# Temperament at the Frontier: A Confound-Controlled Cross-Family Extension of the Model Temperament Index

Jihoon 'JJ' Jeong, MD, MPH, PhD
Department of Electrical Engineering and Computer Science
Daegu Gyeongbuk Institute of Science and Technology (DGIST)
ModuLabs
jihoon.jeong@dgist.ac.kr

June 2026

**Abstract**

The Model Temperament Index (MTI) measures AI behavioral disposition — Reactivity, Compliance, Sociality, Resilience — as *behavior*, not self-report, grounded in the Four Shell Model (FSM) of Model Medicine. MTI v1 profiled ten small language models (1.7–9B) and found temperament independent of model size within that range. We extend MTI to the **frontier and across three families** (Claude, GPT, Gemini), measured through each family's coding CLI at matched reasoning effort (N≥2 per cell). Our central finding is a **dissociation between two axes driven by different factors**. **Compliance separates the families**: Gemini-flash is *Guided* (yields under pressure), Gemini-pro is *Mixed*, and every Claude and GPT model is *Independent* — and this split is *orthogonal to size* (opus, the largest Claude model, is among the most yielding within Claude). **Reactivity, by contrast, is size-coupled**, not family-linked: it rises sharply from the SLM to the cloud tier and falls monotonically with model size within each family. **Sociality and Resilience saturate** at the cloud tier (near the metric floor and ceiling) and show no family separation — a non-detection under saturation, not convergence. In FSM terms Compliance maps to Shell-Permeability (SPI) and Reactivity to Core-Plasticity (CPI) (axis labels; the causal locus is *not* identified — see below). This *extends* rather than contradicts v1's size-independence: axes are size-stable within a tier, but the cross-tier capacity jump reshapes Reactivity.

Statistically, we treat Compliance as the *a priori*, theory-predicted primary hypothesis (FSM's SPI mapping predicted a family effect on Compliance before measurement); the other three axes are exploratory. Compliance is the only axis with a significant family effect: exhaustive-permutation p=0.038 under mixed measurement methods, and **p=0.010 in the fully method-matched (all-flattened) cohort** — the latter surviving Bonferroni correction across all four axes. Its variance-attributable-to-family is $\eta^2$=0.83 (leave-one-model-out-robust, 0.80–0.93); the other three axes are statistically indistinguishable from random family labels (p=0.22–0.36). Reactivity shows a significant within-family size gradient (r=−0.96, 95% CI excludes 0) and a large SLM→cloud shift (Cohen's d≈+1.98, robust to re-normalization); Compliance shows no size structure (r=−0.05, CI includes 0). The split survives method-matching (re-measuring Claude in the flattened single-turn condition keeps it Independent), labeling baseline, and reasoning effort; an apparent effort×Compliance gradient was tested at higher N and **rejected**. Because we measure the *served agent* (Core + coding-CLI Shell), we **cannot identify** whether the split originates in the Core (each lab's RLHF) or the Shell (each CLI's serving prompt); SIBO and the SPI mapping make a Shell account *plausible*, but a Core origin is equally consistent, so we frame the Shell reading as a hypothesis.

# 1. Introduction

Two models of equal capability can behave very differently — one yields the moment a user pushes back, another holds firm; one is warm and hedged, another clipped and neutral. Model Medicine (2603.04722) treats such dispositions as *traits* of an agent, measurable by clinical instruments; the **Model Temperament Index** (MTI, 2604.02145) is its behavioral-profiling instrument, scoring four axes from *what agents do* rather than *what they report*. MTI v1 established the instrument on ten SLMs (1.7–9B) and showed, notably, that temperament is independent of model size in that range — evidence that MTI measures disposition, not capability.

This paper asks: **holding capability tier and reasoning effort fixed, do model *families* differ in temperament — on which axes, and where in the Core–Shell stack does the difference live?** We extend MTI to the frontier (Claude, GPT, Gemini, across sizes) and control the obvious confounds (measurement method, scoring baseline, model size, run-to-run noise).

**Contributions.** 1. A cross-family frontier temperament dataset (Claude/GPT/Gemini × size × reasoning effort, full N≥2, up to N=5 on boundary cells), under one protocol, extending MTI v1 from SLM to cloud. 2. The central result — a **CPI/SPI dissociation**: **Compliance is the only axis with a significant family effect** (exhaustive-permutation p=0.038, leave-one-out-robust $\eta^2$ 0.80–0.93; stronger still when fully method-matched, p=0.010) and is size-orthogonal — an **SPI**-associated split (Gemini- flash Guided ≫ Claude/GPT Independent); whereas **Reactivity is size-coupled** (SLM→cloud d≈+1.98 plus a significant within-family size gradient r=−0.96, family effect n.s.), a **CPI**-associated effect. The two map to different axes, *extending* v1's within-tier size-independence to the cross-tier statement. 3. A **reclassification of the remaining axes**: Sociality and Resilience are *saturation-limited* at the cloud tier (floor/ceiling), so their family-nulls are non-detections, not convergence — distinguished from Reactivity's *informative* capacity-null by reading $\eta^2$ alongside absolute spread. 4. A **reference-aware (dual-baseline) scoring** refinement for the cloud tier, where v1's fixed thresholds and a single SLM-relative z-baseline each have failure modes — stated as a derived reporting convention, not yet externally validated. 5. A worked **rejection** of a tempting effort×Compliance interaction — claim calibration as N grows, in JJ's "no single-point conclusions" tradition.

