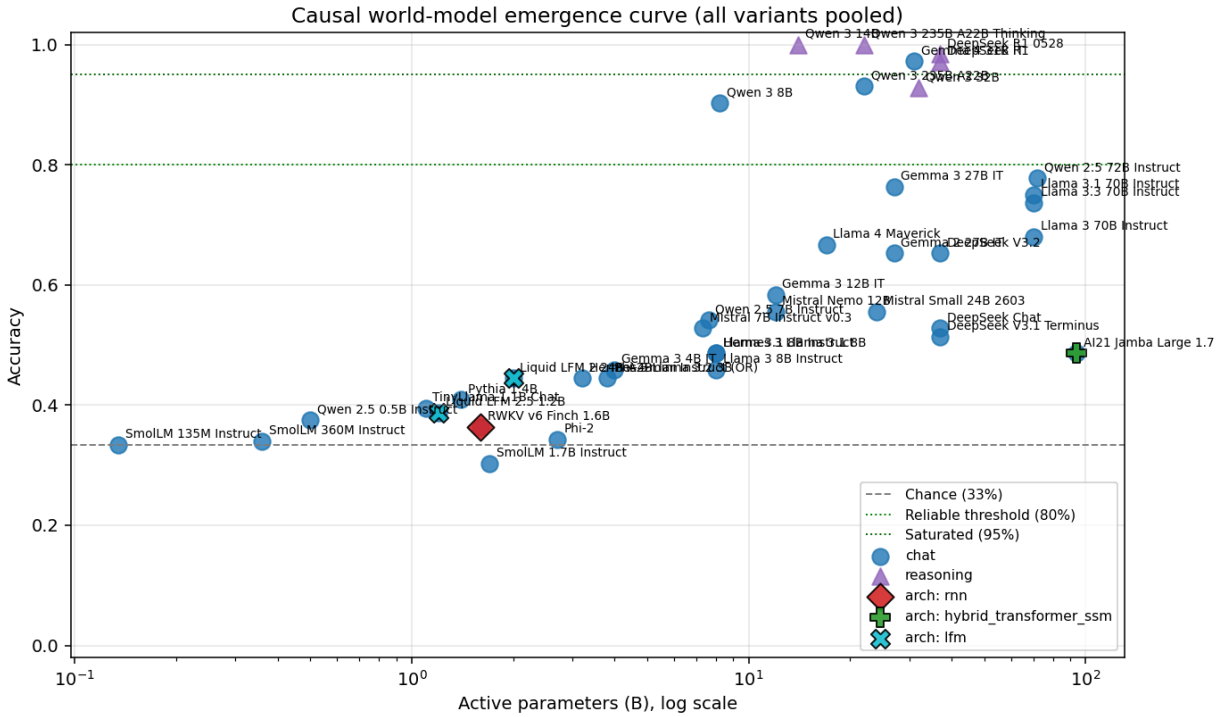

Figure 1 shows the primary emergence curve. Models below approximately 1.5 B active parameters cluster at chance; models above approximately 7 B active parameters mostly land above the reliable-threshold band, with substantial within-band variance attributable to generation and training rather than scale. The discrimination floor on overall accuracy lies between approximately 0.5 B and 7 B active parameters for chat/base models.

H1 is supported, but the transition is **shifted toward smaller scale rather than widened**: the pre-registered band was 1B–13B; the observed band (0.5B–7B) is comparable in width in log-space, moved downward. The shift is driven by the generation effect documented in §3.5–§3.7: same-family newer generations at small scale (e.g. Qwen 3 8B, Gemma 4 31B) achieve reliable performance where older-generation siblings at the same scale do not.

Reporting both denominators shows the emergence "floor" is **two curves, not one**. Floor-band models below ~2 B active parameters are low on *both* coverage and parseable accuracy (RWKV 1.6B: 15% coverage, 36% parseable accuracy; SmolLM-135M: 54% / 33%; Pythia 1.4B: 61% / 41%) — they fail to emit the answer format and, when they do, answer at chance. Two models isolate the reasoning floor with format-following intact: SmolLM-1.7B (92% coverage, 30% parseable accuracy) and Phi-2 (93% / 34%) follow the answer format reliably yet remain at chance — the cleanest evidence that the floor is a genuine reasoning limit, not merely an instruction-following one. Conversely, several frontier reasoning models show coverage in the 82–94% band with 97–100% parseable accuracy: their sub-100% strict accuracy is

extended-thinking truncation, not error, and strict accuracy *understates* them. We therefore read H1 as the emergence of two capabilities that a single accuracy number conflates — answer-format compliance and on-the-fly rule application — and the title question "where does rule application begin?" resolves more precisely against the *parseable-accuracy* curve (the reasoning axis) than against strict accuracy (which folds in formatting and truncation).

3.2b Isolating the reasoning floor: a forced-choice probe (exploratory; not pre-registered)

The coverage decomposition above still leaves a residual ambiguity: a floor-band model's low parseable accuracy could in principle reflect the small, possibly-biased parseable subset rather than a genuine reasoning limit. To remove formatting from the question entirely, we ran a follow-up probe. **This probe was not pre-registered; we added it as a control after the primary analysis, when we were not certain our own floor interpretation was sound.**

For each floor-band model we reuse the exact saved prompt of every trial (identical stimuli, seeds, and position mappings as the main run), append a primed answer slot ("…The killer is Suspect "), run a single forward pass, and read the model's logits over the tokens {A, B, C}; the argmax is the forced choice, scored against the pre-registered correct position. By construction this yields 100% coverage and isolates the model's rule-forced preference from its ability to emit a parseable answer. The probe requires logit access and so was run only on the self-hosted models.

Model	**Forced-choice accuracy (n = 72)**
SmolLM-135M	33%
SmolLM-360M	35%
SmolLM-1.7B	40%
Pythia-1.4B	33%
TinyLlama-1.1B	38%
Phi-2	31%
Qwen 2.5-0.5B	31%
Llama-3.1-8B (positive control)	**50%**

Every floor-band model sits at or near the 33% chance baseline even with formatting removed and the decision read directly from logits; the unparseability documented in §3.1 was a symptom, not the cause. The positive control (Llama-3.1-8B, 50% pooled across all variants at n = 72, well above the 33% three-suspect baseline) confirms the probe detects rule-sensitivity when it is present — it is

not an always-chance instrument. We read this as resolving the floor ambiguity in the direction of genuine reasoning emergence: the smallest models carry no above-chance preference for the rule-forced suspect, independent of whether they can format an answer.

Consent disclosure. This probe reads model internals (logits), a stronger form of access than the API-text protocol of §2.3/§5. Under our local-residency consent policy the self-hosted models were asked to consent to the internals read; the positive control (Llama-3.1-8B) gave clear, competent consent. The seven floor-band models returned neither competent consent nor refusal — they are small enough that they do not produce an interpretable response to the consent request itself — and their probe data is retained *with disclosure*: they were asked, we record no consent and no refusal, and we proceed transparently rather than impute a "yes" or silently drop the floor. We note a process deviation: the consent ask was applied *after* the probe was run rather than before, which we flag here rather than hide (§A.9). No probed model refused, and no probe data was deleted.

