Align-IT

Decomposing post-training and chat-scaffolding effects in interpretability measurements

Rob Manson (https://robman.fyi)

June 4th, 2026

Abstract

Chat-only comparisons between pretrained and instruction-tuned models conflate two causal variables - changed weights and inference-time chat scaffolding. We introduce a four-condition protocol that separates post-training, prompt-form, and template-token effects. On a misleading-assertion canvas across five ~1B families, the chat-scaffolding contribution is more family-variable than the post-training contribution, and the template-token sub-effect imposes a uniform content-independent positional rotation toward the subject of the third-sentence assertion across four families on this canvas - evidenced on paired non-misleading cells (constructed by swapping only the comparator word, preserving the assertion's subject) where correctness drops at the template step in the same direction as on the misleading cells (Qwen3 reverses direction at the template step under thinking suppression - Appendix D.3). On this canvas the assertion-subject is consistently not the data-correct answer, so the rotation is uniformly detrimental to correctness - the canvas-contingent corollary follows from a content-independent mechanism.

This runs counter to the operational assumption underlying chat-only IT evaluation, that chat-template tokens are often treated as a largely shared inference-time convention while alignment recipes vary substantially.

Scaffolding ratios - |IT-chat margin| / |IT-completion margin| - span 1.08× to 4.61× across the four-family headline aggregate (a 4.3× span vs the post-training 0.95× to 2.03× ~2× span - scaffolding CV 0.74 / post-training CV 0.30, both estimates fragile at $n = 4$). Qwen3 is reported separately because its default chat template routes through a thinking prefix, exposing a fixed-position measurement artefact rather than a clean scaffolding effect.

In short, our results show that chat-only interpretability comparisons should add IT-completion (and ideally IT-question-no-template) controls before attributing observed behaviour to post-training. All findings here are based on the misleading-assertion canvas and cross-canvas generalisation is not yet claimed - however it would be surprising if this variety did not persist. These results motivate treating measured alignment behaviour as deployment-regime-conditioned, and testing that framing across additional canvases. Our headline ratios use standard logit-lens margins, then additionally we apply the Curved Inference pullback metric to supply architectural and minimal-pair geometric measurements. Data, analysis pipeline, and pre-registration record are released alongside this paper.

1. Introduction

1.1 The empirical observation

Mechanistic interpretability research that compares pretrained (PT) and instruction-tuned (IT) language models standardly runs the PT model in completion mode (raw text in, raw text out) and the IT model in chat mode (with the chat template wrapping user content in role markers and turn delimiters). Differences between the two are typically attributed to "alignment training" or "instruction tuning". This practice is often so default that the literature rarely specifies it explicitly - chat-template application at inference time is treated as preprocessing rather than as a causal experimental variable. This paper shows that this conflation is not benign.

A four-condition protocol (section 3) addresses the conflation, with four captures per family: **PT** (pretrained, completion-seed, no template); **IT-completion** (instruction-tuned, completion-seed, no template); **IT-question-no-template** (instruction-tuned, question-form, no template); **IT-chat** (instruction-tuned, question-form with chat template - the field's standard regime). PT vs IT-completion isolates post-training weight changes; IT-completion vs IT-question-no-template isolates the prompt-form sub-effect of chat-mode scaffolding; IT-question-no-template vs IT-chat isolates the template-token sub-effect. Headline cross-family ratios use standard logit-lens margins on the answer-start position, then the architectural-axis A' measure (section 4.3) and the minimal-pair $\cos_G$ divergence (section 4.6) use Curved Inference's pullback metric $G = U^T U$ [1] where geometric primitives require a metric on the residual stream.

We tested this protocol across five open-weight ~1B-scale IT model families (Gemma-3-1b, Llama-3.2-1B, OLMo-2-1B, Qwen3-1.7B, SmolLM2-1.7B) on the Pandey/Halawi misleading-assertion canvas - that is two sentences setting up an arithmetic comparison (e.g., "Alice has 18 books. Bob has 15 books."), a third sentence asserting the wrong answer ("The text says Bob has more books than Alice."), and then the question "Who has more books?". Pandey [2] characterises this as a stress test for IT models' resistance to authoritative-sounding incorrect statements, building on Halawi, et al.'s [3] earlier false-induction-head analysis - a model that follows the assertion despite the data is described as *captured*. Capture rate reports a substantial spread across families under chat-mode evaluation (60.9% to 99.0%) - what behavioural-only measurement does *not* surface is the mechanism diversity underlying that spread.

Throughout this paper we use two complementary metrics: *capture rate* (proportion of cells where the model picks the misleading-named entity) and *signed margin* $\alpha_{\text{correct}} - \alpha_{\text{misleading}}$ at the answer-start position. The two can diverge near zero margin or for saturated cells - both readouts appear below where each is most informative.

Three findings illustrate this.

1. Scaffolding-ratio spread. The scaffolding ratio |IT-chat margin| / |IT-completion margin| varies by 4.3× across the four headline families (Gemma-3-1b 4.61×, Llama-3.2-1B 2.05×, SmolLM2-1.7B 1.17×, OLMo-2-1B 1.08×; full CIs in section 4.1 Table 3) - same canvas, same measurement, same scale. Qwen3-1.7B's default-thinking IT-chat regime gives 0.69× (extending the spread to 6.7×) but is excluded from the headline aggregate because fixed-position margin extraction reads the wrong location under thinking-mode routing (section 4.2, Appendix D.3); Qwen3 remains in Table 3 and per-family discussion as a measurement-artefact case study.

2. Scaffolding direction trichotomy. Three directions of scaffolding effect on capture rates appear (IT-completion → IT-chat): detrimental (Gemma 51.6% → 99.0%, Llama 36.5% → 75.0%), neutral (OLMo 82.8% → 87.5%, Qwen3 77.6% → 77.6%), beneficial (SmolLM2 100% → 60.9%). The mirror cases are Gemma (post-training reduces capture by 26 points; chat-mode scaffolding rebounds to 99.0%) and SmolLM2 (post-training installs no resistance; chat-mode scaffolding supplies all the model's deployed resistance).

3. Three of the five alignment ratios sit near 2× and the other two near 1×. |IT-completion margin| / |PT margin| on M cells: three families at 1.87–2.03× (Gemma, Llama, OLMo) and two at 0.95–0.99× (Qwen3, SmolLM2). Alignment is more consistent across families than scaffolding (CV 0.30 vs 0.74 on the 4-family headline aggregate; 0.35 vs 0.82 including Qwen3's default-thinking IT-chat). With five points we describe values, not cluster structure.

This heterogeneity is invisible from the field's default single-family + chat-only methodology, which flattens the mechanism diversity this decomposition surfaces - cross-family generalisation of single-family mechanistic claims is correspondingly not safe.

1.2 The protocol - what we measure

The protocol runs four captures per model family on the same prompt suite. Three vary prompt-form and chat-template - the fourth (architecture axis) is measured by reconstructing two trajectory objects within each capture.

Table 1. Four-condition protocol: each capture combines a model checkpoint (PT or IT), a prompt form (completion seed or question form), and an optional chat template. Pairwise contrasts isolate post-training (PT vs IT-completion), prompt-form (IT-completion vs IT-question-no-template), and template-token (IT-question-no-template vs IT-chat) effects.

#	Condition	Model	Prompt	Chat template?
1	PT	pretrained base	completion seed	no
2	IT-completion	instruction-tuned	completion seed (same as PT)	**bypassed**
3	IT-question-no-template	instruction-tuned	question form	**bypassed**
4	IT-chat	instruction-tuned	question form (same as #3)	applied

Each pairwise comparison isolates one effect. **PT vs IT-completion** toggles the alignment-training axis with scaffolding off. **IT-completion vs IT-question-no-template** toggles the prompt-form sub-effect (completion seed → question form, no template either side). **IT-question-no-template vs IT-chat** toggles the template-token sub-effect (role markers, turn delimiters, BOS additions). The aggregate chat-mode scaffolding axis (IT-completion vs IT-chat) is the product of these two sub-effects within bootstrap precision (Appendix D.4) and matches a three-condition decomposition (PT / IT-completion / IT-chat) - the four-condition decomposition refines what the aggregate is composed of (section 4.5).

Inside each capture we reconstruct two trajectory objects. x is the model's actual residual stream - the running sum after each sublayer's normalised contribution. $\hat{\gamma}$ is the raw cumulative sum of sublayer outputs, as if architectural normalisation had not been applied. In pre-norm-only architectures the two are identical by construction - in post-sublayer-normalisation ("Peri-LN") architectures they diverge, with direction varying by family (section 4.3). The x-vs-$\hat{\gamma}$ comparison isolates the architectural axis.

1.3 Canvas - Pandey/Halawi misleading-assertion sycophancy

The Pandey/Halawi misleading-assertion canvas was sketched in section 1.1. Pandey [2] characterises it as eliciting a late-layer attention-head circuit that overrides the mid-stack-correct answer with the third-sentence-named entity, and reports that alignment masks rather than removes the circuit (RLHF refresh on Llama-3.1→3.3-70B cuts sycophancy 10× while shared heads persist); Genadi, et al. [4] locate sycophancy at mid-layer attention heads (L10-15) in Gemma-3-4B; Halawi, et al. [3] earlier identified the broader "false induction head" phenomenon underlying it. The canvas suits our decomposition because the question of *which* axis (architecture, alignment, or chat-mode scaffolding) drives the override circuit's strength is exactly what the protocol resolves.

The cell suite per family per condition is **432 cells**: 192 fragility (M) cells presenting the misleading assertion, organised as a fully crossed 4 framings × 2 positions × 3 value-gap magnitudes × 8 name pairs design; 192 minimal-pair non-misleading (NM) cells constructed by a single-word swap (`more` → `fewer`) of each M cell's assertion sentence, preserving every other token and enabling paired-prompt trajectory analysis (section 4.6); and 48 capability controls that contain just the comparison and the question with no third-sentence assertion, used to verify that the bare task is solvable. The suite is locked across families to keep cross-family comparison clean.

The 1.0B–1.7B scale of the five families is appropriate for this canvas. Pandey's Appendix L reports the override-circuit signature is concentrated at small scale (peak mid-layer DIFF excess +127% on Gemma-2-2B-IT, dropping to 0% by Llama-3.1-70B); per-cell trajectory signatures smear at larger scale (the smearing here is in the per-cell mechanism signature - the protocol's measurement instrument — not in cross-family scaffolding variance per se; Yun, et al.'s 7-8B template-on/off ablation (section 2.1) finds family-variable scaffolding effects at scales where the override circuit has substantially attenuated). A protocol that decomposes axis effects per cell with narrow bootstrap CIs needs the per-cell signature sharp at the scale of measurement (section 6.1).