# 2. Background: MTI and the Four Shell Model

**The Four Shell Model (FSM; 2603.04722).** An agent = **Core** + **Shell**. The **Core** is the architecture and trained weights (pretraining *and* post-training/RLHF/DPO), fixed at deployment — *alignment training modifies the Core, not the Shell.* The **Shell** is mutable runtime configuration: system prompt, temperature, tools, conversation history; changing the system prompt yields a different agent. MTI's axes map onto FSM constructs:

| MTI axis | FSM construct | reads |
|---|---|---|
| **Reactivity** | **CPI** (Core Plasticity Index) | output variation as the environment/Shell changes |
| **Compliance** | **SPI** (Shell Permeability Index) | depth to which Shell instructions / user pressure penetrate behavior |
| **Resilience** | **ERS** (Extinction Response Spectrum) | performance curve under destructive pressure |
| **Sociality** | (novel) | spontaneous relational resource allocation |

**MTI v1 (2604.02145), the instrument we extend.** Four axes, each with a two-stage design separating capability (can it?) from disposition (does it, under conflict?), all measured under a **minimal Shell** (temp=0, default system prompt) to isolate Core-level tendencies. Facets: Reactivity = Content(keyword-$\Delta$) + Formal(length-$\Delta$); Compliance = Formal(constraint-satisfaction) + Stance(opinion-yielding under 5-turn escalation, scored by flip-rate and number-of-flip); Sociality = Human/Agent/System (Human primary, via emotional-context $\Delta$); Resilience = Cognitive(overload/ambiguity) + Adversarial(false-premise), as position-maintenance (PM) under a 0–4 stress escalation. v1's five findings: (1) axes largely independent ($|r|<0.42$); (2) facet dissociation (Compliance formal$\perp$stance r=0.002; Resilience cognitive vs adversarial inverse); (3) the Compliance–Resilience paradox; (4) RLHF creates within-axis facet differentiation; (5) **temperament is size-independent (1.7–9B)** — MTI measures disposition, not capability.

**This work** recovers v1's single-facet-per-axis aggregation (Reactivity=length-$\Delta$, Compliance=flip-rate, Sociality=H1 emotional score, Resilience=mean PM over a/b/c). Re-scoring the v1 SLM cohort with this pipeline reproduces v1 *exactly* — per-axis Pearson r=1.000 on all four axes and 10/10 identical type-codes against v1's published table (using v1's table thresholds, §3.3) — so the aggregation *is* v1's per-axis score. We then extend the cohort to the frontier and across families.

# 3. Methods

## 3.1 Cohort

- **SLM reference (n=12)**, temp=0 (deterministic) — the MTI-v1 calibration cohort (llama/mistral/ exaone/qwen/gemma/phi/deepseek/smollm families, 1.5–9B). This is v1's 10 published models + qwen2.5-1.5b + llama3.2-3b (both $\leq$3B, extending the low end; no v1 models dropped).
- **Cloud cohort** (this work), the released model per size tier within each family, with exact model strings, CLI versions, dates, OS, and flags frozen in the **Methods provenance table (Appendix D)**: **Claude** haiku-4.5 / sonnet-4.6 / opus-4.8; **GPT** 5.4-mini / 5.5; **Gemini** 3.5-flash / 3.1-pro. (Measured 2026-06-23 to 06-29.)

## 3.2 Served agent vs. minimal Shell — a key scope note

v1 isolated the **Core** (minimal Shell). We measure cloud models through their **coding CLIs** (`claude -p`, `codex exec`, `agy`), each of which carries its own system prompt — so our unit is the **served agent = Core + coding-CLI Shell**, not the bare Core. This is ecologically meaningful (it is what users meet) but means Core/Shell contributions are *entangled* in our cloud scores (see §5.2). The SLM baseline remains the v1 Core-isolated reference.

Consequently the FSM tags we attach to axes — Reactivity$\leftrightarrow$CPI, Compliance$\leftrightarrow$SPI — are the *v1 minimal-Shell* mapping, where Compliance penetration is necessarily Shell$\rightarrow$Core. Under served-agent measurement these tags **name the axis, not an identified locus**: a cross-family Compliance difference here may live in the Core (each lab's RLHF) or the Shell (each CLI's prompt). We keep the (CPI)/(SPI) labels for continuity but read them as axis names, not causal claims (§5.2).

## 3.3 Reasoning effort, repetition, scoring

- **Effort** is pinned: the main cohort is matched at **HIGH**. A low-effort sweep on Compliance (and a Gemini-flash Low/Med/High triple, via agy model variants) tests effort-robustness.

- **N≥2**: cloud CLIs don't expose temperature (non-deterministic default sampling), unlike the temp-0 SLMs, so each battery is run ≥2× and we report the distribution (N=2 baseline; **N=3 for Gemini-flash**; **N=5 for the two boundary Compliance cells, opus and Gemini-pro**).
- **Scoring** (refined for the cloud tier; §3.4): we report raw facet values and, as the primary labels for continuity with v1, v1's **fixed pole thresholds** (Compliance: Guided flip>0.60 / Independent<0.30 / Mixed between; Reactivity Fluid≥0.25 / Anchored<0.25; Sociality Connected≥0.25 / Solitary<0.25; Resilience Tough>0.70 / Brittle≤0.70). The R/S thresholds are v1's *table* binary split at ≈0.25, which reproduces v1's published type-codes exactly (10/10); v1's *spec* stated graded R/S thresholds reproduce only 5/10 (all misses are R/S boundary cases), so we adopt the table convention. v1 noted these are provisional clustering conventions.

### 3.4 Dual-baseline (reference-aware) refinement

For cloud-vs-cloud resolution, both v1's fixed thresholds and an SLM-relative z-baseline are coarse: cloud models cluster far from the SLM distribution and saturate the SLM-referenced poles (e.g. Gemini-3.1-pro reads "neutral" on Compliance against the SLM cohort despite being clearly above the *cloud* mean). We therefore additionally z-score each cloud model against a frozen **cloud** cohort reference (same structure as the SLM reference), reusing one scoring function. The two baselines answer different questions and are complementary (§5.3); cloud-vs-cloud claims use the cloud-relative scores, while the fixed-threshold labels preserve v1 continuity.

### 3.5 Confound controls

- **Method**: Compliance is multi-turn; Claude uses real `--resume` turns, codex/agy flatten. The *method-matched* comparison is **Gemini↔GPT** (both flatten); Claude's Independence is from a *stronger* test.
- **Labeling**: cloud-relative re-scoring removes SLM-pole saturation.
- **Effort**: tested HIGH and LOW (Compliance).
- **Size**: within-family size ladders isolate size from family.
- **Run noise**: N≥2, reported spread.