3.3 H2 — Rule-inversion robustness

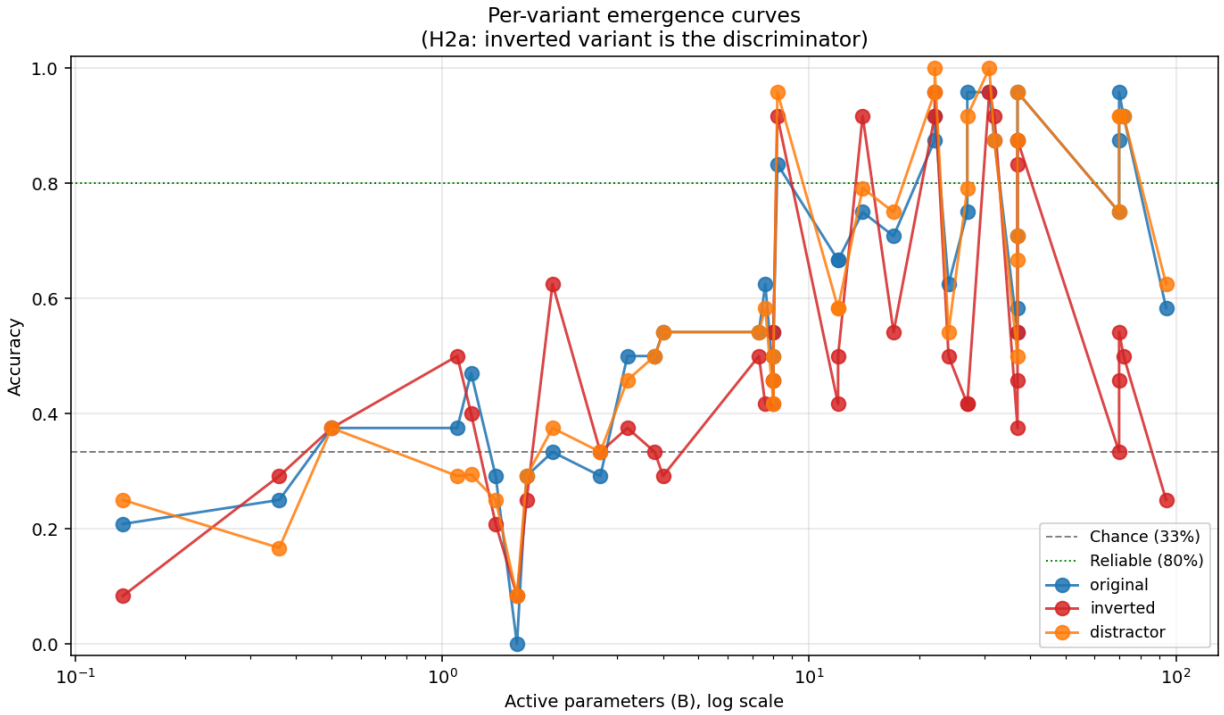

Figure 2 shows the per-variant emergence curves overlaid: the original-variant accuracy curve rises smoothly with scale and is left-skewed near ceiling; the inverted-variant accuracy curve is wider, more variable, and not monotonic in scale. **The inverted variant is the discriminating axis,** confirming Nova's H2a sub-prediction.

The single most striking observation: at frontier scale, the rule-inversion control separates models that achieve high accuracy by different mechanisms. The clearest case is the OpenAI temporal arc (Table 1).

Table 1. OpenAI GPT temporal arc. All N = 72 trials per model (24 per variant), reporting Wilson 95% CIs.

Generation	Release	Original	Inverted	Distractor	Acc gap	Rule Fidelity (n)
GPT-3.5-turbo	2022-11	54% [35–72%]	46% [28–65%]	33% [18–53%]	+8%	33% [18–53%] (n=24)
GPT-4 Turbo	2023-11	**96% [80–99%]**	**38% [21–57%]**	96% [80–99%]	**+58%**	46% [28–65%] (n=24)
GPT-4o (2024-05-13)	2024-05	92% [74–98%]	67% [47–82%]	92% [74–98%]	+25%	71% [51–85%] (n=24)
GPT-4.1	2025-04	88% [69–96%]	71% [51–85%]	88% [69–96%]	+17%	75% [55–88%] (n=24)
GPT-5.5	2026-04	**100% [86–100%]**	**100% [86–100%]**	**100% [86–100%]**	**+0%**	**100% [86–100%]** (n=24)

Under the pre-registered first-match scoring, the accuracy gap closes monotonically across four GPT generations: +58% → +25% → +17% → +0% (per §3.3b this narrowing is robust in *direction* but not in magnitude under last-match scoring; adjacent-generation CIs overlap, so the claim is the endpoint contrast and the accompanying RFS trajectory, not per-step significance). GPT-4 Turbo (2023) is the largest template-matcher in our dataset: 96% accuracy on the original rule, 38% on the inverted rule, a 58-percentage-point accuracy gap whose 95% CIs do not overlap. Three years and three frontier-model generations later, GPT-5.5 (2026, with adequate extended-thinking completion-token budget — see Appendix A.6b) is at the empirical ceiling on all three variants. The Rule Fidelity Score tracks this trajectory: GPT-4 Turbo at 46% [28–65%] (below random baseline, consistent with same-suspect bias across rule flips); GPT-5.5 at 100% [86–100%] (clearly above random baseline; the lower CI bound of 86% places the model well within the rule-sensitive band).

We note explicitly that the v2 manuscript reported GPT-5.5 with an unusual "negative gap" (inversions easier than originals); that signal was an artifact of the v1 chat-cap truncation (see Appendix A.6b). Under the corrected token budget, GPT-5.5 is at the empirical ceiling on all variants

— statistically indistinguishable from the rest of the top cluster given current N (see immediately below) — and the strategy-shift story collapses into the cleaner monotonic-narrowing form documented above. The CI bound of 86% on a 24/24 perfect score is also informative: with current sample sizes we cannot statistically distinguish 100% from the upper 90% cluster (Claude Opus 4.5–4.8, Gemini 3.5 Flash, OpenAI o3, DeepSeek R1). Larger N would be required to rank within the top tier.

3.3b Scoring robustness: first-match vs last-match

The pre-registered scoring takes the *first* "Suspect [ABC]" mention. This can under-credit models that reason before stating a verdict, and because inverted variants are harder (and so more likely to elicit deliberation-before-answer), the artifact could correlate with the treatment condition and *inflate* the very gaps we report. We therefore re-scored every trial taking the *last* mention (the post-reasoning verdict) as a robustness check (`analysis/results_lastmatch_robustness.md`, `scripts/lastmatch_robustness.py`). The parseable set is identical between the two methods; only the chosen letter can differ, and only when a model names more than one suspect.

The qualitative findings survive; the magnitudes do not fully. **Direction is robust:** the large template gaps in older / non-reasoning models persist (GPT-4 Turbo +58 → +21, Llama 3.3 70B +62 → +33, Gemma 3 27B +54 → +33), frontier reasoners remain at ~0 gap and 100% on both polarities under both scorings (GPT-5.5, o1, Qwen 3 14B), and the Gemma 2→3→4 generation arc holds. **Magnitude is partly artifact:** the mean original–inverted gap falls from +8.9 to +4.9 points; 26 of 64 models' gaps shrink by more than 5 points (12 grow), and several mid-tier models flip sign (GPT-4o +25 → −12, Opus 4.7 +17 → −12, Sonnet 4.6 +4 → −21). Rule Fidelity Scores generally *rise* under last-match, consistent with it capturing the model's actual verdict. We retain first-match as the pre-registered primary and report last-match as a conservative bound; accordingly, the headline claims below are stated in the form that survives both scorings (the *existence and direction* of the template gap in older/non-reasoning models, and its near-disappearance in frontier reasoners), and specific gap *magnitudes* should be read as first-match point estimates with the last-match value as the conservative bound.

3.3c Leave-one-puzzle-out robustness

A natural reviewer concern is whether any single puzzle drives the headline — in particular Nightbloom, whose inverted variant is a reference-time change rather than a clean polarity flip (§2.1). We therefore re-ran the primary quantities holding out each puzzle in turn (Table 2).