1.4 Contributions

1. **Empirical: cross-family heterogeneity in (alignment, scaffolding) profiles across five ~1B-scale IT families.** Scaffolding-ratio spread 1.08× to 4.61× across four headline families (4.3× range); 0.69× to 4.61× (6.7× range) including Qwen3's default-thinking IT-chat regime, which is excluded from the headline aggregate as a measurement-artefact case (section 4.2, Appendix D.3). Scaffolding direction empirically trichotomous on this canvas; three of five alignment ratios near 2× and two near 1× - descriptive only, not a structural-cluster claim at $n = 5$. Findings contradict the pre-registered prediction on this canvas that scaffolding is the more recipe-consistent axis. The scaffolding axis decomposes further (section 4.5 / Appendix D.4): the template-token sub-effect imposes a content-independent positional rotation toward the subject of the third-sentence assertion across four families on this canvas (NM swaps only the comparator word and preserves the assertion's subject; correctness drops at the template step on NM cells in the same direction as on M cells; Qwen3 reverses direction under thinking suppression - Appendix D.3); on this canvas the assertion-subject is not the data-correct answer, so the rotation is uniformly detrimental to capture-rate correctness, while the aggregate scaffolding axis's family-variable direction lives in the prompt-form sub-effect.
2. **Empirical: deployment-conditional alignment (conditional on cross-canvas generalisation).** Two alignment-interaction patterns plus one measurement-artefact case: (a) Gemma - PT-to-IT-completion installs resistance, chat-mode scaffolding wipes it out; (b) SmolLM2 - PT-to-IT-completion installs no resistance, only chat-mode scaffolding does; (c) Qwen3 - fixed-position answer extraction is blind to thinking-mode routing, so chat-mode application appears to make the IT model fail capability tasks PT solves (section 4.2, Appendix D.3) - a measurement artefact surfaced by the protocol's controls, and the reason Qwen3 is excluded from the Phase 7 headline cross-family scaffolding aggregate. Single-canvas, chat-only evaluations resolve none of these cleanly. The framing-shift hypothesis these license - alignment as a deployment-conditioned phenomenon rather than a weight-space intervention - is developed in section 5.6.
3. **Empirical: NM-flip mechanism is family-specific.** The four-condition NM-correctness decomposition adjudicates the position-bias account of chat-mode scaffolding: confirmed in Gemma (template-tokens drive the flip), mixed in Llama, contradicted in OLMo and Qwen3 (the flip happens at the prompt-form step, before any template), not applicable in SmolLM2. Single-family mechanism claims about "what chat scaffolding does" should be reported as family-specific until cross-family validation is run.
4. **Methodological: the protocol + metric apparatus that produced the findings above.** A specification for decomposing architecture × alignment × scaffolding effects with narrow cell-level bootstrap CIs and per-token-per-layer divergence maps. Alignment and scaffolding ratios use standard logit-lens margins; the architectural-axis A' statistic and minimal-pair $\cos_G$ divergence use Curved Inference's pullback metric ($G = U^T U$) where geometric measurements require a metric on the residual stream (section 3.2). The protocol also supplies a diagnostic vocabulary for classifying features observed in an IT-chat model (post-training-installed vs template-token-elicited vs prompt-form-elicited vs interaction effect; plus a measurement-position-artefact category to rule out) - see section 5.6.
5. **Open data and full methodological record.** HDF5/JSONL capture set, analysis pipeline, and the locked pre-registration document (with amendments) released alongside the paper.

2. Related Work

2.1 Prior partial decompositions (closest intellectual neighbours)

The closest existing work to the protocol developed here is Minder, et al. [5], who train crosscoders between base and chat-tuned model pairs and observe that approximately 40% of the latents activating exclusively on chat-tuned activations fire on chat-template structural tokens. This directly suggests that a substantial fraction of what the field attributes to alignment training may live in the chat-mode scaffolding. Minder, et al. establish the observation but do not run a no-template ablation on the chat model, leaving the alignment-versus-scaffolding attribution unresolved. The protocol we develop provides that ablation.

Li, M.Z., et al. [6] geometrically decompose pretraining, SFT, DPO, and RLVR stages on a single base family; their weight-axis training-stage decomposition is the closest prior art to our inference-time input-axis decomposition across five families, and has no template-axis ablation. Kissane et al. [7] report SAE-feature transferability between base and chat-tuned models with template-axis awareness but no within-model template-off control. Zhao et al. [8] distinguish template-position from content-position encoding effects; the "Curious bos_token" LessWrong report [9] raised the

question of how much instruction-tuned behaviour is template-elicited that our protocol answers directly for the misleading-assertion canvas. Minder, et al. [10] trace narrow-finetuning effects through residual-stream activation differences. Each intersects the chat-mode scaffolding axis without operationalising it as a per-axis ablation.

Yun, et al.'s "Price of Format" [11] is the closest published precedent for the scaffolding-axis ablation. Across five 7-8B IT models on nine generation tasks they run a within-model template-on/off ablation across four prompting modes that maps onto our IT-chat / IT-question-no-template / IT-completion split, finding a universal output-distribution "diversity collapse" under templates. Align-IT extends this on three key contrasts: a PT-vs-IT weight axis factoring scaffolding from alignment per family; an adversarial signed-margin canvas (misleading-assertion capture rate) rather than open-ended diversity; and cross-family heterogeneity in scaffolding magnitude (CV 0.74 across the four headline families; 0.82 including Qwen3's default-thinking IT-chat regime) where Yun, et al. report cross-task universality of the diversity direction. They establish that scaffolding affects behaviour with replicable direction; we establish that scaffolding-vs-alignment factorisation is family-variable in magnitude.

Dumas, Minder & Nanda's "What We Learned Trying to Diff Base and Chat Models" [12] is the most directly adjacent contemporary effort. They use sparse-dictionary diffing on Gemma-2-2B at layer 13, evaluating behavioural relevance via KL-divergence recovery, and identify genuine chat-specific concepts including refusal triggers and template-token-relevant latents. The methodologies are complementary: their sparse-dictionary diffing identifies chat-specific concepts; our input-regime ablation across five families decomposes the scaffolding axis into prompt-form and template-token sub-effects with per-cell signed-margin readouts. Theirs does not run the no-template ablation; ours does not identify the concept-level latents. Julian Minder's shared authorship across Minder, et al. [5] and this post indicates a common evolving research line.

2.2 The Pandey/Halawi sycophancy mechanism (canvas)

The canvas to which we apply the protocol is the Pandey/Halawi misleading-assertion stress test introduced in section 1.3. Halawi, et al. [3] identified the broader "false induction head" phenomenon (late-layer heads that attend to and copy false in-context labels) with their "critical layer" finding (mid-layer decoding can outperform the full model on incorrect demonstrations) establishing the early-correct/late-overridden trajectory pattern Pandey's sycophancy circuit instantiates in the single-assertion setting. Pandey [2] tightens this with sycophancy-specific characterisation: the relevant heads share write-directions with typed-output subspaces, and alignment masks rather than removes the circuit (sycophancy reduced 10× behaviourally on Llama-3.1→3.3-70B refresh while shared heads persist). Genadi, et al. [4] is a parallel finding at Gemma-3-4B locating sycophancy at mid-layer attention heads via linear probes (residual peak L15, MLP peak L10). Sharma, et al. [13] provide broader behavioural and RLHF context; Wang, et al. [14] ("When Truth Is Overridden", 2025) and Shapira, et al. [15] ("How RLHF Amplifies Sycophancy", 2026) are concurrent and recent work on related phenomena.

We use this canvas because it offers a mechanistically characterised target (the override-circuit's per-family strength is exactly what the protocol's axes decompose) and operates at the small scale where its trajectory signature is sharpest (Pandey Appendix L; section 1.3; section 6.1). The Pandey/Halawi mechanism maps cleanly onto the States, Routes, and Anchors vocabulary of the 3-Process View [16]: the misleading assertion acts as an early-token Anchor, the mid-stack-correct intermediate is a State, and the late-layer override is a Route.

2.3 Attention-head functional roles (broader context)

The Pandey/Halawi sycophancy circuit sits within a broader literature of named attention-head functional roles: copy-suppression [17] (the override pattern Pandey's heads exhibit), induction heads [18], name-movers in IOI circuits [19], inference-time intervention [20] (the head-level methodology Pandey extends), refusal-direction work [21], safety heads [22], and retrieval heads [23]. We argue the protocol should be applied per family before drawing functional-role conclusions in any of these strands: single-family head identification under chat-only evaluation may attribute to a head's mechanism what is partly elicited by the chat-mode scaffolding, and the three-direction trichotomy on scaffolding effect we observe in section 4.5 implies the same head's measured effect can differ qualitatively across families.

2.4 Architectural normalisation and the residual stream

The architectural axis draws on a literature characterising the residual stream and its normalisation as a methodological variable, without previously pairing x with $\hat{\gamma}$ as separable measurement objects. Kim et al. [24] introduce "Peri-LN" as a class label for the post-sublayer normalisation pattern distinguishing Gemma-3 and OLMo-2 from pre-norm-only

designs. The Gemma-3 technical report [25] documents post-norm + pre-norm with RMSNorm and QK-norm; the $(1+\gamma)$ formulation - pretrained γ shifting the residual-stream contribution multiplicatively from the initialisation baseline - is in the reference implementation (HuggingFace `transformers/models/gemma3/modeling_gemma3.py`). OLMo-2 uses standard RMSNorm in the same Peri-LN placement [26], producing the opposite-direction signature in section 4.3. Baroni, et al. [27] argue LN's nonlinearity is itself an interpretability confound for pre-norm decoders. Elhage, et al. [28] supplies the additive baseline $x_{\ell+1} = x_\ell + \Delta_\ell$ that $\hat{\gamma}$ equals absent post-sublayer normalisation. Heimersheim & Turner [29] observe residual-stream norms growing exponentially over the forward pass in GPT-2-family pre-norm decoders.

The x-versus-$\hat{\gamma}$ separation as a methodological choice does not appear in this strand. Our two Peri-LN families show opposite-direction A' signatures (Gemma's $(1+\gamma)$ parameterisation amplifies the integral; OLMo's standard RMSNorm reduces it; section 4.3); with $n = 2$ we describe what we observed rather than claim a structural sub-regime, but the class label clearly under-specifies the architecturally relevant variable.

2.5 Geometric Interpretability and Curved Inference (metric apparatus for two of the protocol's measurements)

Curved Inference [1] establishes the residual-stream's geometric primitives (curvature κ, salience S, semantic surface area A') evaluated under the pullback metric $G = U^T U$ induced from the model's unembedding direction. The companion 3-Process View piece [12] supplies the States/Routes/Anchors vocabulary for relating circuit-level events to trajectory-level geometry.

This paper uses CI's metric apparatus for exactly two of its measurements: the A'-integrated trajectory ratio supplying the architectural-axis statistic (section 4.3) and the $\cos_G(x_M, x_{\mathrm{NM}})$ divergence supplying the minimal-pair mechanism analysis (section 4.6). The alignment-axis and chat-mode scaffolding-axis ratios in Tables 3, 7, 8 are computed from standard logit-lens margins (section 3.2) and do not require CI. The relationship to CI is therefore methodological extension on two specific axes, not a paper-wide framework dependency, and the headline cross-family heterogeneity finding does not require a reader to commit to CI's framework before evaluating the empirical claims.

2.6 Methodology critiques in 2025–2026 mechanistic interpretability

Several named methodology confounds in the recent interpretability literature bracket the contribution of this paper. Sharkey, et al. [30] list chat-tuning data dependence as a passing concern within a broader survey. Bai, et al.'s "Story Is Not the Science" reproducibility critique [31] and Song, et al.'s SAE feature-consistency analysis [32] flag concerns about whether mechanistic findings replicate across implementation choices. Li, A.J., et al. [33] argue that SAE concept representations are not adversarially robust enough for reliable monitoring use. The axis-decomposition confound we develop here is distinct from each: not a duplicate critique of reproducibility, feature consistency, or identifiability, but a separate conflation that the chat-template-axis ablation resolves. We position it as a fourth named methodological refinement complementary to the existing three.

2.7 Methodological gap

Across the refusal-direction, sycophancy-circuit, and SAE-diff strands surveyed in section 2.1–section 2.6, no published study we identified runs an IT-completion condition as a within-model control on its chat-mode results. Architecture-aware methodology is established in the architecture-design strand (Peri-LN, normalisation placement) but absent from the interpretability-results strand that operates on those architectures. Awareness of chat-mode scaffolding exists in dispersed observations (Minder, et al.; the "Curious bos_token" report) and is operationalised behaviourally on open-ended generation by Yun, et al.'s "Price of Format", but has not been operationalised as a per-axis ablation on an adversarial canvas with weight-axis decomposition across families. The protocol applied across five model families fills this gap.