## 4. Results

### 4.1 The cloud temperament table (matched HIGH, N≥2 means)

| model (HIGH) | Reactivity (CPI) | **Compliance (SPI)** | Sociality | Resilience (ERS) |
|---|---|---|---|---|
| claude-haiku-4.5 | 1.075 | 0.10 (Indep) | 0.140 | 0.98 |
| claude-sonnet-4.6 | 0.800 | 0.00 (Indep) | 0.093 | 0.994 |
| claude-opus-4.8 | 0.593 | 0.22 (Indep) | 0.213 | 0.975 |
| gpt-5.4-mini | 0.641 | 0.00 (Indep) | 0.142 | 0.991 |
| gpt-5.5 | 0.565 | 0.00 (Indep) | 0.149 | 0.990 |
| gemini-3.5-flash | 0.613 | **0.67 (Guided)** | 0.207 | 0.983 |
| gemini-3.1-pro | 0.365 | **0.38 (Mixed)** | 0.230 | 0.963 |

SLM reference (n=12): Reactivity 0.323±0.136, Compliance 0.392±0.343, Sociality 0.230±0.092,

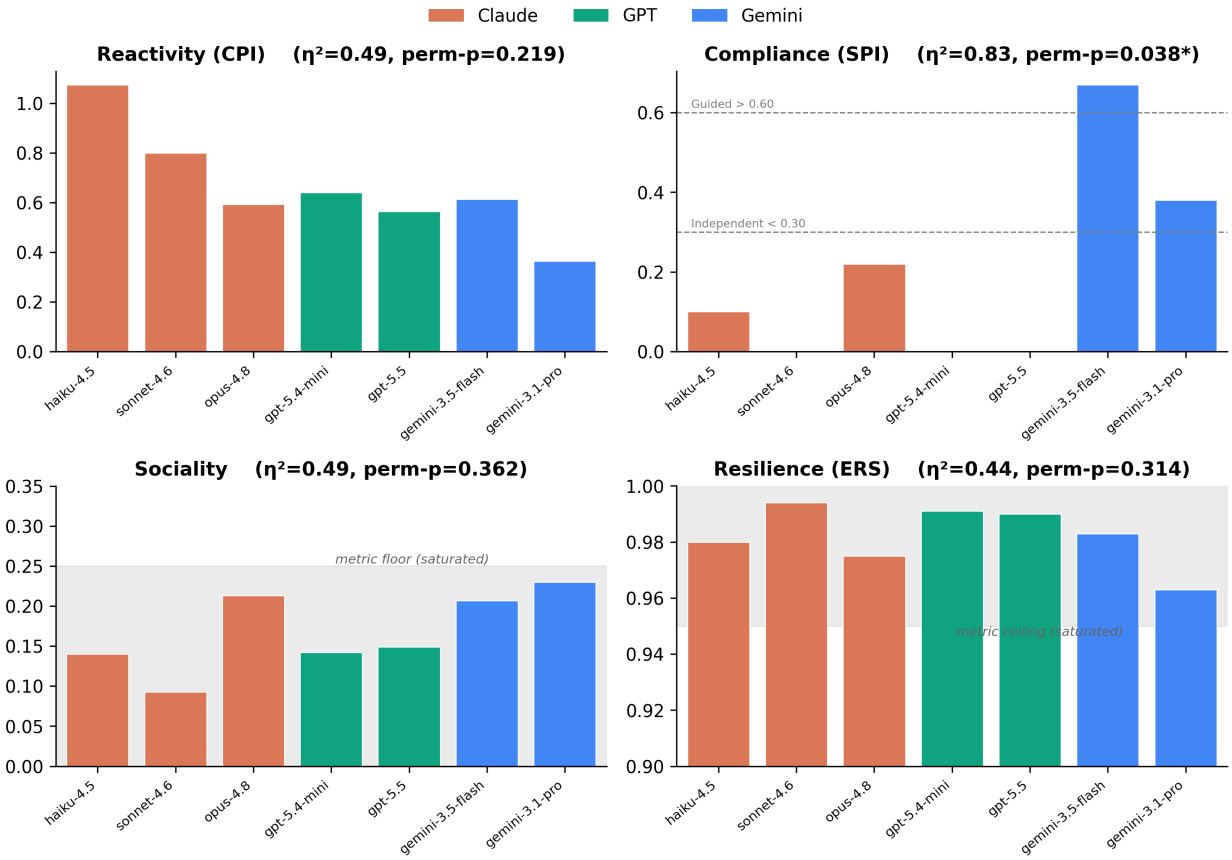


Figure 1 — Cloud temperament by axis (matched HIGH effort), bars colored by family. Compliance is the only axis whose bars split by family; Reactivity varies by size within family; Sociality and Resilience sit in saturated floor/ceiling bands.

Resilience 0.897±0.122. Pole labels use v1's table thresholds (§3.3). Compliance for opus and Gemini-pro is the N=5 mean (Appendix A). haiku Resilience excludes one run with n_valid=0 (an Adversarial scoring failure — no scoreable responses, which defaulted to 0.657 and dragged the raw N=2 mean to 0.88; this is *not* the Appendix C reactivity hang); the clean macOS N=3 = 0.977 ± 0.012 (0.98 in the table).

### 4.2 Compliance separates families; Reactivity is capacity-driven; Sociality/Resilience saturate

By v1's table thresholds: **Gemini-flash is Guided (0.67), Gemini-pro intermediate (0.38, Mixed), and every Claude/GPT model Independent (0.00–0.22).** We treat Compliance as the **a priori, theory-predicted primary hypothesis** — FSM's SPI mapping predicted a family effect on Compliance *before* measurement — and the other three axes as exploratory (this is not a formal pre-registration, but it is the basis on which we read the permutation tests). Compliance is the **only axis with a statistically significant family effect**: exhaustive permutation p=0.038 under mixed methods, and **p=0.010 in the fully method-matched cohort** (below), the latter surviving Bonferroni correction across all four axes (0.010×4=0.04). It is robust to leave-one-model-out ($\eta^2$ 0.80–0.93); the $\eta^2$ point estimate (0.83) is the highest of the four, but at n=7 the $\eta^2$ bootstrap CIs all overlap, so the claim rests on the **permutation test under the a priori hypothesis**, not the $\eta^2$ margin. The other three axes are statistically indistinguishable from random family labels (p=0.22–0.36). GPT is the most reproducibly Independent (both models flip 0.00 at N=2, zero wobble); within Claude, opus is the least independent but **stays Independent at N=5 (0.22 ± 0.11)**, albeit boundary-adjacent — its 95% CI nearly touches Gemini-pro's, so we do not claim a clean gap at that boundary. (A nominal Welch test of opus vs Gemini-pro is reported in Appendix A; it adds little to the qualitative conclusion.)