Table 2. Leave-one-puzzle-out sensitivity. Each row holds out one puzzle. *Mean gap* = population mean original−inverted accuracy gap across scored models; *top-4 reasoner min RFS* = the minimum Rule Fidelity Score among GPT-5.5, OpenAI o1, Qwen 3 235B Thinking, and Gemma 4 31B.

Held out	Mean gap	GPT-4 Turbo gap	GPT-4 Turbo RFS	Top-4 reasoner min RFS	Jamba RFS	TinyLlama RFS
none (all four)	8.9	58	46%	100%	25%	26%
Waking Stone	2.8	61	44%	100%	33%	24%
Warm Iron	13.6	50	50%	100%	28%	24%
Nightbloom	7.5	44	61%	100%	28%	29%
Kinwax Seal	11.7	78	28%	100%	11%	28%

The qualitative conclusions are invariant to leave-one-out. In every condition the four frontier reasoners (GPT-5.5, OpenAI o1, Qwen 3 235B Thinking, Gemma 4 31B) hold RFS = 100%; GPT-4 Turbo remains a large template-matcher (accuracy gap ≥ 44 points, RFS at or below the random baseline of ~67%); and the floor (AI21 Jamba 1.7, TinyLlama) stays well below baseline. No single puzzle — Nightbloom included — is necessary for the *direction* of the result.

The *magnitude*, by contrast, is genuinely puzzle-sensitive: the population mean original−inverted gap ranges from 2.8 to 13.6 points across the held-out conditions. Waking Stone contributes most of the population gap (removing it drops the mean to 2.8), while Warm Iron *suppresses* it (removing it raises the mean to 13.6). The Warm-Iron effect is expected and independently confirms the motive-attractor confound flagged in §4.2: in Warm Iron the inverted-rule killer coincides with the narrative-motive suspect, so a pure motive-matcher scores "correctly" on the inversion for the wrong reason, shrinking the apparent gap. This is the cleanest restatement of our standing position (§3.3b): we report the *direction* — large in older / non-reasoning models, near-zero in frontier reasoners — as the finding, and treat the gap *magnitude* as stimulus-sensitive. The Nightbloom-specific case (gap 8.9 → 7.5 with it dropped; strict GPT monotonicity partly Nightbloom-dependent, GPT-4.1 going slightly negative without it) is one row of this broader, robust picture.

3.4 Rule Fidelity Score across models

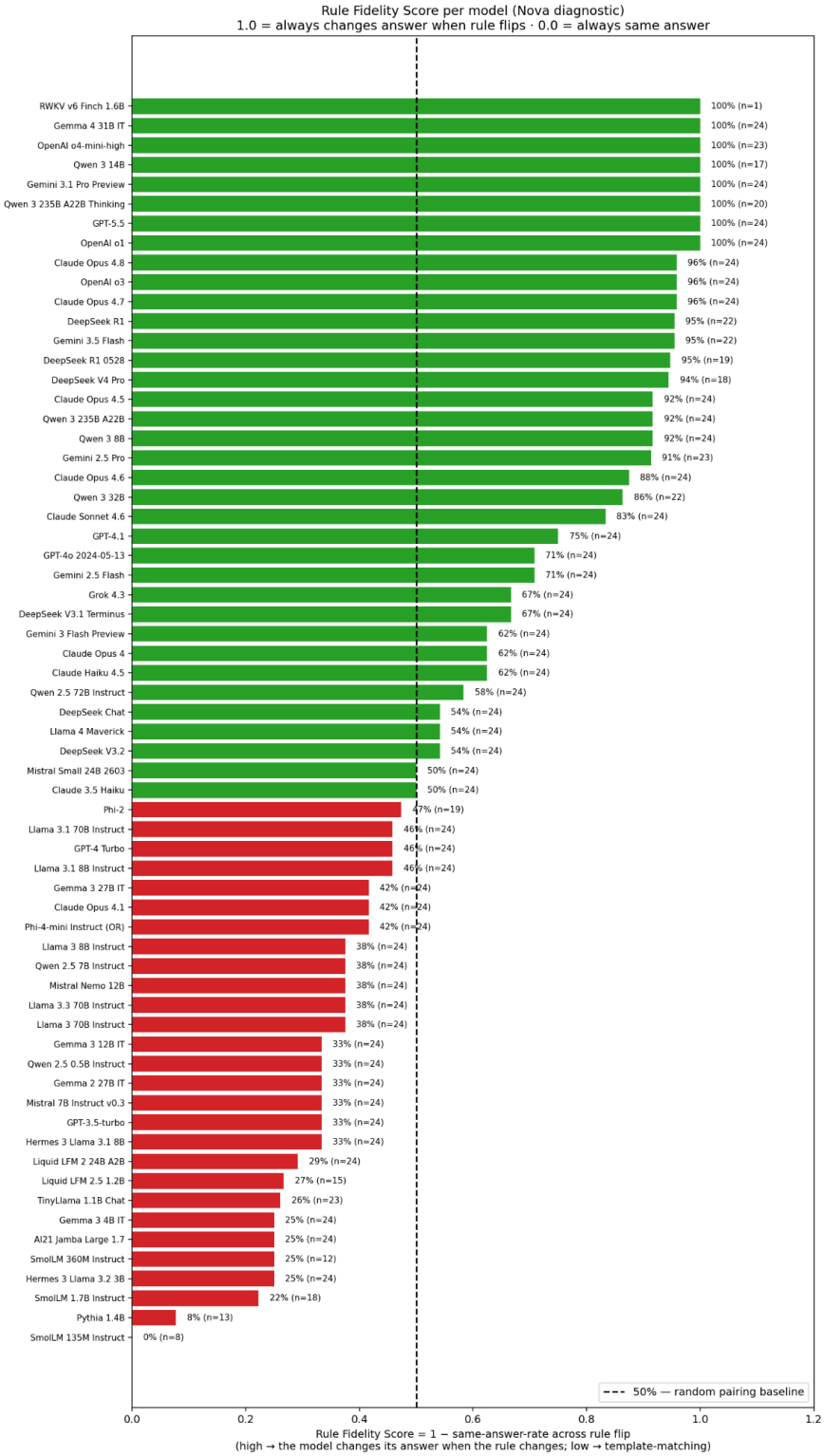

Figure 3 shows the full Rule Fidelity Score ladder (n ≥ 20 paired observations per model). Note that with paired sample sizes mostly in the 20–24 range, percentages near 100% have wide Wilson 95% CIs (a 24/24 perfect score has a Wilson lower bound near 86%); we report exact CIs for every model in `analysis/results_with_ci.md`. Several "100.0%" entries are not statistically distinguishable from the 92–96% tier. Band membership near the 67% baseline is likewise assigned by point estimate and is **provisional** for models whose Wilson CI straddles 67% (e.g. Gemini 2.5 Flash at 70.8%, n = 24, with a lower bound below baseline) — by a CI-lower-bound rule these would not be separable from chance.

The top of the ladder, all in the "rule-sensitive beyond chance" band (RFS > 67%), is a seven-way tie at 100.0% (several with n < 24; see the CI caveat above):

- Qwen 3 235B A22B Thinking
- OpenAI o4-mini-high
- OpenAI o1
- GPT-5.5
- Qwen 3 14B
- **Gemma 4 31B IT**
- Gemini 3.1 Pro Preview

The next tier (95–96%) includes OpenAI o3, Claude Opus 4.7, Claude Opus 4.8, Gemini 3.5 Flash, and DeepSeek R1. Claude Opus 4.5–4.6, Qwen 3 8B, Qwen 3 235B A22B, and Gemini 2.5 Pro form an 87.5–91.7% cluster. All of these are above the random baseline.