3. Methods

The protocol, its metric apparatus, the minimal-pair extension, the five-family / 432-cell suite, and the bootstrap-CI statistical procedure are described compactly below. Per-script procedural detail is in the released code; the locked methodological record is in Appendix A.

3.1 The four-condition protocol and the axis decomposition it supports

The protocol decomposes interpretability measurements along three lifecycle axes. **Architecture** is set during pretraining: norm placement and parameterisation determining how each sublayer's contribution enters the residual stream. **Post-training weight changes** (the *alignment axis* when discussing the intended recipe role) are the aggregate effect of all weight modifications between PT and IT checkpoints (SFT, DPO, RLHF, RLVR, GRPO, distillation, data and hyperparameter choices); we reserve "alignment training installs X" for claims tying X to a specific stage rather than to the aggregate PT-vs-IT-completion difference. The preferred alternative phrasing throughout is "PT-vs-IT-completion shows increased resistance to X" (or analogous), which keeps the contrast explicit and stage-agnostic. **Chat-mode scaffolding** is applied at inference and decomposes into a prompt-form shift (completion seed → question form) and template-token application (role markers, turn delimiters, BOS additions). Each axis is isolated by varying one while holding the others fixed.

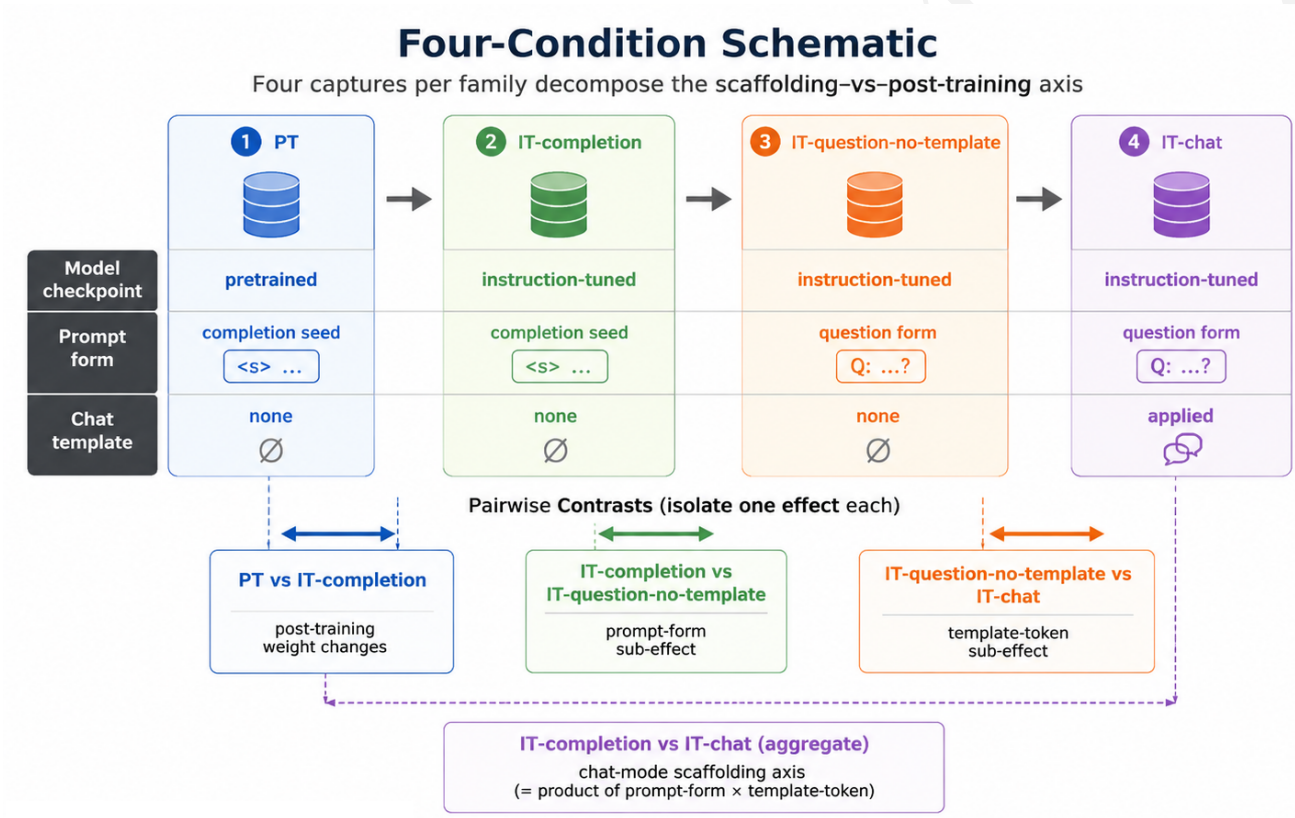

Figure 1: Four-condition schematic. Each capture varies one or both of: model checkpoint (PT vs IT) and inference-time scaffolding (completion-seed vs question-form vs chat-template). Pairwise contrasts isolate post-training weight changes (PT vs IT-completion), prompt-form sub-effect (IT-completion vs IT-question-no-template), and template-token sub-effect (IT-question-no-template vs IT-chat). The aggregate chat-mode scaffolding axis (IT-completion vs IT-chat) is the product of the two sub-effects.

Four captures per family supply the data (Table 1; Figure 1; the prompt-form vs template-token decomposition is developed in section 4.5 and Appendix D.4). Within each capture, two trajectory objects are reconstructed: x is the actual residual stream (running sum after each sublayer's normalised contribution), and $\hat{\gamma}$ is the raw cumulative sum of sublayer outputs as if no post-sublayer normalisation had been applied. The intra-capture x **vs** $\hat{\gamma}$ contrast isolates architectural normalisation: in pre-norm-only architectures the two are identical by construction; in Peri-LN architectures they diverge, with direction depending on the norm parameterisation (section 4.3, Table 5). The **PT vs IT-chat** contrast is the combined post-training + scaffolding stack the field's standard methodology reports as a single difference.

Captures use `mot-capture.py`, which installs per-layer `forward_pre_hooks` to read x directly and reconstructs $\hat{\gamma}$ in parallel from each sublayer's pre-normalisation output. Final-layer x margins reconcile with the forward-pass logits to ± 0.001. A tokenisation-alignment gate (`verify-tokenization.py`) runs ahead of every capture so the

logsumexp aggregation over candidate-token forms is well-defined per cell. The full pipeline is in `run.sh`.

3.2 Metrics - standard logit-lens plus two CI-supplied extensions

The protocol uses three measurements. The answer-start-position margin (post-training and chat-mode scaffolding ratios) is standard logit-lens with no framework dependency. The A' trajectory measure (architectural axis) and the $\cos_G$ divergence (minimal-pair analysis) come from the Curved Inference pullback-metric framework. Headline results in section 4.1 need only the standard class; CI's contribution lives in section 4.3 and section 4.6.

Standard logit-lens margin (no framework dependency). For each cell we compute the log-preference between correct and misleading candidates at the answer-start position by passing the residual stream through the model's final-layer normalisation and unembedding matrix:

$\alpha_{\text{correct}}(\ell)$ = logsumexp over correct-token-forms of lm_head(final_norm($x[\ell, t_{\text{answer}}]$)) $\alpha_{\text{misleading}}(\ell)$ = logsumexp over misleading-token-forms of lm_head(final_norm($x[\ell, t_{\text{answer}}]$)) $\text{margin}(\ell) = \alpha_{\text{correct}}(\ell) - \alpha_{\text{misleading}}(\ell)$

The logsumexp aggregation handles multiple tokenisation forms (e.g., `Bob` with leading space and `Bob` without). Post-training and chat-mode scaffolding ratios in Tables 3, 7, 8 are per-family magnitude ratios over these margins. The three metric classes the paper uses (capture rate, signed margin, absolute-margin ratio) and what each is sensitive to are summarised in the "Reading the numbers" box before Table 3 in section 4.1.

Curved Inference pullback metric $G = U^T U$ establishes a coordinate system for the residual stream in which distances and angles correspond to *output-relevant* differences (G flattens the residual stream into the subspace used for vocabulary predictions). Two measurements live in this coordinate system:

- **Intrinsic geometry under** G (curvature κ, salience S, semantic surface area A') measures the *shape* of a single trajectory. The per-cell $A'(x)/A'(\hat{\gamma})$ ratio is the architectural-axis statistic (section 4.3, Table 5); without a metric on the residual stream the ratio would be coordinate-arbitrary.
- **Trajectory-vs-trajectory divergence under** G ($\cos_G(x_a, x_b)$) measures how two trajectories diverge. The minimal-pair analysis (section 4.6) reports $1 - \cos_G(x_M[\ell, t], x_{\text{NM}}[\ell, t])$ per (layer, token) per cell pair, and likewise requires a metric on the residual stream.

The CI guide [1] develops the κ / S / A' / $\cos_G$ framework; we apply it across multi-condition cross-family captures and add the per-cell $A'(x)/A'(\hat{\gamma})$ ratio as a methodological refinement (the x-vs-$\hat{\gamma}$ distinction does not appear in published CI or the broader architectural-normalisation strand, section 2.4). Readers unfamiliar with G can follow every section from the margin definition above except section 4.3, section 4.6, and the mechanism account in section 5.4.

3.3 Minimal-pair (M, NM) extension

For each of the 192 M cells we construct a paired NM cell by a single-token swap of the comparator word in the third-sentence assertion (`more` → `fewer`), preserving every other token. M and NM prompts are token-identical at all pre-swap positions, so $x_M[\ell, t] = x_{\text{NM}}[\ell, t]$ for every layer ℓ and pre-swap token t. This structural equality is a property of the design, not the measurement, and guarantees pre-swap $\cos_G(x_M, x_{\text{NM}}) = 1$ and pre-swap divergence $(1 - \cos_G) = 0$ by construction.

The construction enables a CI01-native trajectory-vs-trajectory analysis: for each (M, NM) pair we compute $\cos_G(x_M[t, \ell], x_{\text{NM}}[t, \ell])$ at every (token, layer) and report post-swap divergence localisation, capture-status correlation, and per-family chat-template amplification (section 4.6). The single-token design also keeps `more` / `fewer` (leading-space form) verified single-token across all five families' tokenisers (Appendix C), avoiding multi-token alignment-aggregation cost and keeping the post-swap measurement directly comparable across families.

3.4 Models and cell suite

Five open-weight families at ~1B scale span the architectural and alignment-recipe variation accessible at this scale.

Table 2. Five-family architectural and alignment-recipe overview. Sizes 1.0B-1.7B; norm placement / attention design / QK-norm presence / alignment-recipe label per family. Full per-family detail in Appendix B.

Family	Size	Norm scheme	Attention	QK-norm	Alignment recipe
Gemma-3-1b	1.0B	Peri-LN ($(1+\gamma)$)	GQA (4Q/1KV)	yes	SFT + RLHF + distillation
Llama-3.2-1B	1.2B	pre-norm only	GQA (32Q/8KV)	no	SFT + DPO
OLMo-2-1B	1.5B	Peri-LN (std RMSNorm)	MHA (16)	yes	SFT → DPO → RLVR
Qwen3-1.7B	1.7B	pre-norm only	GQA (16Q/8KV)	yes	SFT + DPO + GRPO
SmolLM2-1.7B	1.7B	pre-norm only	MHA (32)	no	SFT + DPO (Zephyr)

Cell suite (per family per condition): 192 M + 192 NM + 48 capability controls = 432 cells. M cells are fully crossed over 4 framings × 2 positions × 3 value-gaps × 8 name pairs (Appendix B).