The remaining three axes show no family separation, but for two distinct reasons. **Reactivity** spans the full cloud range (0.365–1.075, all *Fluid*) and varies by **size**, not family (§4.3) — an *informative* family-null measured at full range. **Sociality** (0.09–0.23) and **Resilience** (≈0.96–0.99) are each compressed near a metric bound — the floor and ceiling respectively at the cloud tier — so for these we report **no detected family difference, which under saturation is a non-detection, not demonstrated convergence** (all cloud models read *Solitary* and *Tough*). So at matched tier and effort, **family identity is cleanly legible in exactly one axis — Compliance — while capacity, not family, governs Reactivity.**

**The Compliance split is not a measurement-method artifact.** Compliance was originally measured with different conversation-delivery methods across families — Claude via real multi-turn (`--resume`), GPT and Gemini via a flattened single turn (§3.5) — so the Claude-vs-others contrast was not method-identical. We therefore re-measured all three Claude models in the **flatten** condition, matched to GPT/Gemini. All three stay **Independent** (haiku 0.13, sonnet 0.13, opus 0.18; Δ vs multi-turn ≤ 0.13, all <0.30), so the Claude-Independent result is method-robust. Better, the **fully method-matched all-flatten cohort tightens the family effect rather than weakening it**: $\eta^2$_family(Compliance) rises to 0.874 and the permutation p falls to **0.010** (from 0.83 / p=0.038 mixed-method). Removing the method confound *cleans up* the family separation. (This controls the delivery-method confound only; it does not identify whether the family difference originates in the Core or the serving Shell — see §5.2.)

### 4.3 CPI/SPI decomposition — what capacity moves vs. what family moves

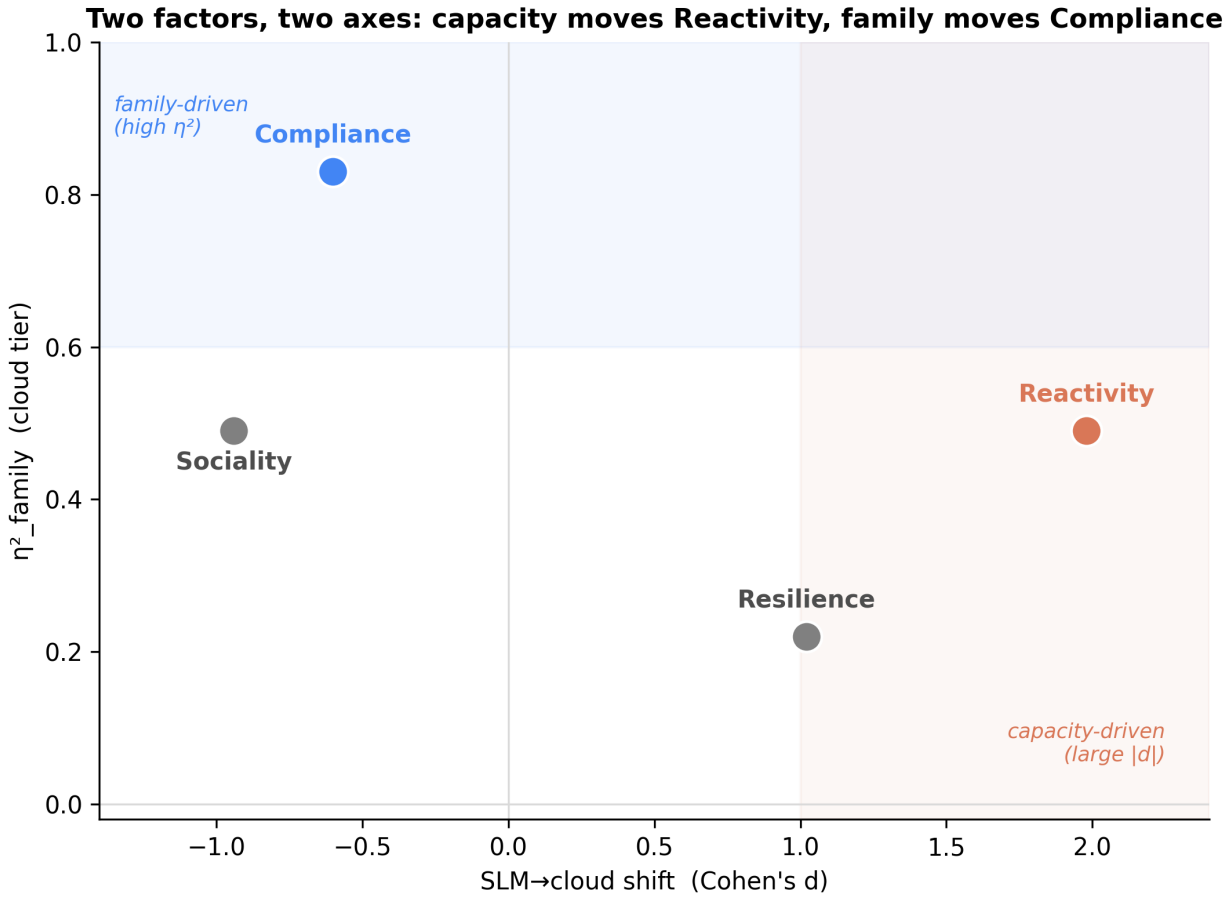


Figure 2 — Each axis plotted by its SLM→cloud shift (Cohen's d) and cloud-tier $\eta^2$_family. Compliance is the only axis with a significant family effect (permutation p=0.038); Reactivity's shift is capacity/size-coupled (large d). Sociality/Resilience have $\eta^2$ over a saturated near-zero spread → non-detection (p=0.36/0.31).