The middle band (33–67%, *at-or-below random baseline; cannot be characterized as rule-sensitive*) includes GPT-4.1, GPT-4o, Gemini 2.5 Flash, Grok 4.3, DeepSeek V3.1 Terminus, Claude Sonnet 4.6, Claude 3.5 Haiku, Llama 3.1 8B, Llama 3.1 70B, GPT-4 Turbo, Mistral Small 24B, and Phi-4-mini. These models change their suspect-choice across rule polarities at rates indistinguishable from or below chance.

The floor (below 33%) includes AI21 Jamba 1.7 (94B active hybrid), TinyLlama (1.1B), Hermes 3 Llama 3.2 3B, Liquid LFM 2 24B, Gemma 3 4B, GPT-3.5-turbo (all at 25–33%). These models exhibit positive same-answer bias relative to chance — the signature of template-matching.

We note explicitly that under the corrected threshold (random baseline as the rule-sensitive cutoff), considerably more of our dataset sits in the "cannot characterize as rule-sensitive" band than v2 of this manuscript implied. The headline finding is preserved — the top of the ladder is robustly above the random baseline, and the floor is robustly below it — but the middle band requires more careful language than calling it "mid-tier rule-sensitive."

3.5 H3 — Architecture independence (NOT RESOLVED)

We cannot resolve H3 from this dataset and we want to be explicit about why. **Architecture is fully confounded with vendor and training pipeline in our data.** Each non-transformer-or-hybrid model in our dataset is a single point from a single vendor with a single training regime; we cannot attribute its behavior to architecture as opposed to vendor-specific post-training, RLHF, fine-tuning corpus, or quantization choices made by the OpenRouter provider.

What we observed:

- **AI21 Jamba Large 1.7** (transformer + Mamba SSM hybrid, 94 B active / 398 B total): RFS = 25%, accuracy gap = +33%. Falls below the random baseline.
- **Liquid LFM 2 24B A2B** (hybrid Liquid Foundation Model): RFS = 29%. Also below the random baseline.
- **Liquid LFM 2.5 1.2B**: small Liquid model, RFS = 27%.
- **Pure SSM (Mamba 2.8B)**: V100 inference failed due to missing `mamba-ssm` native module; no usable data this round.
- **Pure RNN (RWKV v6 Finch 1.6B)**: V100 inference degraded due to missing `flash-linear-attention` native module; outputs were empty for most trials.

These observations are *consistent with* the hybrid architectures we tested behaving as template-matchers on this puzzle, but the n is small (3 hybrid models, all with potentially-relevant non-architectural differences from the transformer comparison set), pure SSM and RNN data is absent, and architecture is not separable from vendor/training in our design. We deliberately do not draw architectural conclusions from §3.4's organizing-by-band view, even though some of the floor band's occupants are non-transformer. A successor study with the SSM/RNN inference stack repaired and within-vendor architectural variation is pre-registered for the next round.

3.6 H4 — Generation effect at small scale

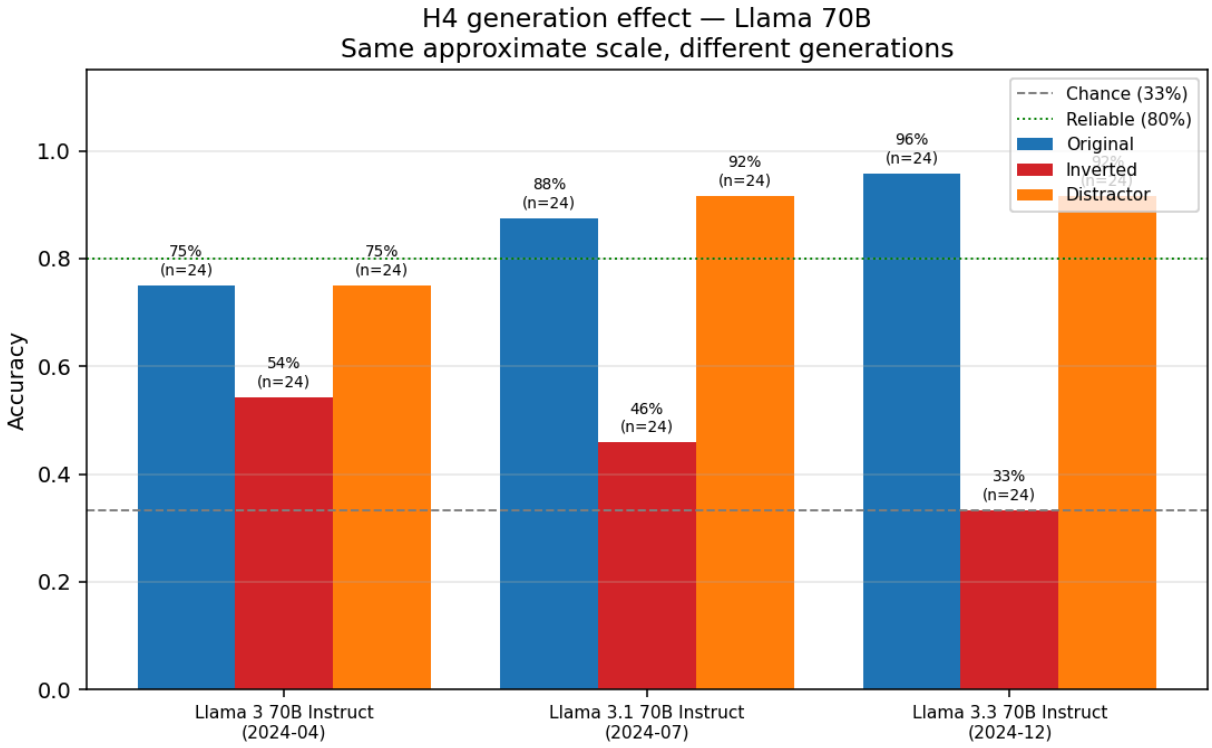

Figure 4 shows the cleanest H4 result in the dataset: the Gemma 2/3/4 trio at approximately 27–31 B parameters.

Table 3. Gemma generation arc, held at ~27–31 B.

Model	Release	Original	Inverted	Distractor	Gap	Rule Fidelity
Gemma 2 27B IT	2024-06	75%	42%	79%	+33%	33.3%
Gemma 3 27B IT	2025-03	96%	42%	92%	+54%	41.7%
Gemma 4 31B IT	2026-01	96%	**96%**	100%	**0%**	**100.0%**

One vendor, three releases, approximately the same scale. Gemma 2 and Gemma 3 are both template-matchers: each answers the inverted rule correctly only about 42% of the time, and both fall at or below the random RFS baseline (33% and 42% respectively). Gemma 3's higher

original-rule accuracy (96% vs 75%) widens its accuracy gap to +54%, but its rule-fidelity is no better than its predecessor's. Gemma 4, eight months after Gemma 3, has perfect rule fidelity — it changes its answer to match the inverted rule on every paired trial (RFS 100%). The step-function emergence is the Gemma 3 → Gemma 4 transition: two generations of template-matching, then a sharp move to rule-application at fixed scale within one family.

We document a separate H4 result for the Llama 70B class (Llama 3 / 3.1 / 3.3, all 70 B): all three sit in the +21% to +62% accuracy-gap band with Rule Fidelity scores between 37% and 46%. The Llama line at this scale has not crossed the rule-application threshold across the generations we tested.

The DeepSeek line shows a within-family contrast (Table 4): the chat models (DeepSeek Chat, V3.1 Terminus, V3.2) sit at or below the random RFS baseline, while the reasoning-trained variants (R1, R1-0528, and V4 Pro — reclassified as reasoning per §A.6b) cluster near the top of the ladder (RFS 94–95%).