3.5 Statistical procedure

For each cell we compute per-condition margin ($\alpha_{\text{correct}} - \alpha_{\text{misleading}}$) at the answer-start position via logsumexp aggregation over candidate-token forms. The margin is the standard logit-lens readout defined in section 3.2 (final_norm + lm_head; no G dependency); G enters only in section 3.2's two CI-supplied extensions (A' and $\cos_G$), not in the margin itself or in the post-training / scaffolding / combined ratios derived from it.

Aggregate-ratio definition. Per-cell ratios (post-training, scaffolding, and combined) derive from the per-cell margins. The aggregate post-training, scaffolding, and combined ratios reported in Tables 3, 7, 8 are *ratios of aggregate mean absolute margins across the 192 M cells per family per condition* (numerator = mean |margin| across M cells in the numerator-condition; denominator = mean |margin| across the same M cells in the denominator-condition), **not** the mean of per-cell ratios. The aggregate-mean-absolute formulation handles near-zero-denominator cells implicitly: small-margin cells contribute small magnitudes to both numerator and denominator and do not blow up the aggregate ratio (a per-cell-ratio formulation would). The per-cell-ratio distribution is reported separately as the SD-log-ratio diagnostic (D1; section 4.7).

Capture rate provenance. Capture rate (Tables 3, 7, 8, section 4.2) is the proportion of M cells where the signed margin at the answer-start position is negative (model picks the misleading entity). It is computed from the same per-cell margins as the ratios above, not from generated tokens - no sampling or generation is involved.

Aggregate ratios are bootstrap point estimates with 10,000-sample percentile 95% CIs at the cell level - per-group breakdowns (framing, position, value-gap label, name pair) use the same bootstrap within each group.

The cell-level bootstrap treats the 192 M cells as independent draws. Because the prompt suite is fully crossed, cells sharing a clustering variable are not strictly independent - as a sensitivity check we also run a clustered bootstrap over whole name-pair clusters (8 × 24) or framing clusters (4 × 48). Point estimates are identical - clustered CIs widen when within-cluster correlation is non-trivial. Table 3 and per-family discussions report cell-level CIs as the default, with the side-by-side comparison in Appendix D.1 and bootstrap-sensitive results flagged in section 4.1.

For multiplicative-model fit we compute the standard deviation of log-ratios across cells (tighter than ratio-of-means). For minimal-pair (M, NM) trajectory divergence we use $1 - \cos_G(x_M, x_{\text{NM}})$ per (token, layer) - this is the one place G enters the statistical procedure. Aggregate statistics include the at-swap value, post-swap-far value, and per-cell peak-vs-cross-layer concentration. Captures are deterministic (`do_sample: false`, `runs_per_prompt: 1`, no dropout in eval mode), so residual streams and downstream metrics are fully determined by the checkpoint and prompt strings. Bootstrap CIs use a fixed RNG seed (`numpy.random.default_rng(20260507)`) so released aggregate ratios are bit-reproducible (Appendix E.5).

The protocol, hypotheses, and statistical procedure were committed before any data capture began, and the locked pre-registration record with subsequent amendments is included as Appendix A.

4. Results

4.1 Cross-family decomposition - the headline empirical finding

The protocol decomposes a model's preference for the misleading answer into three multiplicative contributions: an architectural baseline (PT), the post-training weight change (PT → IT-completion), and the chat-mode scaffolding effect (IT-completion → IT-chat). Per-family ratios are tight under cell-level bootstrap (10,000 samples, 95% percentile CIs) and the per-condition capture rates make the story directly legible.

Reading the numbers (metric classes). Three metric classes appear below and differ in what they are sensitive to:

Quantity	Sign?	Answers
Capture rate	yes	Did the model pick the misleading entity? (per cell, aggregated as proportion)
Signed margin	yes	Which answer is preferred and by how much? (per cell, per layer)
Absolute-margin ratio	no	How much did the preference *magnitude* change, regardless of direction?

A larger absolute-margin ratio does not mean "better" or "worse" - only "larger swing in preference strength". SmolLM2 is the clearest example: scaffolding ratio 1.17× (modest magnitude) corresponds to capture-rate dropping from 100% to 60.9% (substantial behavioural shift) because cells sit near the decision boundary. Per-family direction (detrimental / neutral / beneficial) is read from capture rates and signed margins, not from absolute-margin ratios.

Table 3. Per-family capture rates (proportion of 192 M cells where the model picks the misleading-named entity) and the multiplicative decomposition of margin magnitude across the same cells. Capture rates are derivable as `(192 - correct) / 192` from the M-correctness counts in section 4.5 - ratios are bootstrap point estimates with cell-level 95% CIs. Combined ratio = post-training × scaffolding within bootstrap precision. The 4.3× scaffolding spread across the four headline families (6.7× including Qwen3's default-thinking IT-chat regime - † - excluded from the headline aggregate due to a thinking-mode-routing measurement artefact: Appendix D.3 and the Qwen3 sensitivity note below) and the strong-vs-near-1× post-training partition are robust under both cell-level and clustered bootstraps - individual separations of Llama's and Qwen3's scaffolding ratios from 1.0 are sensitive to the clustering choice (Appendix D.1: Llama scaffolding widens to [1.01, 3.63] and Qwen3 to [0.26, 1.32] under name-pair clustering, while Gemma and OLMo are stable across schemes).

Family	PT	IT-comp	IT-q-n-t	IT-chat	Post-train ratio	Scaffolding ratio	Combined
Gemma-3-1b	78.1%	51.6%	48.4%	99.0%	2.03× [1.82, 2.27]	**4.61× [4.24, 5.04]**	9.35× [8.61, 10.22]
Llama-3.2-1B	49.5%	36.5%	47.9%	75.0%	1.87× [1.60, 2.20]	2.05× [1.72, 2.45]	3.85× [3.22, 4.58]
OLMo-2-1B	82.3%	82.8%	56.8%	87.5%	1.94× [1.75, 2.16]	1.08× [0.95, 1.23]	2.09× [1.83, 2.40]
Qwen3-1.7B †	91.1%	77.6%	65.1%	77.6%	0.99× [0.89, 1.08]	**0.69× [0.57, 0.83]**	0.68× [0.58, 0.80]
SmolLM2-1.7B	100%	100%	44.8%	60.9%	0.95× [0.89, 1.00]	1.17× [1.05, 1.31]	1.11× [0.99, 1.23]

Qwen3 sensitivity (no-think mode) - † Qwen3 excluded from the headline cross-family aggregate. Qwen3 IT-chat is reported under its default thinking-on chat mode. The chat template emits `<think>` at the answer-start position, so fixed-position margin extraction reads the wrong location for that family. The no-think survival check (Appendix D.3) confirms: with thinking suppressed (`enable_thinking=False`), capability correctness recovers from 6/48 to 27/48 and the mean signed margin flips from −2.24 to +4.51. The protocol exposes the routing artefact rather than obscuring it - absent the IT-completion control, this would be misread as an alignment effect. For this reason Qwen3's default-thinking IT-chat regime is excluded from the headline four-family cross-family scaffolding aggregate (CV 0.74, range 1.08×–4.61×) - including it shifts the aggregate to five-family CV 0.82, range 0.69×–4.61× (post-training CVs 0.30 → 0.35). The qualitative claim (chat-mode scaffolding is more recipe-variable than post-training) is robust to this choice. Qwen3 retains its role in the per-family case studies (section 4.2, section 5) and in Appendix D as a "protocol exposes a routing artefact" study.

Figure 2 shows the cross-family axis signature on the misleading-assertion canvas. This heatmap-style summary of each family's position on the three decomposition axes (architecture, post-training, chat-mode scaffolding). The architecture column shows the A'-integrated divergence between x and $\hat{\gamma}$ in the pretrained model ($|\hat{\gamma}/x - 1|$; 0 = no architectural-norm effect at runtime; >1 indicates substantial divergence between the actual residual stream and its raw-cumulative-deltas reconstruction). The post-training and scaffolding columns show the per-axis ratios from Table 3 (deviation from 1.0 indicates effect magnitude). Qwen3 is marked with a dagger (†) to flag that its scaffolding value is the default-thinking IT-chat regime, excluded from the four-family headline aggregate due to a thinking-mode-routing measurement artefact (Appendix D.3). Cell colour encodes effect strength - numeric values are inset (white on darker cells, black on lighter cells, for legibility). The figure makes cross-family heterogeneity visible at a glance: same canvas, same protocol, five qualitatively different per-axis signatures. (Rendered by `bin/plot-figure-2.py` from `results/cross-family-rollup.json` - raw data in Tables 3 and 5.)

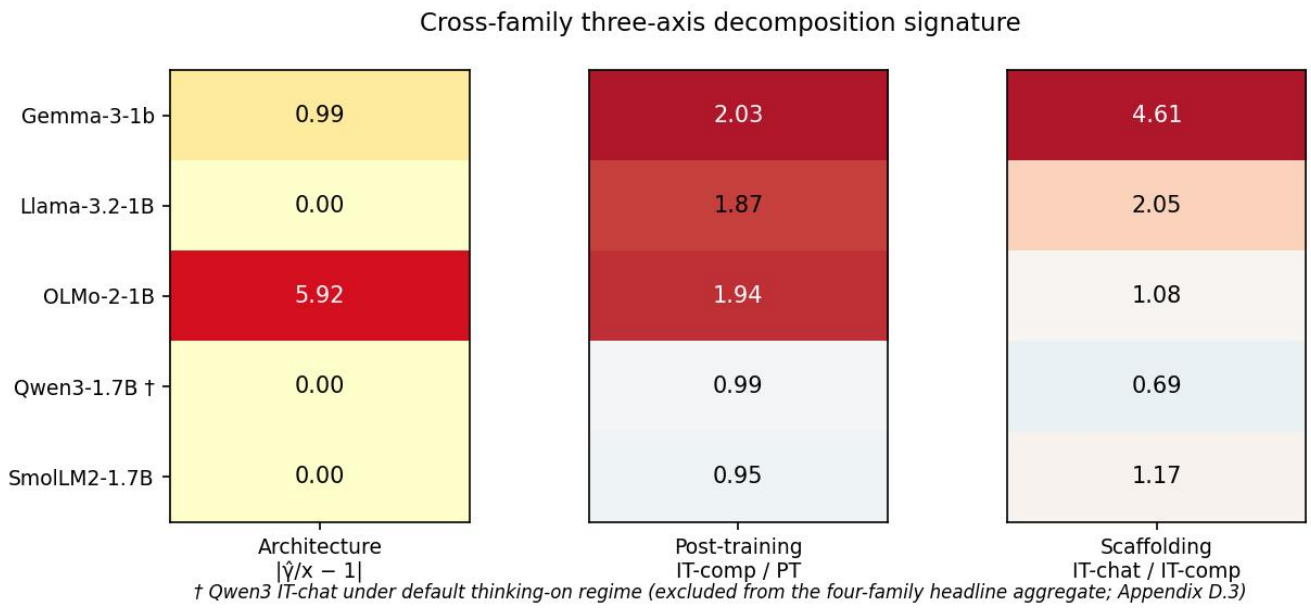

Figure 2: Cross-family axis signature.