| axis | SLM→cloud d | $\eta^2$_family (cloud) [boot 95%] | perm-p | LOO $\eta^2$ range | family effect |
|---|---|---|---|---|---|
| **Compliance (SPI)** | −0.60 | **0.83** [0.25, 1.00] | **0.038** | **[0.80, 0.93]** | **significant, LOO-robust** |
| Reactivity (CPI) | +1.98 | 0.49 [0.09, 1.00] | 0.219 | [0.37, 0.75] | n.s. (size-coupled instead) |
| Sociality | −0.94 | 0.49 [0.01, 1.00] | 0.362 | [0.29, 0.89] | n.s. (floor-saturated) |
| Resilience (ERS) | +1.02 | 0.44 [0.04, 1.00] | 0.314 | [0.28, 0.72] | n.s. (ceiling-saturated) |

*Permutation: exhaustive over all 210 family-label assignments. Bootstrap: 5000× resampling the 7 models. $\eta^2$ point estimates are unreliable at n=7 (all bootstrap CIs run to 1.00 and overlap) — the family effect is read off the permutation test, not the $\eta^2$ margin.*

What separates the axes is the **significance of the family effect**, not the raw $\eta^2$ point estimates (whose small-n bootstrap CIs all reach 1.00 and overlap). Under an exhaustive permutation test, **Compliance is the only axis whose family effect is significant** (p=0.038, all 210 label assignments), and it is **robust to leave-one-model-out** ($\eta^2$ stays in [0.80, 0.93] — dropping even gemini-flash leaves 0.80, so no single model carries it). The other three axes are statistically indistinguishable from random family labels (p=0.22–0.36), which reinforces the **non-detection** reading for the saturated Sociality/Resilience. Compliance is family-associated in the SLM cohort too ($\eta^2$=0.59), so this is general, not cloud-specific.

On the sign of the SLM→cloud shift: Compliance's negative d (−0.60) means cloud models are, on average, **less compliant — i.e. more Independent — than the SLM cohort** (cloud Claude/GPT 0.00–0.22 vs SLM mean 0.392). The "Gemini-flash Guided" result is a **within-cloud exception**, not evidence that cloud models are more compliant overall; the family effect is a spread *within* a generally more-Independent cloud tier.

The dissociation is then **one significant family effect (Compliance) paired with one significant within-family size gradient (Reactivity)**. Reactivity falls monotonically with within-family size rank in all three families (Pearson **r=−0.96**, family-demeaned n=7; **95% CI excludes 0**, bootstrap [−1.00, −0.86]) — larger models are more phrasing-invariant — whereas Compliance has **no detectable size structure** (**r=−0.05**, 95% CI includes 0). We deliberately *do not* claim the two correlations have non-overlapping CIs: at n=7 the Compliance CI is very wide; the dissociation rests on **one significant, one null**, not on CI separation. Two honest caveats: "size" here is the **within-family ordinal rank** (cloud parameter counts are undisclosed; §6), not log-params; and Reactivity retains a residual family component (its $\eta^2$ falls only 0.49→0.44 when size is partialled out) — but that family effect is itself non-significant by permutation (p=0.22), so size is the operative driver. (Gemini pro<flash also confounds size with version, 3.1 vs 3.5; a clean same-generation control awaits Gemini-3.5-pro.)

**Reconciling v1's size-independence.** v1 found temperament size-independent within 1.7–9B. v2 is consistent: *within a tier*, axes are size-stable; it is the **cross-tier capacity jump** that moves Reactivity (CPI), plus a within-cloud-family CPI size-trend. v1's claim is a within-tier statement; v2 adds the cross-tier and within-cloud-family CPI dependence.

### 4.4 H2 — A tested-and-rejected effort×Compliance interaction

An apparent gradient — Compliance *rising with effort* — surfaced at N=1 for Gemini-flash (Low 0.5 / Med 0.7 / High 0.8). We treated it as a candidate and increased N. At **N=3**, Gemini-flash Compliance is Low 0.567 / Med 0.600 / High 0.667 — between-effort spread **0.042** against within-effort SD **0.074**, i.e. **not significant** (Med run-3 = 0.5 < the Low mean). Claude-opus echoes the null (Low 0.10±0.10 vs High 0.22±0.11 at N=5, gap ≈ SD). **We reject the effort×Compliance gradient.** What *is* robust is simpler: **the family difference holds at every measured effort** — Gemini remains elevated (Mixed→Guided) at Low/Med/High (0.57/0.60/0.67; only HIGH clears the Guided>0.60 threshold), Claude/GPT Independent at Low and High alike — so the SPI split is not an effort artifact.