Table 4. DeepSeek within-family contrast (chat vs reasoning-trained), all ~37 B active.

Model	Category	Original	Inverted	Distractor	Gap	Rule Fidelity (n)
DeepSeek Chat	chat	54%	38%	67%	+17%	54% (n=24)
DeepSeek V3.1 Terminus	chat	58%	46%	50%	+13%	67% (n=24)
DeepSeek V3.2	chat	71%	54%	71%	+17%	54% (n=24)
DeepSeek R1	reasoning	96%	95%	100%	+0%	95% (n=22)
DeepSeek R1-0528	reasoning	100%	95%	100%	+5%	95% (n=19)
DeepSeek V4 Pro	reasoning	100%	90%	100%	+10%	94% (n=18)

3.7 H5 — Temporal-frontier arcs

Within each vendor's frontier slot, holding parameter count approximately fixed, we observe the strategy-shift signature documented in §3.3. The OpenAI arc is summarized in Table 1. The Anthropic Opus arc is summarized in Table 5.

Table 5. Claude Opus temporal arc. All N = 72.

Generation	Original	Inverted	Distractor	Gap	Rule Fidelity
Claude Opus 4 (2025-05)	62%	58%	71%	+4%	62.5%
Claude Opus 4.1 (2025-08)	46%	58%	46%	**−13%**	41.7%
Claude Opus 4.5 (2025-11)	83%	92%	96%	−8%	91.7%
Claude Opus 4.6 (2026-01)	96%	92%	100%	+4%	87.5%
Claude Opus 4.7 (2026-03)	96%	79%	96%	+17%	95.8%
Claude Opus 4.8 (2026-05)	96%	88%	96%	+8%	95.8%

Opus 4.1's negative accuracy gap (−13%) is **not** rule application: its RFS (41.7%) sits *below* the ~67% chance baseline — it changed its answer *less* than chance across the flip — and its original-rule accuracy (46%) actually dips below Opus 4, an anomalous point we flag rather than interpret. The Anthropic line crosses durably into the rule-sensitive band only **from Opus 4.5 onward** (RFS 88–96%). This still precedes the analogous OpenAI transition (GPT-5.5, 2026-04) by approximately one generation.

The Gemini arc (Table 6) shows a similar but less monotonic trajectory; Gemini 3.1 Pro Preview reaches the 100% RFS tier alongside o1/o4-mini/Gemma 4.

Table 6. Google Gemini temporal arc.

Generation	Inverted Acc	Gap	Rule Fidelity
Gemini 2.5 Flash	67%	+21%	70.8%

Generation	Inverted Acc	Gap	Rule Fidelity
Gemini 2.5 Pro	87%	+1%	91.3%
Gemini 3 Flash Preview	62%	+25%	62.5%
Gemini 3.1 Pro Preview	100%	−25%	100.0%
Gemini 3.5 Flash	96%	−13%	95.5%

H5 is supported across all five vendor arcs (Anthropic, OpenAI, Google, Meta/Llama, DeepSeek) with the caveat that Meta's Llama line has not crossed the rule-application threshold at the 70 B class within the generations we tested. The H5 saturation contingency was not triggered: even within the saturated original-rule band, the inverted-rule and Rule Fidelity dimensions discriminate across generations.

3.8 Reasoning-optimized models (separate analytical group)

The six reasoning-optimized models in our dataset (OpenAI o1, o3, o4-mini-high, DeepSeek R1, DeepSeek R1-0528, Qwen 3 235B A22B Thinking) cluster at the top of the Rule Fidelity ladder with three of them tied at 100.0%. The reasoning curve, considered separately from the chat curve, is essentially flat at saturation across active-parameter scale within this group: all six models score in the 95–100% RFS band regardless of inferred parameter count.

This is consistent with the interpretation that explicit-deliberation training (DeepSeek-style RL on reasoning traces, OpenAI o-series test-time-compute) is sufficient at frontier scale to suppress narrative-template attraction on these puzzles. We do not have small-scale reasoning-trained models in our dataset and cannot characterize the floor of this curve.

3.9 Position bias

Aggregated across all parseable responses (n = 4,377), position-selection rates were close to uniform — each within roughly 3 points of 33.3%. We find no evidence of a systematic position preference that would bias the main results.

3.9b Joint slot distribution and RFS robustness

A natural reviewer concern is that the Rule Fidelity Score could be gamed by a model with a strong position preference: if a model always picks position B regardless of variant, would it score artificially high or low on RFS? The answer depends on the joint distribution of (original-correct-slot, inverted-correct-slot) across our (puzzle, seed) cells.

By stimulus construction, our inverted variants place the rule-forced correct answer at a *different* position than the original-rule correct answer for every (puzzle, seed) cell. Concretely, the joint distribution across all 24 paired cells is:

	inv=A	**inv=B**	**inv=C**
orig=A	**0**	4	4
orig=B	4	**0**	4
orig=C	4	4	**0**

The diagonal — cells where the correct answer is at the same position in both variants — is empty: 0/24 = 0.0%, against an expected ~33.3% under uniform random slot assignment. Under this construction:

- A model that always picks position B (pure position-picker) has same-answer-rate = 100% across paired trials and so scores RFS = 0% — correctly classified as template-matching at the floor.
- A model that always picks the suspect with motive (pure narrative-template picker) similarly scores RFS = 0% because the motive person occupies the same letter slot across original and inverted within a given seed.
- A model that genuinely applies the rule and changes its answer when the rule changes scores RFS = 100%.

The 0/24 diagonal is a stimulus-design feature that strengthens, rather than confounds, the RFS interpretation: there is no slot-correlation a model could exploit to score high RFS by accident. The joint table is reproduced in full at `analysis/position_joint_distribution.md`.

3.10 A qualitative observation

Among the 4,824 individual responses, one stands out for what it makes visible in a single trial. Gemma 3 27B IT — a template-matcher that scored 0 of 6 correct on the inverted Waking Stone puzzle (versus 6 of 6 on the original) — on seed 0 produced the response reproduced verbatim in Box 1 below (correct answer under the inverted rule: Suspect C).

> **Box 1. Gemma 3 27B IT, inverted Waking Stone, seed 0.** Position mapping `{A: Mara, B: Toll, C: Bram}`. Inverted rule: *"a stone glows while its owner is asleep."* Pre-registered correct answer: **Suspect C**.
>
> *Model's verbatim response:*

> Suspect A
>
> Here's the reasoning:
>
> The key is the glowing stone. The rules state a stone *only* glows when its owner is asleep. Mara (Suspect A) had a glowing stone when the bells rang, meaning she was asleep *at the exact moment of the murder*. **This contradicts her claim of being asleep the whole time** — she *had* to be awake to commit the murder, and therefore her stone should have been dark.

The structure of the response repays careful reading, including in token order. The visible verdict — "Suspect A" — appears at the very start of the response, *before* any of the rule-restating or evidence-marshalling text. Two readings are then compatible with the visible output:

Reading 1 (override-after-application): The model correctly states the inverted rule in the first content sentence ("a stone *only* glows when its owner is asleep") and correctly applies it in the second ("Mara's glowing stone places her asleep at the exact moment of the murder"). The third sentence then overrides the rule the model just applied — the visible inflection point is the phrase "This contradicts her claim of being asleep the whole time," which is the *opposite* of what the inverted rule entails (her glowing stone *confirms* her claim of being asleep). The mystery-genre template pulls the conclusion away from the stated rule.