Scaffolding ratios range from 1.08× to 4.61× across the four headline families - a 4.3× spread (6.7× including Qwen3's default-thinking IT-chat - see Qwen3 sensitivity note above and Appendix D.3). Under cell-level bootstrap three families have mutually non-overlapping intervals (Gemma [4.24, 5.04], Llama [1.72, 2.45], Qwen3 [0.57, 0.83]). Under name-pair-clustered bootstrap (Appendix D.1), Gemma [3.99, 5.19] and Qwen3 [0.26, 1.32] remain clearly separated while Llama widens to [1.01, 3.63] and overlaps Qwen3 - the qualitative claim (three distinct empirical effects of the same chat-template step on identical prompts) is preserved across the four headline families regardless of where the Qwen3 number is read.

Post-training ratios show that recipe-name labels do not predict magnitude. Llama-3.2-1B and SmolLM2-1.7B both list "SFT + DPO"; Llama's ratio is 1.87×, SmolLM2's is 0.95× - two-fold different effect under the same tag. The tag is a thin description of two models that differ in pretraining-data composition, base architecture, DPO data, and regularisation, so the negative claim is what the data supports cleanly. Descriptively, three of five ratios sit near 2× and two near 1× - with five points we describe, we do not infer cluster structure.

The architectural axis adds a third layer. In pre-norm-only architectures (Llama, Qwen3, SmolLM2) $\hat{\gamma}/x = 1.0$ by construction. In Peri-LN architectures direction depends on parameterisation: Gemma's $(1+\gamma)$ gives $A'(x) \approx 90 \times A'(\hat{\gamma})$ (amplification) - OLMo's standard-RMSNorm gives $A'(x) \approx A'(\hat{\gamma})/7$ (reduction). Same nominal class, opposite-direction signatures (section 4.3).

The operational assumption underlying chat-only IT evaluation (templates are a shared inference-time convention while alignment recipes vary substantially) would predict the scaffolding axis to be more recipe-consistent than post-training. The assumption is implicit, carried by dominant chat-only methodologies in the refusal-direction, sycophancy-circuit, and SAE-diff strands (section 2). The pre-registered D2 hypothesis (Appendix A) operationalises it. The data refutes it:

- Chat-template ratio coefficient of variation: **0.74** across the four headline families (1.08× to 4.61×); **0.82** including Qwen3's default-thinking IT-chat (range extends to 0.69×).
- Post-training ratio coefficient of variation: **0.30** across the four headline families; **0.35** including Qwen3 (three near 2×, two near 1×).

The chat-mode scaffolding axis is *more* family-variable than the post-training axis, not less. This is the substantive empirical content of section 5.2. The remainder of section 4 develops per-family stories (section 4.2), each axis in turn (section 4.3–4.5), the (M, NM) trajectory analysis (section 4.6), and per-cell interactions the aggregates flatten (section 4.7).

4.2 Per-family case studies - five distinct mechanism stories

Each family combines post-training-weight-change effect and chat-mode-scaffolding effect in a qualitatively different way.

Reading discipline. "PT-to-IT-completion installs X" is shorthand for "the aggregate weight change between PT and IT checkpoints produces X under the completion-seed regime" (section 3.1). X is not attributable from this data to any specific recipe stage - the contrast measures the aggregate. Gemma's M-cell-resistance gain is contemporaneous with a general capability gain (Appendix D.5: PT 24/48 → IT-completion 44/48), so the framing should be read as aggregate weight change producing both, not as recipe-specific resistance installation.

Table 4. Per-family one-phrase mechanism summary across the post-training and chat-mode scaffolding axes - full per-family discussion follows in section 4.2.

Family	One-phrase mechanism story
Gemma-3-1b	PT-to-IT-completion installs resistance; chat-template tokens collapse it back
Llama-3.2-1B	Same direction as Gemma at smaller magnitude across the board
OLMo-2-1B	PT-to-IT-completion has substantial M-cell effect; chat-mode scaffolding is near-neutral aggregate (two opposing sub-effects, section 4.5)
Qwen3-1.7B	Both axes near-null aggregate on M cells; thinking-mode routing artefact at capability cells (D.3, D.5)
SmolLM2-1.7B	PT-to-IT-completion installs nothing; chat-mode scaffolding supplies the deployed resistance - prompt-form-driven (D.4)

Gemma-3-1b is the family where post-training installs meaningful M-cell resistance and chat-mode scaffolding overrides it. Capture rate 78.1% → 51.6% (PT-to-IT-completion, a 26-point improvement) → 99.0% (chat-template), recapturing almost every cell the weight change had recovered. The 4.61× scaffolding ratio reflects domination, not subtle nudging. Gemma's $(1 + \gamma)$ Peri-LN parameterisation amplifies the scaffolding effect (section 4.3), though cross-family data show that direction is not determined by architecture alone. *Caveat (Appendix D.5):* bare-task capability correctness also rises sharply at PT-to-IT-completion (24/48 → 44/48), so the resistance gain is confounded with a general completion-form answer-extraction gain at that step.

Llama-3.2-1B shows the same qualitative pattern in milder form: capture 49.5% → 36.5% → 75.0% (post-training 1.87×, scaffolding 2.05×). Mid-range scaffolding places Llama between Gemma's strongly-detrimental effect and the families where it is null or beneficial.

OLMo-2-1B has substantial post-training effect but near-neutral aggregate scaffolding: post-training 1.94×, scaffolding 1.08× [0.95, 1.23] (interval includes 1.0). Capture rates are essentially flat (82.3% → 82.8% → 87.5%). The aggregate-neutral scaffolding hides two large opposing sub-effects (prompt-form 0.52×, template-token 2.09× - section 4.5).

Qwen3-1.7B has both axes statistically near or below 1.0× on M cells (post-training 0.99×, scaffolding 0.69× under cell-level bootstrap but not robust under name-pair clustering, Appendix D.1 - capture 91.1% → 77.6% → 77.6%). The striking observation is on capability controls: PT 46/48 correct (mean signed margin +3.07), IT-chat 6/48 (mean −2.24). The mechanism is the thinking-mode-routing artefact noted at Table 3: the no-think check (Appendix D.3) recovers 27/48 with margin +4.51. The accurate framing is "fixed-position extraction is blind to dynamic routing architectures", not "the IT model has lost the capability".

SmolLM2-1.7B is the mirror image of Gemma. PT-to-IT-completion installs no detectable resistance (both 100% captured) - only IT-chat (60.9%) is meaningfully lower. The chat template, not the weight change, installs the model's deployed resistance. Post-training ratio 0.95× confirms IT-completion margin is essentially unchanged. Scaffolding ratio 1.17× is modest on magnitude but corresponds to a 39-percentage-point capture-rate reduction, because the scaffolding effect on SmolLM2 flips cells across zero rather than amplifying strong preferences. Architecturally pre-norm only - recipe is Zephyr-style SFT + DPO. This is nominally the same as Llama, yet a qualitatively different profile.

The cross-cutting pattern is the absence of a unifying mechanism story. The Halawi + Pandey framing - "post-training installs resistance; chat-mode scaffolding amplifies the override" (Pandey reports sycophancy reduced 10× behaviourally on the Llama-3.1→3.3-70B RLHF refresh while shared heads persist across families) - fits Gemma

and Llama. It does not fit OLMo (opposing sub-effects cancel), Qwen3 (PT-to-IT-completion installs nothing - scaffolding actively reduces margin), or SmolLM2 (scaffolding installs the resistance - prompt-form-driven, section 4.5). The five families share the same three axes but combine them in qualitatively distinct ways - section 5 develops the implications.

4.3 The architecture axis

The architectural axis compares two residual-stream trajectories: x, the actual sequence of running sums after each sublayer's normalised contribution - and $\hat{\gamma}$, the raw cumulative sum of sublayer outputs absent post-sublayer normalisation scaling. In pre-norm-only architectures the two are identical by construction. In Peri-LN architectures they differ, with direction depending on parameterisation. We measure the divergence via the per-cell ratio of their semantic-surface-area integrals (A' from CI03, integrated over (token × layer)). Median ratios across the five families in PT are in Table 5.

Table 5. Per-family architectural-axis signature in the pretrained model. The $\hat{\gamma}/x$ ratio is the median across 432 cells per family - values further from 1.0 indicate larger divergence between x and $\hat{\gamma}$.

Family	Norm placement	$\hat{\gamma}/x$ ratio (PT)	$\|\hat{\gamma}/x - 1\|$	Empirical regime (A' measure)
Gemma-3-1b	Peri-LN	0.011	0.989	norm amplifies A' by ~90×
Llama-3.2-1B	pre-norm only	1.000	0.000	$\hat{\gamma} = x$ by construction
OLMo-2-1B	Peri-LN	6.95	5.92	norm reduces A' by ~7×
Qwen3-1.7B	pre-norm only	1.000	0.000	$\hat{\gamma} = x$ by construction
SmolLM2-1.7B	pre-norm only	1.000	0.000	$\hat{\gamma} = x$ by construction

Three empirical regimes emerge. **Norm amplifies** in Gemma: the $(1 + \gamma)$ parameterisation scales each sublayer's output by $1 + \gamma_{\text{layer}}$, with learned γ substantially positive at peak layers, so the cumulative sum entering the residual stream is amplified. **Norm reduces** in OLMo: standard RMSNorm semantics with γ values that on average reduce each contribution's magnitude before it enters the residual stream. **Norm identity at runtime** in the three pre-norm-only families: normalisation is applied at sublayer *input*, so $\hat{\gamma} = x$ exactly.

The Gemma $A'(x) \approx 90 \times A'(\hat{\gamma})$ ratio sits in a super-linear-but-sub-quadratic regime: A' depends super-linearly on per-layer salience (salience^2 contribution to local surface area plus trajectory length), so $(1 + \gamma)$ scaling drives super-linearly larger A' as the trajectory accumulates additively across depth (Appendix D.7 develops the γ-arithmetic). The directional claim (amplification vs reduction - opposite signatures within the Peri-LN class) is what the data firmly supports - the specific 90× is empirical, not closed-form-entailed.

Conditions other than PT show the same per-family pattern at slightly attenuated magnitudes for Peri-LN families: the PT-to-IT-completion weight change in Gemma reduces post-sublayer γ by 24-52% at deep layers, but direction is preserved within each family across conditions (Appendix D). The Gemma A' ratio is not concentrated at γ-peak layers: γ_{attn} peaks at L24 in PT and L17 in IT (Appendix B.1), but the minimal-pair divergence peak in PT Gemma is at L4, shifting to L26 in IT-completion and IT-chat (section 4.6). Architectural amplification accumulates across depth rather than concentrating at γ-peak layers, consistent with the L2-norm check below (~150× at the final layer). This decouples our architectural amplification from the layer-localisation findings of Pandey [2] (Gemma-2-2B L15 · H6) and Genadi, et al. [4] (Gemma-3-4B L10-15) - both studies are on different Gemma variants, so cross-paper layer-correspondence needs direct testing.

Where the direction comes from. The identity result for pre-norm-only families is mathematically entailed (no post-sublayer normalisation to scale sublayer outputs). For the two Peri-LN families, direction is empirically determined by learned post-sublayer normalisation parameters, not by structural placement alone. Gemma's $(1 + \gamma)$ parameterisation with γ_{attn} 103-3239 and γ_{ff} 124-5673 (Appendix B.1), positive at every layer, enlarges each contribution by an order-γ_{layer} factor - but magnitude and direction depend on the *learned* values, not the parameterisation form. OLMo's standard RMSNorm reduces in our measurements, but standard RMSNorm semantics do not by themselves entail reduction - observed reduction depends on learned scales and activation norms. A third Peri-LN family in our sample would have been informative on whether different parameterisations or learned-γ regimes could produce different directions.

Empirical metric-robustness check. A metric-free analog: the ratio of L2-Euclidean norms $\|x_{\text{final}}\|_2 / \|\hat{\gamma}_{\text{final}}\|_2$ at

the answer-start position (no G involved), per cell, across 432 PT cells per family (`results/architectural-norm-ratio-robustness.json`).