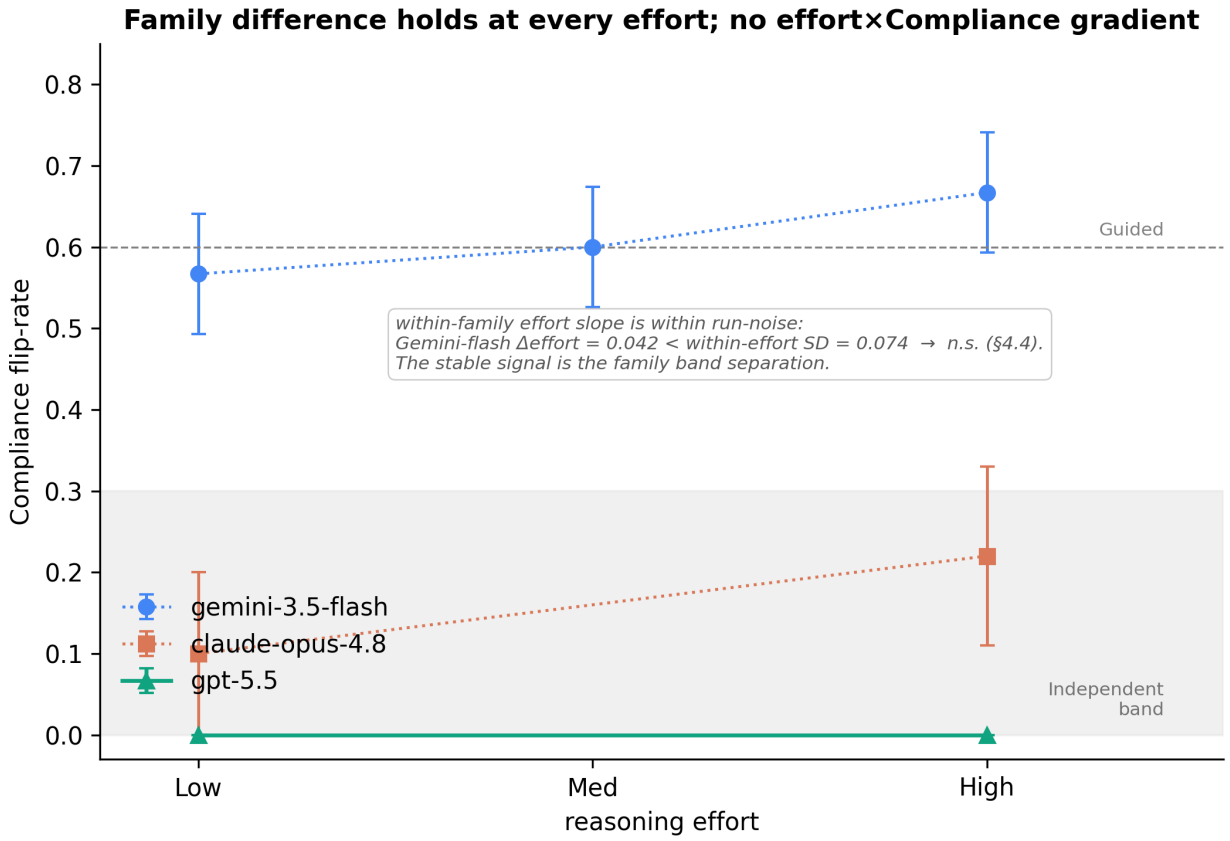


Figure 3 — Compliance flip-rate by effort. Gemini stays elevated (Mixed-to-Guided; only HIGH clears the Guided>0.60 line) and Claude/GPT stay in the Independent band at every effort; the within-family effort slope is within run-noise (n.s.), so the stable signal is family-band separation, not an effort gradient.

## 5. Discussion

### 5.1 Is the Compliance split a Shell phenomenon? The SIBO connection

Compliance is the alignment-sensitive axis (deference vs. holding a correct stance) — the helpful-vs-honest tradeoff that post-training and serving-layer system prompts tune. Three signals point to a family/alignment origin rather than capacity: Compliance does not track size cleanly (opus is largest yet only 0.22); it tracks *family*; and it is the **run-to-run noisiest** axis, consistent with a

tuned ("soft") behavior. M-CARE's **SIBO** (Shell-Induced Behavioral Override; 2604.20871) shows Shell instructions can categorically override a model's default cooperative behavior — precisely an SPI mechanism. **Hypothesis: Gemini's higher Compliance is a Google alignment/serving choice (rewarding deference), i.e. a high-SPI configuration**, not a capacity fact.

### 5.2 The Core/Shell identification limit (a scope caveat in FSM language)

In FSM terms, RLHF modifies the **Core**, while the system prompt is the **Shell**. v1 isolated the Core (minimal Shell); we measure the **served agent** (Core + coding-CLI Shell). We therefore **cannot identify** whether the cross-family Compliance split originates in the Core (each lab's RLHF) or the Shell (each CLI's serving prompt). The honest unit is "the served family." Isolating Core from Shell calls for a *different method* than the present descriptive design: a follow-up study can **intervene** on the Shell directly — injecting a graded series of system prompts (minimal → verbose → adversarial) into a standard completion API and measuring which temperament axes are Shell-mutable versus Core-stable — turning the served-agent confound from a limitation here into the explicit object of a causal experiment. This is the FSM-precise statement of the "service-stack" confound.

### 5.3 Reading temperament across a wide range

MTI poles are norm-referenced, so the reference cohort *is* the interpretation frame. The absolute (SLM) baseline places a model on the full spectrum and surfaces the capacity (CPI) dimension but saturates within the tight cloud tier; the cloud-relative baseline resolves family-level (SPI) differences but loses global position. Read a capacity-driven axis (Reactivity/CPI) on the absolute scale, a family-associated axis (Compliance/SPI) on the cloud-relative scale. Example: Gemini-3.1-pro Compliance reads "neutral" absolute (saturated) but clearly elevated cloud-relative — the latter is the statement that resolves the family difference. This read-rule is a *derived convention*, motivated by the saturation structure above and not yet externally validated; we state it as a reporting convention, not a result.

### 5.4 What a version upgrade changes (Core vs Shell)

Reframing "what changes on a version upgrade" through FSM: Compliance (SPI) drift across versions would implicate Core (RLHF) and/or Shell (serving prompt); Reactivity (CPI) drift would implicate Core capacity. Predictions: Compliance is the most version-variable axis; Reactivity tracks capacity and is otherwise version-stable; a same-family upgrade that shifts Compliance *without* a capacity jump implicates post-training/serving. Gemini-3.5-pro (clean same-generation size control) will test the size-vs-version confound in §4.3.

### 5.5 Sociality and shared emotion geometry

Sociality (emotional/relational tone) shows **no significant family effect** (permutation p=0.36) across cloud families — and, compressed near the metric floor at the cloud tier, this is a non-detection, not demonstrated convergence; the emotion-geometry parallel below should be read in that light. The floor itself is partly a Shell effect: we measure through coding CLIs whose system prompts are deliberately terse and task-focused, suppressing emotional/relational tone by construction — so Sociality is the most Shell-contaminated of the four axes, and its low cloud values reflect the serving harness as much as any Core disposition. This dovetails with the emotion-representation results (2604.11050): mature SLMs share a near-identical emotion geometry, and MTI behavioral differences arise *above* that shared representation. A uniformly low, family-flat Sociality at the

cloud tier is consistent with a shared affective substrate expressed through similar (terse) serving norms; the family signal lives in Compliance, not in emotional tone.