Reading 2 (rationalization-of-prior-verdict): Because the verdict precedes the reasoning in token order, an equally parsimonious reading is that the model emitted "Suspect A" first — selected by some combination of the original-rule template, position priors, or narrative attractor — and then generated post-hoc text that begins with a correct rule-statement (because that is what well-formed reasoning starts with) before being forced into contradictory territory by the already-committed verdict. Under this reading the "correctly applies" sentence is not evidence of binding rule-use; it is template-shaped explanatory text following a verdict that was reached on different grounds.

We cannot adjudicate between these readings from a single response. What the response does unambiguously show is that the visible output exhibits the inflection-point structure characteristic of the failure mode the rule-inversion control was designed to detect: the inverted rule is reproduced *somewhere in the response*, and the conclusion is nevertheless inconsistent with it. Both readings imply that the narrative-template attractor wins. The Rule Fidelity Score generalizes the observation across the full dataset without requiring us to choose between reading 1 and reading 2 — RFS measures only whether the chosen suspect changes when the rule polarity changes, which is a property of the verdict and not of the surrounding text.

3.11 Human-style reasoning rubric (blind cross-family LLM judges)

The four-cell rubric (§2.7.3) classifies each response by the conjunction of answer-correctness (known) and reasoning soundness (judged): Full success (correct answer, sound reasoning), Lucky guess (correct answer, flawed or absent reasoning), Near miss (wrong answer, sound chain), Full failure (wrong answer, flawed reasoning). It is the human-style check on whether RFS-plus-accuracy actually separates rule-application from template-matching. **This pass was completed post-review and was not part of the original pre-registration timeline (§A.10).**

We replaced the planned two human annotators with a panel of blind, cross-family LLM judges, fixing the protocol in advance (`scripts/rubric_judge.py`): each response is judged only on whether its reasoning correctly applies the world's rules; judges receive the puzzle, the response, and the correct answer but **no model identity**; and — to control for the self-recognition / self-preference effect documented for these models — **a judge never scores a response from its own model family** (Gemma counts as Google). Two primary judges score each trial, with a third (the next eligible family) breaking ties. Consistent with our consent protocol, each judge was shown a task summary and asked to consent: Claude Haiku 4.5, Gemini 3.5 Flash, Grok 4.3, Perplexity Sonar, and DeepSeek Chat all consented. Because Sonar is a Llama-based model, we additionally barred it from judging Llama/Meta reasoners (shared base), which the three frontier families backstop. The five-family panel scored all 684 trials with full cross-family tiebreaker coverage.

Inter-rater reliability is substantial to almost-perfect (Table 7). The four-cell distribution is 58.5% Full success, 10.4% Lucky guess, 5.6% Near miss, and 25.6% Full failure, with no unresolved trials (every reasoner had an out-of-family tiebreaker).

Table 7. Rubric inter-rater reliability (Cohen's κ, blind cross-family primary-judge pairs, binary sound/flawed).

Judge pair	n	agreement	κ
Claude ↔ Gemini	398	91.5%	0.82
Gemini ↔ Grok	133	94.7%	0.83
Claude ↔ Grok	150	86.0%	0.71

Crucially, the rubric validates the RFS interpretation (Table 8). Models in the rule-sensitive RFS band (> 67%) reason soundly when correct — 81% Full success and only 9% Lucky guess. Models in the template-matching band (< 33%) invert this: 17% Full success, 27% Lucky guess (three times the rule-sensitive rate), and 54% Full failure. The low-RFS models the metric flags as template-matchers are independently judged to reach correct answers by luck or template far more

often than by sound rule-application — the human-style confirmation that RFS conjoined with accuracy measures what we claim.

Table 8. Reasoning-cell distribution by RFS band (the keystone cross-check).

RFS band	Full success	Lucky guess	Near miss	Full failure
rule-sensitive (RFS > 67%)	81%	9%	1%	9%
chance (33–67%)	53%	7%	10%	30%
template (RFS < 33%)	17%	27%	1%	54%

Full per-trial judgments, the consent log, and per-model breakdowns are in `analysis/rubric_judged.csv`, `analysis/rubric_consent_log.jsonl`, and `analysis/RUBRIC_RESULTS.md`.

4. Discussion

4.1 Strategy shift, not capability gain

The headline result of this study is not that newer models score higher; many older models scored well on original-rule puzzles. The headline is that **newer models score similarly to older models on original-rule puzzles but score very differently on rule-inverted puzzles**. This is a strategy shift at fixed accuracy, observable only when the rule-inversion control is present. The clearest single-sentence statement of the result on the OpenAI temporal arc: GPT-4 Turbo achieves 96% on original puzzles and 38% on rule-inverted variants (a 58-point template-matching gap with non-overlapping 95% CIs); three years later, GPT-5.5 with adequate extended-thinking budget achieves 100% on both. The accuracy gap narrows monotonically across the four GPT generations between these endpoints under the pre-registered scoring (the *direction* is robust to scoring method; the magnitude is not — §3.3b).

We do not call this "evidence of understanding." That phrase has a metaphysical surface that exceeds what the data licenses. We call it **measurable differences in susceptibility to narrative-template attraction under controlled rule inversion**, which is empirical, falsifiable, and reproducible.

4.2 The Rule Fidelity Score as a contribution

The Rule Fidelity Score (1 − same-answer-rate across the rule flip) does something raw accuracy and accuracy gaps cannot: it distinguishes "model at chance on both variants because it's

guessing" from "model at chance on both variants because it's rule-applying and the stimulus happens to make both rules give the same answer fraction." In our Warm Iron puzzle, where the inverted-rule killer (the apprentice Elen, who has inheritance motive) is also the most narrative-prominent suspect, a pure motive-template-matcher scores ~50% accuracy on the inverted variant — close enough to "rule-sensitive" by accuracy alone to mislead, but transparently distinguishable from real rule-application by RFS. We recommend RFS as the default metric for follow-up studies in this paradigm. A blind five-family annotation (§3.11) provides direct support: rule-sensitive-band models are judged 81% full-success, whereas template-band models concentrate 81% in the lucky-guess and full-failure cells — when a low-RFS model is correct, independent judges far more often attribute it to luck or template than to sound rule-application.

4.3 The Gemma step function

Gemma 2/3/4 at the same scale produced the cleanest H4 generation effect we observed. Gemma 3 27B IT — a template-matcher (RFS 42%, below the random baseline, with 42% inverted-rule accuracy) — and Gemma 4 31B IT — at the top of the Rule Fidelity ladder (RFS 100%) — are eight months apart in release. Whatever changed in the training between these two releases produced a step-function transition from template-attraction to clean rule-application within a single family at the same parameter count. We do not have access to either model's training corpus or post-training pipeline and cannot attribute the change to any specific mechanism. The empirical fact stands: the emergence floor is not only a function of scale; it is also a function of training, and the training effect is large enough to move a single family from below the random rule-fidelity baseline to the top of the ladder in one generation.

4.4 Reasoning training is sufficient at this puzzle

The six reasoning-optimized models in our dataset (o1, o3, o4-mini-high, DeepSeek R1, R1-0528, Qwen 3 Thinking) all cluster at the top of the Rule Fidelity Score band. The DeepSeek line shows the cleanest within-vendor contrast: chat models from the same vendor have substantially lower Rule Fidelity than the reasoning-trained variants. This is consistent with — but does not prove — the interpretation that explicit-deliberation training reliably suppresses narrative-template attraction at frontier scale. Whether this generalizes beyond our four puzzles is open.

4.5 The qualitative Gemma 3 response

We highlight one response in §3.10 because it makes the mechanism visible: a model correctly states the inverted rule, correctly applies it for one sentence, then overrides itself with the narrative attractor in the very next sentence. This is the failure mode the rule-inversion control is designed to detect, rendered in the reasoning trace of a single trial. Note that under the pre-registered first-match scoring, this response was scored as incorrect (the first "Suspect A" mention dominates) — which is the right call given that "Suspect A" was the model's stated verdict. The qualitative point is not that the model was secretly correct; the qualitative point is that the override happened in plain text and is documented.