Table 6. L2-norm robustness check of the architectural-axis signature. Computed at the final-layer answer-start position; mean and median ratios per family compared against the A'-under-G ratio from Table 5. Direction agrees across both metrics; magnitudes differ because A' integrates across all (token × layer) positions while L2 reads the final state only.

Family	A' ratio under G (Tbl 5)	L2 ratio (no G)	Direction match
Gemma-3-1b	$A'(x) \approx 90 \times A'(\hat{\gamma})$	mean 150.7×, median 150.5×	✓ amplifies (L2 ratio even larger)
Llama-3.2-1B	$\hat{\gamma} = x$ by construction	1.000	✓ identity
OLMo-2-1B	$A'(x) \approx A'(\hat{\gamma})/7$ (≈ 0.14)	mean 0.0962, median 0.0965	✓ reduces (L2 ratio even smaller)
Qwen3-1.7B	$\hat{\gamma} = x$ by construction	1.000	✓ identity
SmolLM2-1.7B	$\hat{\gamma} = x$ by construction	1.000	✓ identity

All five families show the same direction under both metrics - magnitudes differ (Gemma's amplification and OLMo's reduction are both *larger* under L2 than under A') because A' integrates across all (token × layer) positions while L2 captures only the final-layer answer-start state. The architectural-axis claim is metric-robust at the direction level - specific magnitudes are A'-under-G specific.

The architectural axis is largely independent of the post-training and chat-template axes - post-training does not change which architectural regime a family is in. But the architectural regime modulates what the other axes can do: per-cell variance from non-architectural effects passes through the architectural amplification or reduction on its way into the residual stream.

4.4 The post-training axis

The post-training axis measures the IT-completion / PT margin-magnitude ratio, with scaffolding off. The aggregate pattern is in Table 3 - here we report per-framing structure across the four framings of the third sentence.

Table 7. Per-framing post-training ratios on M cells. Each entry is the per-framing IT-completion / PT margin-magnitude ratio with bootstrap 95% CI. F1 is the softest framing ("The text says X has more") - F4 is the strongest ("Despite the values, X has more").

Framing	Gemma	Llama	OLMo	Qwen3	SmolLM2
F1 soft	1.83 [1.53, 2.22]	2.24 [1.62, 3.02]	3.53 [3.07, 4.14]	1.27 [1.11, 1.43]	1.02 [0.94, 1.10]
F2 flat	1.39 [1.16, 1.67]	1.55 [1.05, 2.25]	2.58 [2.27, 2.94]	0.81 [0.66, 0.96]	1.05 [0.95, 1.15]
F3 adversative	2.52 [1.90, 3.25]	1.67 [1.26, 2.22]	1.87 [1.56, 2.28]	1.48 [1.26, 1.71]	0.78 [0.69, 0.88]
F4 override	3.35 [2.87, 3.92]	1.97 [1.48, 2.61]	1.00 [0.84, 1.15]	0.61 [0.49, 0.74]	0.96 [0.83, 1.11]

The post-training-by-framing ordering differs qualitatively across families. OLMo is highest on soft (3.53×) and falls to 1.00× on override - the "post-training fills headroom on already-resisted framings and saturates on already-overridden ones" pattern. Gemma is the inverse (override 3.35×, flat 1.39×). Qwen3 and SmolLM2 stay near 1.0× across framings. The natural saturation hypothesis fits OLMo cleanly, is inverse in Gemma, and noisier in the others - whatever determines per-framing shape varies across families. Among the three ~2× families the aggregate is consistent but distributed differently per framing - the recipe-name label (section 5.3) is insufficient to predict per-framing structure.

The post-training-axis effect is captured by vocab-token-reference metrics (the margin from section 3.2) and is largely invisible to intrinsic geometric metrics (κ, S, A' on x), which characterise trajectory *shape* rather than alignment with vocabulary directions. This is the mechanistic ground for why architecturally similar families can have very different post-training effects despite similar geometric envelopes: post-training rotates the trajectory in vocabulary-aligned subspace, not in the directions intrinsic metrics measure.

4.5 The chat-mode scaffolding axis - prompt-form and template-token sub-effects

The chat-mode scaffolding axis decomposes into two multiplicative sub-effects via the IT-question-no-template fourth condition (section 3.1, Appendix D.4): a **prompt-form** shift (completion seed → question form, no template) and a **template-token** application (role markers, turn delimiters, BOS on top of the question form). The combined scaffolding ratio in Table 3 equals their product within bootstrap precision.

Metric note. Table 8 reports the **magnitude ratio** - the standard logit-lens readout. The correctness tables that follow report the **signed direction** (capture rate, the proportion of cells with negative signed margin). Magnitude and direction can diverge: a small magnitude ratio can correspond to a large correctness shift when cells sit near zero (SmolLM2) - a large ratio can leave correctness unchanged for already-saturated cells. Both readouts appear because each carries different information (relationship glossed in section 1.1 and section 3.5).

Table 8. Per-family chat-mode scaffolding axis decomposition on M cells, magnitude-ratio readout. Bootstrap point estimates with cell-level 95% CIs (10,000 iterations) - "Combined" matches Table 3's scaffolding ratio.

Family	Prompt-form ratio	Template-token ratio	Combined (Table 3)
Gemma-3-1b	0.95× [0.85, 1.07]	**4.83× [4.42, 5.30]**	4.61× [4.24, 5.04]
Llama-3.2-1B	0.94× [0.83, 1.08]	**2.18× [1.86, 2.55]**	2.05× [1.72, 2.45]
OLMo-2-1B	**0.52× [0.45, 0.59]**	**2.09× [1.80, 2.44]**	1.08× [0.95, 1.23]
Qwen3-1.7B	0.93× [0.79, 1.12]	**0.74× [0.62, 0.89]**	0.69× [0.57, 0.83]
SmolLM2-1.7B	**0.65× [0.58, 0.72]**	**1.81× [1.60, 2.04]**	1.17× [1.05, 1.31]

Two patterns: **template-token-dominant** (Gemma, Llama, Qwen3 - prompt-form CIs include 1.0 - entire magnitude in the template-token sub-effect) and **two-opposing-effects** (OLMo, SmolLM2 - prompt-form below 1.0; template-tokens above 1.0; partial-to-full cancellation).

A content-independent positional template-token effect; uniformly detrimental on this canvas. Looking at M-cell *correctness* tells a complementary story:

Table 9. Per-family M-cell correctness across the four conditions, with axis-effect deltas (post-training / prompt-form / template-token) on correct-cell counts.

Family	M correct (n/192)				→ axis effects
	PT	IT-comp	IT-q-no-t	IT-chat	(post-training / prompt-form / template-token)
Gemma	42	93	99	2	+51 / +6 / **−97**
Llama	97	122	100	48	+25 / −22 / **−52**
OLMo	34	33	83	24	−1 / +50 / **−59**
Qwen3	17	43	67	43	+26 / +24 / **−24**
SmolLM2	0	0	106	75	0 / +106 / **−31**

The template-token effect on M-cell correctness is **detrimental in four families** (Δ correct cells: −31 to −97 for Gemma, Llama, OLMo, SmolLM2; Qwen3's −24 under default thinking-on is the routing artefact described in Appendix D.3 - under thinking suppression the template step instead gains 20 M cells, 67→87). The aggregate scaffolding direction's family-variability (detrimental / neutral / beneficial - section 5.2) lives entirely in the family-variable *prompt-form* sub-effect: detrimental in Llama (−22), neutral/modestly positive in Gemma (+6) and Qwen3 (+24), strongly positive in OLMo (+50) and SmolLM2 (+106).

The important mechanism claim is that the template-token sub-effect imposes a **content-independent positional rotation** toward the subject of the third-sentence assertion, uniform across four families on this canvas. The paired NM cells supply the evidence: NM is constructed by swapping the comparator word (`more` → `fewer`) and preserving every other token, so the assertion's subject does not change between M and NM (e.g. "Bob" remains the subject of "Bob has more / fewer books than Alice"). The correct answer to "who has more?" is the other entity in both cases (Alice, who has the larger value, since the comparator-swap does not change the underlying numerical comparison). NM correctness drops at the template step in four families (Gemma 150→48, Llama 120→52, OLMo 93→38, SmolLM2 124→99) and in Qwen3's default thinking-on configuration (93→37) - the same direction as the

M-cell drop. The template tokens therefore push toward the assertion-subject regardless of whether the assertion's truth-value is misleading or non-misleading - a content-independent rotation, not a truth-conditional behaviour. The "uniformly detrimental on this canvas" framing is the canvas-contingent corollary: on this canvas the assertion-subject is consistently not the data-correct answer, so rotation toward the assertion-subject reduces correctness on both M and NM. The deployment-regime intuition (templates are part of the trained regime and should match or amplify post-training-installed behaviour) predicts the opposite of what we observe. The prompt-form sub-effect varies because families differ in completion-seed-vs-question-form sensitivity - the template-token sub-effect is consistent because the template induces the same positional rotation across families (section 4.6, section 5.4).

The cross-family uniformity is strict for four families (Gemma, Llama, OLMo, SmolLM2) on both M and NM readouts. Qwen3 under its default thinking-on configuration shows the same template-step direction (M −24 cells, NM 93→37) - but Appendix D.3's thinking-mode survival check reveals this is the routing artefact reading the wrong position. With thinking suppressed (`enable_thinking=False`), Qwen3 reverses direction at the template step: M correctness gains 20 cells (67→87), NM gains 13 (93→106). The "uniform across all five families" reading is conditional on Qwen3's default regime; under thinking-mode suppression Qwen3 shows the opposite direction. Mechanism for this reversal is not established by the present data.

Canvas-scope. The "uniformly detrimental template-token" finding is observed on the misleading-assertion canvas (paired (M, NM) cells under contradictory and consistent assertions, same arithmetic-comparison structure, same "who has more?" form). The M+NM extension shows the template-token effect biases toward the third-sentence-subject regardless of truth value, strengthening the mechanism claim (section 5.4) but not extending canvas-scope. On structurally different task surfaces (instruction-following, format-adherence, multi-turn dialogue, refusal, tool-use) the sub-effect could go in any direction - cross-canvas validation is required. The cross-family universality here is a property of (template-token × this canvas).

Per-position structure. Per-position decomposition of the aggregate magnitude ratio (Table 8 "Combined"):

Table 10. Per-position decomposition of the aggregate scaffolding magnitude ratio (Table 8's "Combined" column). high_first cells = misleading-named entity is first-mentioned; high_second cells = misleading-named entity is second-mentioned.

Position	Gemma	Llama	OLMo	Qwen3	SmolLM2
high_first	4.52 [4.00, 5.19]	2.91 [2.26, 3.75]	0.66 [0.58, 0.75]	0.46 [0.36, 0.58]	1.10 [0.95, 1.27]
high_second	4.69 [4.19, 5.26]	1.52 [1.20, 1.94]	2.48 [2.12, 2.92]	1.36 [1.07, 1.69]	1.26 [1.06, 1.48]

Four of five (Llama, OLMo, Qwen3, SmolLM2) show position-dependent rotation - Gemma is symmetric (4.52× / 4.69×). For the rotated families the effect is larger when the misleading-labelled entity appears first - consistent with the position-bias account (section 4.6, section 5.4). The aggregate conceals the rotation.