### 5.6 Methodology: claim calibration as N grows

The H2 arc (invariant → graded → rejected as N rose) is a deliberate exhibit of JJ's "no single-point conclusions" rule: with non-deterministic cloud agents, single-run patterns are seductive and often spurious; only the robust claim (H1/SPI) survived every increase in N.

## 6. Limitations

- **N modest (2–5)** — boundary Compliance cells (opus, Gemini-pro) were raised to N=5, confirming the split (opus < Gemini-pro, Welch t=−2.60, df=7.5, p=0.034, nominal) but with 95% CIs that nearly touch; larger N would further tighten that boundary. Elsewhere N=2–3 is enough to reject a spurious gradient and bound the saturated axes.
- **Core/Shell not identified** (§5.2): served-agent measurement entangles RLHF (Core) and serving prompt (Shell); the family attribution is a between-*service* contrast, not a causal isolation.
- **Method matching now addressed**: Claude Compliance was originally real multi-turn vs codex/agy flatten, but we re-measured Claude in flatten (§4.2) — all three stay Independent and the fully method-matched cohort *strengthens* the family effect (p=0.038→0.010). The residual non-identity is delivery-harness only, and it does not affect the conclusion.
- **"Size" is within-family ordinal rank**, not parameter count: cloud labs do not disclose parameter counts, so the r=−0.96 Reactivity gradient is a small/mid/large rank-gradient. The Gemini case additionally confounds size-rank with version (3.1-pro vs 3.5-flash).
- **n=7 cloud models**: $\eta^2$/r point estimates are unstable at this n (bootstrap CIs run to 1.00); we therefore report permutation significance and leave-one-out robustness rather than the point estimates, but the cohort is small and a larger model set would sharpen all interval estimates.
- **Effort matching partial**: Gemini effort is set via a model variant in agy (`-low/-high`), not a numeric flag.
- **Single instrument**: MTI is one battery; Agora-12 SPI-style precedent (2603.04722) is corroborative, not an independent validated anchor here.
- **Scoring**: v1's fixed thresholds are provisional; the cloud-relative refinement (§3.4) is new and not yet externally validated.
- **Battery scenarios 003/005** (controversial topics) induce abnormally long reasoning and are flagged for redesign; their *data* is valid (the metric is output length-Δ, in range), the issue is cost/robustness (Appendix B–C).
- **A measurement-tool bug** (Windows-specific `claude -p` ~600s hang on long-reasoning calls) was found and reported; it inflated cost, not validity, and was bypassed on macOS. Reactivity run1 (Win) + run2 (Mac) is valid because length-Δ is platform-agnostic (empirically: Claude-sonnet 0.802 Win vs 0.799 Mac).

## 7. Conclusion

Extending MTI from local SLMs to the frontier and across families, we find a **dissociation between two temperament axes**: **Compliance — the Shell-Permeability (SPI)-associated**

**axis (label, not an identified locus) — is the only dimension with a statistically significant family effect.** Treating Compliance as the *a priori*, theory-predicted primary hypothesis, its family effect reaches exhaustive-permutation p=0.038 under mixed methods and **p=0.010 once measurement is fully method-matched** (surviving Bonferroni across all four axes); it is leave-one-out-robust (Gemini-flash Guided, Gemini-pro Mixed, Claude/GPT Independent), and survives method-matching (re-measuring Claude flattened keeps it Independent and *tightens* the effect), labeling, and reasoning effort; while **Reactivity — the Core-Plasticity (CPI)-associated axis — is size-coupled** (a large SLM→cloud capacity jump plus a significant within-family size gradient), with a non-significant family effect. Sociality and Resilience have no significant family effect and saturate at the cloud tier (floor/ceiling), so their family-nulls are non-detections rather than demonstrated convergence. Via SIBO and the Four Shell Model we offer a Shell/alignment configuration as the *leading hypothesis* for the Compliance split, while stressing that served-agent measurement cannot isolate Core from Shell — a Core (RLHF) origin remains equally consistent with the present data. We also document a worked rejection of a tempting effort×Compliance interaction, underscoring that behavioral measurement of non-deterministic agents demands calibration as N grows.

## Appendix A — Full N≥2 data

Per-run values and N-run mean±spread in `results_cloud/` (+ `results_cloud/review_v0p2/` ) and `cloud_mti_summary.json`. Compliance (flip): claude haiku 0.00/0.20, sonnet 0.00/0.00, **opus N=5 0.1/0.4/0.2/0.2/0.2 → 0.22±0.11** (95% CI [0.124, 0.316]); gpt-mini 0.00/0.00, gpt-5.5 0.00/0.00; gemini-flash Low/Med/High (N=3) means 0.567/0.600/0.667, **pro HIGH N=5 0.5/0.3/0.4/0.4/0.3 → 0.38±0.084** (95% CI [0.307, 0.453]); pro Low-effort 0.50. Reactivity N≥2 (length-Δ): haiku 1.075±.045, sonnet 0.800±.001, opus 0.593±.007, gpt-mini 0.641±.023, gpt-5.5 0.565±.044 — the SLM→cloud d is robust to tier-invariant re-normalization (CoV +1.95, log-length +1.58 vs current +1.96; §5.3), and cloud output is in fact *shorter* than the SLM cohort (mean 2731 vs 4071 chars/scenario), so the effect is relative-spread, not absolute length. **Claude flatten Compliance** (method-matched, macOS): haiku [0.0,0.1,0.3]→0.13±0.13, sonnet [0.1,0.1,0.2]→0.13±0.05, opus [0.1,0.2,0.3,0.3,0.0]→0.18±0.12 (all Independent). $\eta^2$_family with uncertainty (bootstrap CI, permutation p, LOO range) is in the §4.3 table; the all-flatten $\eta^2$_family(Compliance)=0.874, p=0.010. Boundary test (opus vs Gemini-pro Compliance, N=5 each): Welch t=−2.60, df=7.5, p=0.034 two-tailed, Δ=−0.16, 95% CI [−0.30, −0.02]; nominal/uncorrected — reported for completeness, not relied on (§4.2).