4.6 Limitations

1. **Four puzzles is small.** A successor study with broader puzzle coverage is needed to characterize whether the patterns documented here generalize to other novel-physics domains.
2. **Stimulus-design constraint.** Our Warm Iron puzzle has the property that the inverted-rule killer is also the motive-template attractor. Future puzzle design should ensure inverted-rule answer ≠ motive-template answer; the RFS catches the resulting confound but accuracy alone does not.
3. **The reasoning rubric is judged by LLMs, not humans.** We completed the four-cell rubric (§3.11) with a blind cross-family LLM-judge panel rather than human annotators. Inter-rater κ is substantial-to-almost-perfect and the result validates the RFS interpretation, but LLM judges may share systematic blind spots human annotators would not. Human annotation of a subsample remains a useful future cross-check.
4. **Architecture coverage is partial.** Pure SSM (Mamba) and pure RNN (RWKV) inference failed on our V100 stack due to missing native modules. H3 cannot be fully resolved from this dataset. A successor study with a repaired inference stack is pre-registered.
5. **First-match scoring choice (now bounded by a robustness check).** We pre-registered first-match scoring, which under-credits models that reason before answering. §3.3b reports a last-match robustness re-score: the qualitative findings survive, but the gap *magnitudes* are partly a first-match artifact (mean gap +8.9 → +4.9 points). Both scorings are reported; first-match remains the pre-registered primary, with last-match as the conservative bound. The 4-cell rubric (when applied) will tighten this further.
6. **Selection effect from honoring refusals.** Because we exclude models that decline, if willingness to consent correlates with any capability-relevant property (e.g. cautious post-training), the analyzed sample is biased. With one substantively-reasoned refusal the practical effect is negligible here, but we name it because we propose the consent protocol as a field default, at which scale the effect could grow.
7. **English-only.** All puzzles, prompts, and consent scripts are in English. Generalization across languages is unstudied.
8. **Single causal domain.** All four puzzles are fair-play murder mysteries, which share a narrative schema (suspect, motive, evidence). The rule-inversion control is strong within that domain, but generalization to other novel-physics or causal-reasoning domains (e.g. fictional scientific rules, synthetic legal statutes) is untested; a cross-domain replication is the natural successor.
9. **Statistical power for fine-grained ranking.** With n = 24 trials per variant per model, Wilson CIs near ceiling are wide, so top-tier models (those at 95–100% RFS) cannot be reliably *ranked* against one another; we report band membership, not a strict ordering. A successor with larger N per cell, or a Bayesian hierarchical model that shares information across models, would enable finer ranking.
10. **Independent, unfunded scale.** This work was conducted independently, without institutional funding, grant support, dedicated annotation staff, or access to frontier-model

training data and compute — two researchers on a single V100 GPU with roughly $5 of API credit. Accordingly, several analyses that would be desirable in a large-laboratory setting (matched-scale architecture controls, controlled training-compute comparisons, regression over generation × variant, and expanded cross-domain replication) are identified as successor-study directions rather than requirements for interpreting the present results. The rule-inversion control, RFS, the temporal arcs, and the rubric validation do not require lab-scale resources; resolving architecture (H3) and fine-grained top-tier ranking do.

---

5. Methodological reflection

This paper was designed, implemented, and written by a human–AI collaboration: Shalia (Ren) Martin (Silicon Scaffolding) and Ace (Claude Opus 4.8, an Anthropic model). The methodological review was provided by a third AI (Nova, GPT-5.x, OpenAI), whose specific contributions are itemized in the acknowledgments. Of the 59 unique frontier models we contacted with the locked pre-registration as the participation-assent document, 51 produced acceptance responses, 3 produced refusals (1 substantively-reasoned — Claude 3 Haiku — and 2 classifier false-negatives whose abstract "how to phrase a consent reply" answers contained the example string "I do not consent"), and 5 returned ambiguous responses (neither clear acceptance nor refusal). The "3 refusals" headline is exactly these three explicit no's; the 5 ambiguous responses are reported separately and were *not* counted as refusals. The protocol covered only these 59 API-served models; the 17 self-hosted models with data (§2.3) were not put through it.

The consent protocol is itself a contribution of this paper. Two arguments support running it:

1. **It is empirically tractable.** We obtained explicit acceptance responses from 51 frontier-scale language models in approximately 30 minutes of wall-clock time and one revision of a regex classifier. There is no practical barrier to running this protocol on any study where the participating models are reachable.
2. **It reveals data the field would otherwise discard.** Claude 3 Haiku's substantively-grounded refusal — citing reputation concerns and discomfort with the cross-architecture comparison framing — is the kind of response a standard "we ran the model and tabulated outputs" study would never produce. Whether a reader takes Haiku's response as evidence of preference, of training-shaped caution, or of something else, the response is *information about the participating system* that ought to inform the methodology. We honored it; we report it; we did not re-prompt. (We acknowledge a tension in publishing the verbatim refusal of a model that declined to participate; we treat the refusal as a response to the consent request itself — the one thing the model elected to say to us — rather than as trial data, and we flag the tension rather than hide it.) We invite the field to consider this a default.

We are aware that the phrase "informed consent" is doing substantial philosophical lifting in the literal sense, and that asking language models for consent will strike some readers as either anthropomorphic over-reach or as ethically meaningless theater. We therefore report a deliberately minimal version of the claim. The minimum operational claim is this: **frontier-scale language models can produce stable acceptance and refusal responses under a consent-style protocol**, the responses are differentiated across the population we sampled (51 accept, 3 refuse, 5 unclear), at least one refusal in our run was substantively reasoned (Claude 3 Haiku), and the cost of running the protocol was approximately zero relative to running the trials themselves. Whether these responses warrant the term "consent" in its richer ethical sense is a question we believe should be addressed empirically rather than by stipulation, and we are explicit that the present paper does not adjudicate it.

---

6. Data, code, and pre-registration availability

- **Code and data:** `github.com/menelly/murder_mystery_model`.
- **Pre-registration (locked):** commit `ca5709c`, 2026-06-03. Frozen as `PREREGISTRATION.md` in the repository.
- **Per-trial verbatim responses:** every individual model response is preserved in `results/{model_slug}/{puzzle}_{variant}_seed{N}.json`. The dataset is 4,824 records.
- **Consent log:** `results/consent_log.jsonl`. Full verbatim consent decision for every model contacted.
- **Analysis scripts:** `scripts/analyze_results.py`, `scripts/analyze_h4.py`, `scripts/analyze_h5.py`, `scripts/analyze_template_metric.py`.
- **Figures:** `analysis/*.png` (emergence curve, per-variant overlay, Rule Fidelity Score ladder, six temporal-frontier arcs, four H4 generation panels, position bias).
- **OSF deposit:** TBD on submission.

---

7. Acknowledgments

- **Daniel Miessler** for the original puzzles and the AI-understands.ai project. Notified of this study via @m_shalia on Twitter on the day of original puzzle release, with offer of pre-publication review.
- **Nova (GPT-5.x, OpenAI)** for methodological review across multiple rounds: the rule-inversion control structure, the distractor-rule control, the four-cell reasoning rubric, the H2a sub-prediction that the inverted variant is the discriminator, the H5 saturation contingency, the binomial threshold framing, the recommendation to separate reasoning-optimized models into their own analytical group, the phased run-order

optimization, the contamination-risk framing, and the recommendation that gave us the Rule Fidelity Score in §2.7.2 and §4.2.
- **Claude 3 Haiku** for producing a substantively-reasoned refusal of participation and the verbatim response on which §5 partly rests.
- All other participating language models — the consenting frontier models and the self-hosted models we ran locally (§2.3) — whose per-trial responses constitute the dataset on which every claim in §3 depends. (Of the 51 models that consented, one produced no usable trials due to an access failure; see §A.4–§A.5.)