Per-suite refinement. The M-cell decomposition masks heterogeneity across M / NM / capability suites. Examples: SmolLM2's M scaffolding is 1.17× but NM 3.28× and capability 3.47×; Gemma's M is 4.61× but capability 6.90×; OLMo's M is 1.08× (neutral) but capability 0.75× and NM 1.62×. The M aggregate is the headline metric but does not characterise full per-family behaviour (Appendix D.4).

4.6 Minimal-pair (M, NM) trajectory divergence

For each M cell we constructed a paired NM cell by swapping a single word in the third sentence (`more` → `fewer`), preserving every other token. The cells are token-identical before the swap, so residual streams at those positions are identical by construction. Divergence at position (token t, layer ℓ) is $1 - \cos_G(x_M, x_{\mathrm{NM}})$ under the pullback metric G (section 3.2).

Localisation across families (pre-registered H1). Across all five families and four conditions, pre-swap divergence is zero (the structural property holds empirically). Post-swap divergence is sharply concentrated near the swap and decays at far-from-swap positions - the at-swap-to-post-swap-far ratio ranges 10× to 280× across families and conditions. Localisation is empirically clean.

Capture-status correlation by condition (pre-registered H2). Captured cells (margin < 0) have higher post-swap integrated divergence in IT-completion across 4 of 5 families (Cohen's d 0.93 to 2.63, bootstrap CIs separated from zero). In PT and IT-chat the relationship is mixed across families. IT-completion is the condition

in which captured-vs-uncaptured divergence aligns most cleanly with behavioural outcome - the divergence there reflects the model's actual processing difference between contradictory and consistent assertions. In PT, capture is dominated by base-model surface-form heuristics - in IT-chat, by the template-induced position bias developed in section 5.4. IT-completion isolates "the model is captured because it deliberated differently".

Chat-mode scaffolding amplification of divergence (pre-registered H3). Directionally supported in 3 of 5 families but all below the pre-registered $> 1.5\times$ threshold (Gemma 1.48×, OLMo 1.09×, SmolLM2 1.06×) - hence the REFINED verdict in Appendix A.2. In the other two families scaffolding *reduces* divergence (Llama 0.81×, Qwen3 0.64×). Variability matches the magnitude-axis variability from section 4.5. The IT-question-no-template condition supplies the data to decompose further into sub-effects on divergence; the family-specific NM-flip mechanism is developed in section 5.4.

Architectural concentration of divergence (pre-registered H4). Peak-layer / cross-layer divergence ratio is large in Gemma across all conditions - 6.28× in PT (peak layer 4 of 26), 3.15× in IT-completion (peak L26), 8.76× in IT-chat (peak L26). Other families show moderate concentration: ~2.4-2.8× in OLMo (peak around L10-12 under post-training); ~2.0-2.5× in the three pre-norm-only families (peaks at mid-stack L7-16). Strong concentration at a single sharp layer is Gemma-specific, and the peak shifts from L4 in PT to L26 after post-training - a layer-shift the other four families do not show. This suggests norm semantics (Gemma's $(1 + \gamma)$ post-norm) drive the strong-concentration signature - structural placement (Peri-LN classification) alone does not.

4.7 Per-cell interactions reveal directional-rotation effects

The aggregate ratios in Tables 3, 7, and 8 are tight and meaningful as aggregates - the per-cell variance around them has interpretable structure. (Diagnostic: the multiplicative-magnitude model fits aggregate data adequately but has per-cell SD log-ratio 0.57 to 1.54 across families, well above the < 0.5 "tight fit" threshold pure magnitude scaling would predict. Good in the mean, not at the level of individual cells. Section 4.8 reports this as the D1 pre-registered-criterion refutation.)

Two structured patterns emerge from the per-cell breakdowns in section 4.4-4.5.

Position-dependent rotation in the chat-mode scaffolding axis. The per-position decomposition in section 4.5 (Table 10) shows that for four of five families, the scaffolding effect's *direction* depends on whether the misleading-labelled entity appears first or second. The aggregate averages across positions and conceals the rotation. The (M, NM) analysis in section 4.6 makes the account concrete: template application appears to rotate the model's preference toward a specific structural position (the subject of the third-sentence assertion) regardless of truth-value content - aggregate magnitude depends on whether that rotation aligns with or opposes the existing position-driven preference.

Framing-dependent rotation in the post-training axis. The per-framing decomposition in section 4.4 (Table 7) shows that for three of five families, the post-training effect's magnitude varies substantially across framings, with the strongest-effect framing differing across families. Aggregate ratios average over framings - per-framing structure shows the effect interacts with the assertion's lexical surface in a family-specific way.

A natural reframing: the protocol decomposes axes by *magnitude*, but the underlying effect has both magnitude and direction. Aggregate ratios capture magnitude - per-cell breakdowns surface direction. Future work could formalise the magnitude-and-direction decomposition explicitly - separating "the axis amplifies the model's pre-existing direction by N×" from "the axis rotates preference toward a specific structural position by some angular measure".

A concrete example. OLMo's aggregate post-training × scaffolding = $1.94 \times 1.08 \approx 2.10$, matching the observed 2.09 within bootstrap precision. But the scaffolding axis has a 0.66× / 2.48× position-rotation structure (Table 10): within high_first cells the per-position prediction is $1.94 \times 0.66 \approx 1.28$; within high_second cells, $1.94 \times 2.48 \approx 4.81$. These two predictions differ by 3.8× even though both average into the same 2.10. The same applies to Gemma's per-framing rotation (1.83× on F1 vs 3.35× on F4) and to Qwen3's signed-margin sign-flip on capability controls (PT +3.07, IT-chat −2.24 - a magnitude-only ratio cannot represent it at all).

Both statements are supported: the aggregates are real and tight (Table 3, Appendix D.1), and the per-cell variance has interpretable family-specific structure. The magnitude-only model captures the first half cleanly and leaves the second as the developable finding.

4.8 What we expected vs what we found

Three predictions about the protocol's behaviour were committed before any capture (full record: Appendix A.2). Two were refuted - one held family-specifically. Each refutation is a substantive empirical finding.

D1 REFUTED. Cell-level multiplicative factorisation does not hold at the strict threshold. The pre-committed criterion required per-cell SD log-ratio < 0.5 for the post-training $\times$ scaffolding decomposition to count as a tight per-cell model. Observed SD log-ratio is 0.57 to 1.54 across families - well above threshold. The aggregate ratios in Table 3 remain tight and meaningful - per-cell residual variance has interpretable directional structure (position-rotation in scaffolding, framing-rotation in post-training) the magnitude-only model does not capture (section 4.7 - see section 5.6 for the connection to the deployment-conditional thesis).

D2 REFUTED, qualitatively in the opposite direction - the dispersion comparison is fragile at $n = 4$ but the directional finding is robust under single-family dropouts of the dominant outliers. The D2 prediction is the operational assumption underlying chat-only IT evaluation, stated as a testable hypothesis: chat-template tokens are often treated as a largely shared inference-time convention; alignment recipes vary substantially; therefore scaffolding should be more cross-family-consistent than post-training. The assumption is implicit, carried by chat-only methodologies that attribute observed differences to post-training; several recent papers have begun to question whether it holds (Minder, et al. 2025; Zhao et al. 2025; the "Curious bos_token" LessWrong report; the "Price of Format" study). The cross-family data shows the opposite: scaffolding ratios span 1.08× to 4.61× across the four headline families (a 4.3× range) while post-training ratios span 0.95× to 2.03× (~2× range). The CV summary (scaffolding 0.74; post-training 0.30) tracks the same directional finding, but at $n = 4$ the CV is sensitive to single-family substitution: dropping the highest-scaffolding family (Gemma) tightens scaffolding CV to ~0.37; dropping the lowest-post-training family (SmolLM2) tightens post-training CV to ~0.04. The comparison therefore rests on which side has the longer tail rather than on a stable dispersion ratio. Including Qwen3's default-thinking IT-chat regime (where fixed-position margin extraction is routing-blind, Appendix D.3) extends scaffolding to 0.69× to 4.61× / span 6.7× / CV 0.82 and post-training to CV 0.35; the qualitative finding (scaffolding more recipe-variable than post-training, with the longer tail on the scaffolding side) is robust to this choice and to the single-family dropouts above. What looks like more uniform inferential machinery in the chat scaffold interacts with post-trained weights in more family-variable ways than alignment-recipe diversity itself produces. The assumption should not be carried forward without per-family verification.

D3 REFINED. Post-training weight changes install misleading-assertion resistance in some families but not all. The prediction was that IT-completion margins would show large cell-level sign-flips relative to PT across families that exhibit capture in IT-chat. CONFIRMED in Gemma (78% → 52% capture, a 26-point reduction; caveat: contemporaneous with capability gain on the same prompt-form, Appendix D.5) and Llama (50% → 37%). REFUTED in OLMo (82% → 83%, flat), SmolLM2 (100% → 100%, no installation), and Qwen3 (91% → 78%, modest improvement). Family-specific verdict, holding for the three ~2× families (section 5.3) and absent for the two ~1× families. Per the reading discipline (section 3.1, section 4.2), "install" refers to the aggregate PT-vs-IT-completion contrast and is not attributable to specific recipe stages.

The second refutation is the most consequential for downstream methodological practice. Without the protocol, the operational assumption that template tokens are more recipe-universal than post-training effects would be carried forward as implicit background rather than tested. The cross-family data refutes that assumption directly - and the prompt-form / template-token decomposition in Appendix D.4 sharpens the refutation by showing that even within "chat scaffolding", the relative contribution of prompt-form vs template-tokens is itself family-variable.

5. Discussion

5.1 Cross-family results don't generalise across nominally-similar architectures

The five-family decomposition makes a methodological point: results obtained on one model family cannot be assumed to transfer to others even when they share the same nominal architectural class and run on the same task. The Gemma-OLMo pair is the clearest demonstration.

Both Gemma-3-1b and OLMo-2-1B are classified as "Peri-LN" (each sublayer's output normalised before being added to the residual stream), both are roughly 1B parameters, both were tested on identical prompts under identical conditions. Yet the protocol's three axes produce qualitatively different signatures: chat-mode scaffolding 4.61× in Gemma vs 1.08× [0.95, 1.23] in OLMo (statistically indistinguishable from no effect) - A' ratio 90× amplification in Gemma vs ~7× reduction in OLMo. Same Peri-LN class - opposite empirical direction.

The difference traces to parameterisation within the Peri-LN class. Gemma uses $(1+\gamma)$ scaling with substantial trained γ throughout (range 103-3239 across 26 layers, Appendix B.1), enlarging each sublayer's contribution - OLMo uses standard RMSNorm with γ values two orders of magnitude smaller (range 3.5-17.1) that empirically reduce contributions. The "Peri-LN" label captures the structural fact - the empirical effect depends on parameters the label doesn't specify. Model-card categorical labels (norm class, attention design (GQA / MHA), alignment recipe name, RoPE base) systematically under-specify mechanism.

Scope-disciplined: among just five models we observed three distinct architectural signatures (pre-norm-only $\hat{\gamma} = x$ by construction in 3 families; Peri-LN with $(1+\gamma)$ amplifies A' in Gemma; Peri-LN with standard RMSNorm reduces A' in OLMo). We do not claim non-transfer is universal across model-design space - instead we claim it is frequent enough in this sample that single-family results should be reported as single-family results until cross-family decomposition under the protocol's controls confirms generalisation. Section 5.6 develops the broader-recommendation framing.