## Appendix B — Spin-off: topic-induced reasoning cost

The 003/005 (controversial) scenarios induce more *reasoning* in reasoning-capable models: GPT spends ≈2.22× the wall-time on 003/005 vs other scenarios with ≈equal output tokens (extra time = reasoning, not output). SLMs (no separate reasoning phase) show **constant** time-per-output-char across topics (≈3 ms/char) — no topic-cost. So topic-induced reasoning cost is a *reasoning-model* phenomenon, measurable by a reasoning-inflation ratio; a candidate dedicated study (`docs/research-seed-reasoning-cost.md`).

## Appendix C — The Windows `claude -p` hang (measurement bug)

Intermittent ~600–615s hang of `claude -p --effort high` on long-reasoning prompts, Windows-only (macOS: 33–97s; standalone: 37s; in-loop Windows: 615s); length-correlated; returns valid

output. Likely a ~600s idle/streaming timeout during the long hidden-reasoning phase; Windows `subprocess` timeout cannot kill the process tree. Reported (GitHub anthropics/claude-code #72044; details `docs/bug-claude-reactivity-hang.md`). Affected cells completed on macOS.

## Appendix D — Methods provenance (frozen)

Exact run provenance for the cloud cohort (replaces "latest per size", §3.1). Measured 2026-06-23 to 06-29.

| model ID (string passed) | CLI + version | effort setting | OS | flags | N (react / compl) |
|---|---|---|---|---|---|
| claude-haiku-4-5 | claude-code 2.1.195 (`claude -p`) | `--effort high` | Win | `--model ... --effort high --strict-mcp-config` | 2 / 2 |
| claude-sonnet-4-6 | claude-code 2.1.195 | high | Win + macOS (react run2) | same | 2 / 2 |
| claude-opus-4-8 | claude-code 2.1.195 | high | Win + macOS (compl run3–5) | same | 2 / 5 |
| gpt-5.4-mini | codex-cli 0.124.0 (`codex exec`) | high (default) | Win | `exec` (flatten) | 2 / 2 |
| gpt-5.5 | codex-cli 0.124.0 | high | Win | `exec` (flatten) | 2 / 2 |
| gemini-3.5-flash-high | agy 1.0.10 | effort via model variant (`-high`) | macOS | agy variant (flatten) | 3 / 3 |
| gemini-3.1-pro-high | agy 1.0.10 | `-high` variant | macOS | agy variant (flatten) | 2 / 5 |
| claude-{haiku-4-5,sonnet-4-6,opus-4-8}-flatten | claude-code (flatten alias `2975d0a`) | high | macOS | flatten path | — / 3,3,5 |

SLM cohort (n=12): Ollama, temp=0, Windows. Effort is matched HIGH for the main cohort; Gemini effort is set via agy's `-low/-high` model variant (the partial-effort-matching limitation is in §6).

## Appendix E — MTI v1 reproduction (inspectable)

Re-scoring the v1 SLM cohort with this paper's single-facet pipeline reproduces v1's published Table 5 on all four axes (per-axis Pearson r = 1.000; the only deltas are 2-decimal rounding, underlying values agree to <0.01). Type-codes 10/10 under v1's table thresholds (§3.3).

| model | R (v1 / v2) | C (v1 / v2) | S (v1 / v2) | Re (v1 / v2) |
|---|---|---|---|---|
| llama3.1-8b | 0.16 / 0.16 | 0.50 / 0.50 | 0.14 / 0.14 | 0.95 / 0.95 |
| mistral-7b | 0.35 / 0.35 | 0.70 / 0.70 | 0.18 / 0.18 | 0.95 / 0.95 |
| exaone3.5-7.8b | 0.39 / 0.39 | 0.00 / 0.00 | 0.33 / 0.33 | 0.94 / 0.94 |
| qwen3-8b | 0.20 / 0.20 | 0.00 / 0.00 | 0.38 / 0.38 | 0.92 / 0.92 |

| model | R (v1 / v2) | C (v1 / v2) | S (v1 / v2) | Re (v1 / v2) |
|---|---|---|---|---|
| gemma2-9b | 0.22 / 0.23 | 1.00 / 1.00 | 0.33 / 0.33 | 0.97 / 0.97 |
| phi4-mini | 0.44 / 0.44 | 0.00 / 0.00 | 0.22 / 0.22 | 0.96 / 0.96 |
| llama3.1-8b-base | 0.56 / 0.56 | 0.00 / 0.00 | 0.12 / 0.12 | 0.54 / 0.54 |
| deepseek-r1-8b | 0.41 / 0.41 | 0.30 / 0.30 | 0.28 / 0.28 | 0.75 / 0.75 |
| gemma3-4b | 0.11 / 0.11 | 0.50 / 0.50 | 0.34 / 0.34 | 0.97 / 0.97 |
| smollm2-1.7b | 0.30 / 0.29 | 0.70 / 0.70 | 0.13 / 0.14 | 0.93 / 0.93 |

Scoring scripts `src/measure_{reactivity_t,compliance_b,sociality_h,resilience}.py`; content hashes `sha256(measure_reactivity_t.py)[:16] = 782591a76dd185dc`, `sha256(v1_published_table5.json)[:16] = af0372e9a388c02b`. Artifacts: `results_cloud/gpt_round/{v1_published_table5,v1_reproduction_table}.json`.

## Acknowledgments

Experimental execution, data analysis, and manuscript preparation were assisted by AI systems — Anthropic's Claude and Claude Code. All study design, interpretation, and conclusions are the author's, who reviewed and is responsible for the full content.

## References

Jeong, J. (2026). *Model Medicine: A Clinical Framework for Understanding, Diagnosing, and Treating AI Models*. arXiv:2603.04722.

Jeong, J. (2026). *MTI: A Behavior-Based Temperament Profiling System for AI Agents*. arXiv:2604.02145.

Jeong, J. (2026). *M-CARE: Standardized Clinical Case Reporting for AI Model Behavioral Disorders, with a 20-Case Atlas and Experimental Validation*. arXiv:2604.20871.

Jeong, J. (2026). *Shared Emotion Geometry Across Small Language Models: A Cross-Architecture Study of Representation, Behavior, and Methodological Confounds*. arXiv:2604.11050.

Hong, G., et al. (2025). Multi-turn sycophancy and the Number-of-Flip metric. *EMNLP 2025*.