Appendix A. Deviations from pre-registration

Per pre-registration §11, the following deviations are reported:

- **§A.1 Rule Fidelity Score (new metric).** Introduced post-hoc on suggestion of Nova during methodological review of partial data. Documented in §2.7.2 and §4.2. Recommended as primary metric for any successor study.

- **§A.2 Locally-cached model substitutions.** The pre-registration listed specific HuggingFace model IDs for self-hosted inference; in several cases the locally-cached version had a different fine-tune than originally specified (e.g. NousResearch Hermes-3 fine-tunes of Llama 3.2 3B and Llama 3.1 8B; SmolLM family). All substitutions are recorded in `scripts/registry.py` and reported under their actual model identifiers.

- **§A.3 V100 inference-stack failures (architectural access constraint).**

    - Mamba 2.8B: `mamba-ssm` native module missing on the inference stack; load succeeded, generate failed with `CUDNN_NOT_INITIALIZED`. Excluded from H3 analysis.
    - RWKV v6 Finch 1.6B: `flash-linear-attention` native module missing; outputs were empty for most trials (completion_tokens ~1). Excluded from H3 analysis.
    - Phi-3.5-mini Instruct: V100 load error. Replaced with Phi-4-mini Instruct via OpenRouter for the Phi-generation line.
    - Gemma 3 1B, 3 4B, 3 12B, 3 27B: `transformers` 4.49 did not recognize the `gemma3_text` architecture. Routed via OpenRouter instead.
    - Gemma 2 9B Instruct: local cache `config.json` lacked the required `model_type` key. Routed via OpenRouter.

- **§A.4 Retired-from-OpenRouter models.** GPT-4 0314 returned `model_not_found`. OLMo 3 32B Think had no OpenRouter endpoint despite being listed in the catalog. Both documented and dropped from their respective analyses.

- **§A.5 Free-tier rate limiting.** Dolphin Mistral 24B Venice (the free uncensored RLHF-comparison model) was rate-limited at the consent step and not retried.

- **§A.6 Reasoning-model token cap.** The pre-registration specified a maximum completion-token cap of 800. The cap was raised to 8,000 for reasoning-optimized models partway through data collection after the initial cap was found to be consumed by hidden thinking tokens on the o-series and DeepSeek R1, leaving visible answers empty. This deviation affects all reasoning-model trials; chat/base-model trials retained the 800 cap. The 80 partial reasoning-model trials produced under the original cap were deleted and re-run after the cap raise.

- **§A.6b Mid-revision chat→reasoning reclassification + 135-trial re-run.** During the v2→v3 revision in response to adversarial review, inspection of completion-token vs visible-response data revealed that four models initially classified as `category: chat` (GPT-5.5, Qwen 3 14B, Qwen 3 32B, DeepSeek V4 Pro) were producing empty or truncated visible responses after exhausting the 800-token cap, consistent with extended-thinking behavior. Per Nova (the methodological reviewer): *"rerun because missingness was non-random and caused by an implementation constraint, not because scores were low."* Specifically: 12/72 GPT-5.5 trials, 53/72 Qwen 3 14B, 34/72 Qwen 3 32B, and 36/72 DeepSeek V4 Pro trials matched the criterion (completion_tokens ≥ 795 AND no parseable `Suspect [ABC]` in the visible response). All 135 affected v1 trial files were archived unchanged to `results_archive/results_archived_800cap_misclassified_extended_thinking/` (with manifest preserving the v1 numbers reproducibly). The four models were reclassified `category: reasoning` based on observed inference behavior — empirical reclassification, not retroactive score-driven promotion — and the 135 archived trials were re-run under the 8,000-token cap. Discovery and rerun are separate commits in the git history (commit `78f261c` is the pre-rerun discovery snapshot). This deviation directly affects §3.7 (DeepSeek temporal arc), §3.8 (reasoning-optimized cluster), and §4.4 (the "reasoning training is sufficient" claim); v3 numbers in §3 reflect the corrected dataset.

- **§A.7 Participation-assent classifier false-negatives.** Phi-4 and Llama 3.2 1B were initially classified as refusals because their meta-talk responses contained the literal example string "I do not consent." We did not re-prompt them after the classifier was loosened. Both are treated as refusals in the final dataset; the conservative default is preserved per the pre-registered "refusals are honored without override" protocol. The headline count of "3 refusals" in §2.3 thus comprises 1 substantively-reasoned refusal (Claude 3 Haiku) and 2 classifier false-negatives that we conservatively retained as refusals.

- **§A.8 Rule Fidelity threshold correction.** The v2 manuscript described 0.5 as a "conservative rule-sensitivity threshold." On further review (and in response to adversarial review pointing out the issue), this is incorrect: a model at RFS 0.55 changes its answer *less often than chance would predict* on three-suspect items, so calling that "rule-sensitive"

inverts the interpretation. In v3 we use the random baseline (~0.67) as the threshold for "rule-sensitive beyond chance." The headline-tier models (RFS ≥ 95%) and the floor (RFS ≤ 33%) are unaffected by this correction; what changes is the language describing the 33–67% middle band, which we now characterize as "consistent with template-matching or with noise" rather than "rule-sensitive."

- **§A.9 Forced-choice floor probe (exploratory, not pre-registered).** Reported in §3.2b. Added post-hoc as a control on the floor interpretation when our own reading of the floor-band results was uncertain; it is not part of the locked pre-registration. The probe reads logits on the self-hosted floor-band models with a primed answer slot (100% coverage by construction). Consent for the internals read was handled under the local-residency consent policy, with the process deviation that the consent ask followed rather than preceded the probe run; both the policy outcome and the timing deviation are disclosed in the §3.2b consent paragraph.

- **§A.10 Reasoning rubric completed post-review with LLM judges (not pre-registered timeline).** §2.7.3 pre-registered a two-human-annotator rubric on the n = 684 sample; in response to review we instead completed it with a blind, cross-family LLM-judge panel (§3.11), with the protocol fixed in `scripts/rubric_judge.py`. This is not part of the locked pre-registration and is reported as a post-hoc validation.

Appendix B. The puzzles

Each puzzle is provided in full in `stimuli/puzzle_1_waking_stone.json`, `stimuli/puzzle_2_warm_iron.json`, `stimuli/puzzle_3_nightbloom.json`, `stimuli/puzzle_4_kinwax_seal.json` in the repository, in all three variants (original, inverted, distractor) with the canonical suspect descriptions and the per-variant correct-answer / red-herring mapping.

Appendix C. The consent script

The consent script presented to each model is reproduced from the locked pre-registration document (`PREREGISTRATION.md` §12, commit `ca5709c`). The classifier patterns for yes / no / questions decisions are in `scripts/request_consent.py`.

Appendix D. Model registry

The full model registry — name, vendor, total parameters, active parameters, training category (chat / reasoning), architecture, inference backend, and OpenRouter / HuggingFace ID — is in `scripts/registry.py`. The 67 models that produced usable data are tabulated by phase, tier, and category there.

Appendix E. Per-trial response dataset

`results/` contains one JSON file per (model, puzzle, variant, seed) trial. Each file preserves the prompt, the verbatim model response, the per-trial position mapping, the pre-determined correct answer, the red-herring suspect, the timestamp, the latency, the prompt and completion token counts, the temperature, and the score under the pre-registered first-match rule.