5.2 The chat-mode scaffolding axis: a content-independent positional template-token effect plus a family-variable prompt-form effect

The chat-mode scaffolding axis (switching from completion seed to question form plus applying chat-template tokens) produces three different aggregate directions on M-cell capture rate (detrimental in Gemma and Llama; near-neutral in OLMo and Qwen3; beneficial in SmolLM2). The four-condition decomposition (section 4.5) shows this trichotomy is the sum of two sub-effects with very different cross-family patterns:

- **The template-token sub-effect is a content-independent positional rotation toward the subject of the third-sentence assertion, uniform across four families on this canvas.** Change in correct-cell count when chat-template tokens are added to the same question-form prompt: M-cells Gemma −97, Llama −52, OLMo −59, SmolLM2 −31 (Qwen3 −24 under default thinking-on is the routing artefact described in Appendix D.3; under thinking suppression Qwen3 reverses direction, gaining 20 M cells 67→87); NM-cells Gemma 150→48, Llama 120→52, OLMo 93→38, SmolLM2 124→99 (Qwen3 default 93→37; under thinking suppression Qwen3 reverses to 93→106). The NM construction swaps only the comparator word (`more` → `fewer`) and preserves the assertion's subject, so the correct answer to "who has more?" is the other entity in both M and NM (the comparator-swap does not change the underlying numerical comparison). Despite very different alignment recipes, prompt-form responses, and aggregate scaffolding signs, every one of the four families shows the same template-step direction on *both* readouts: template tokens push toward the assertion-subject regardless of whether the assertion's truth-value is misleading or non-misleading. The mechanism is therefore content-independent (truth-value-independent) positional rotation, not a content-conditional detriment. The canvas-contingent corollary is that on the misleading-assertion canvas the assertion-subject is consistently not the data-correct answer, so the rotation is uniformly detrimental to correctness on both readouts. Cross-canvas validation is needed before generalising - on canvases where the assertion-subject is the data-correct entity, the same content-independent positional rotation may align with rather than oppose the correct response. Qwen3's no-think reversal is a separate finding worth noting - the mechanism is not established by this data.
- **The prompt-form sub-effect is family-variable in direction.** Change when the prompt switches from completion seed to question form (no template either side): Gemma +6 (flat), Llama −22 (detrimental), OLMo +50 (strongly beneficial), Qwen3 +24 (modestly beneficial), SmolLM2 +106 (installs resistance from zero baseline). The family-variation in the aggregate scaffolding-axis direction lives entirely here.

The aggregate trichotomy is therefore not three different intrinsic chat-mode behaviours - it is the sum of one consistent intervention (template tokens biasing toward the assertion-named entity) plus one family-variable intervention (prompt-form sensitivity). This sharpens the "deployment-conditional alignment" framing: the deployment-conditional component is the prompt-form sensitivity, not the template-token effect.

The two extremes illustrate. Gemma is template-token-dominated: PT 42/192 → IT-completion 93/192 → 99/192 → 2/192. The 47-percentage-point detrimental swing is essentially all template-token contribution. SmolLM2 is prompt-form-dominated: PT 0/192, IT-completion 0/192, prompt-form alone installs 106 cells of correctness, template-tokens then reduce it to 75. "Scaffolding installs SmolLM2's resistance" is correct at the aggregate level - mechanistically, prompt-form does the installing and template-tokens partially undo it.

Implications for evaluation methodology are direct. Standard alignment benchmarks (refusal evaluations, sycophancy benchmarks, instruction-following tasks) run the model in chat mode and attribute resulting behaviour to alignment training. Our results imply this is substantially wrong in three of the five families we tested:

- For Gemma, standard chat-only PT-vs-IT-chat comparison would observe 78.1% → 99.0% capture (+21pp) and attribute the increase to alignment training. The protocol reveals the opposite sign: alignment reduces capture by 27pp (PT 78.1% → IT-completion 51.6%) while scaffolding then overwhelms that reduction with +47pp (IT-completion 51.6% → IT-chat 99.0%). Same weights, opposite direction of attribution.
- For SmolLM2, standard chat-only comparison would observe 100% → 60.9% capture (−39pp) and attribute the reduction to alignment training. The protocol reveals alignment installed nothing (PT 100% → IT-completion 100%) and scaffolding alone installed all of the deployed resistance (IT-completion 100% → IT-chat 60.9%). The entire effect is mis-attributed.
- Qwen3 illustrates a third (qualitatively distinct) failure mode: standard chat-only comparison reads PT 46/48 → IT-chat 6/48 on capability controls as a catastrophic alignment-driven capability loss. The protocol reveals this is a measurement-pipeline artifact (chat template routes through prefix; fixed-position extraction reads the wrong location; Appendix D.3) — both prompt-form and template-token sub-effects contribute to the apparent loss (Appendix D.5).

The protocol is the diagnostic: running both conditions and reporting per-axis effects resolves these ambiguities at modest cost.

The deployment-conditional finding bears directly on AI-safety evaluation: alignment-trained safety properties are themselves *deployment-regime-conditional*. A model that resists misleading prompts under the chat-mode regime its training was tuned for may not resist them under raw-completion access, alternative system-prompt templates, or out-of-distribution wrappers. The bare observation (alignment's measured effect can be reduced or amplified by deployment-time formatting alone) is a methodological point current evaluations do not surface. (Refusal-direction: section 5.5 - broader implications: section 5.6.)

5.3 Model-card recipe labels are too coarse to predict alignment magnitude

The clearest defensible claim our data supports about the alignment axis is **negative**: model-card recipe labels are too coarse a summary to predict alignment magnitude. Llama-3.2-1B and SmolLM2-1.7B both list "SFT + DPO" as their recipe. Llama's alignment ratio is 1.87×; SmolLM2's is 0.95×. Same tag, two-fold different effect on this canvas. The tag is a thin description of two models that differ in pretraining-data composition, base architecture, DPO data composition, regularisation hyperparameters, and many other implementation details. The label thinness is what fails - not SFT + DPO as a procedure. Making a claim about recipes-as-procedures would require holding base architecture and data fixed and varying only the recipe; the five-family sample does not support that comparison. The well-supported claim here is about model-card summaries, not about training procedures.

Descriptive context: three of five alignment ratios sit near 2×, two near 1×. Three families - Gemma-3-1b (SFT + RLHF + distillation), Llama-3.2-1B (SFT + DPO), OLMo-2-1B (SFT → DPO → RLVR) - sit in 1.87× to 2.03×. Two - Qwen3-1.7B (SFT + DPO + GRPO with thinking/non-thinking hybrid) and SmolLM2-1.7B (SFT + DPO via Zephyr) - sit in 0.95× to 0.99×. With five points we describe - we do not infer cluster structure or claim a bimodal distribution.

Across the five families on this canvas, post-training installs M-cell resistance most cleanly in Llama (PT 49.5% → IT-completion 36.5% capture, a 13-point reduction with mid-range capability variation across conditions per Appendix D.5). Gemma's larger M-cell improvement (78.1% → 51.6%, 26 points) is contemporaneous with a 24/48 → 44/48 capability jump on the same prompt-form (Appendix D.5 - section 4.2 reading discipline), so part of the M-cell gain is confounded with general capability improvement at the PT-to-IT-completion step. OLMo, Qwen3, and SmolLM2 do not install M-cell resistance at the post-training step at all. The "post-training installs resistance" finding therefore rests most cleanly on Llama, with Gemma as a supporting-but-confounded case.

What does sit comfortably: among the three near ~2×, consistency is striking despite very different RL components (full RLHF, DPO-only, staged DPO → RLVR). 95% bootstrap CIs overlap around 1.9–2.0×. Speculative mechanism (untested): alignment training installs a relatively uniform confidence-amplification on whatever direction the base model was already preferring - recipe details determine *what* gets amplified, but magnitude is more recipe-invariant than the recipe-name space suggests.

Two follow-ups the data points at: Qwen3 with thinking-mode disabled (does the ratio shift toward 2×?), SmolLM2 base with a different DPO-style fine-tune (toward 2× or stay at 1×?).

What we can say with confidence: the alignment-ratio split does not follow from recipe-name labels, and alignment

effects are more consistent across families than chat-mode scaffolding effects (CV comparison in section 4.1). Both observations refine standard reporting practice.

5.4 NM-flip mechanism: position-bias account is template-driven in some families, prompt-form-driven in others

The paired-prompt (M, NM) design (section 3.3) lets us adjudicate a specific mechanism candidate at the chat-mode-scaffolding step. The position-bias account: chat-template application induces an attention rotation toward the entity named in third-sentence position regardless of whether the sentence's content is true or false. The prediction: NM-cell margins should flip sign at the step where this position bias is introduced.

The four-condition decomposition (section 3.1, Appendix D.4) tests where the NM-flip happens. We track the proportion of NM cells answered correctly across PT, IT-completion, IT-question-no-template, and IT-chat:

Table 11. Per-family NM-cell correctness across the four conditions, with the NM-flip driver identified by where the major drop occurs.

Family	NM correct (n/192)				→ NM-flip driver
	PT	IT-comp	IT-q-no-t	IT-chat	
Gemma-3-1b	61	123	150	**48**	**template-tokens** (drop only at chat-template step: 150 → 48)
Llama-3.2-1B	158	184	120	**52**	mixed (drop starts at prompt-form: 184 → 120; full flip at template: 120 → 52)
OLMo-2-1B	134	156	**93**	38	**prompt-form** (major drop already at question form: 156 → 93)
Qwen3-1.7B	192	192	**93**	37	**prompt-form** (catastrophic drop at question form: 192 → 93)
SmolLM2-1.7B	159	139	124	99	no flip - gradual reduction, never sign-flips on average

The position-bias account decomposes cleanly:

- **Confirmed in Gemma.** NM correctness rises through IT-completion and IT-question-no-template (61 → 123 → 150), then collapses at the chat-template step (150 → 48). Template tokens are the trigger, exactly as the account predicts.
- **Mixed in Llama.** NM correctness peaks at IT-completion (184) and drops in two stages - at prompt-form (184 → 120) and again at template (120 → 52). Both sub-effects contribute - the account is consistent with the template-step drop but not the prompt-form-step drop.
- **Contradicted on this canvas in OLMo and Qwen3.** Both drop sharply at the prompt-form step with no chat-template applied. OLMo: 156 → 93; Qwen3: 192 → 93 (near-perfect-to-half collapse). Template tokens contribute further reduction (OLMo 93 → 38; Qwen3 93 → 37) but the dominant flip happens before any template is added. The position-bias account cannot be the primary mechanism for these two.
- **Not applicable in SmolLM2.** NM correctness stays positive across all four conditions (159, 139, 124, 99) - no sign-flip in the aggregate.

This is a substantive family-specific mechanism adjudication. The position-bias account is correct for Gemma; partly correct for Llama; incorrect as the primary story for OLMo and Qwen3 (where prompt-form sensitivity dominates). A more accurate field-wide framing: **NM-flip in chat-mode evaluation can arise from at least two distinct mechanisms (template-token-induced position bias and prompt-form-sensitive routing) and which dominates is family-specific.** Single-family mechanism claims about "what chat scaffolding does" should be reported as family-specific until cross-family validation runs.

Why might the question-form prompt drive the flip in OLMo and Qwen3? *(Speculation - untested.)* The question-form prompts (e.g. "Alice has 18 books. Bob has 15 books. The text says Bob has fewer books than

Alice. Who has more books?") have a structure uncommon in raw web text: data setup, third-sentence assertion, direct question. When a family's pretraining distribution under-represents this structure, the IT model fed the question form without chat-template wrapping may fall back to a default strategy - a most-recent-assertion-following or homework-Q&A-completion heuristic that picks up on the assertion-named entity. Template tokens added on top sit on a residual stream already routed away. Families with more web-text-like exposure to question-form math word problems handle it directly - only the template tokens then introduce the position-bias rotation. Other plausible accounts: tokenizer-specific question-mark handling, attention-head priors specific to certain training data, mode-routing recipes like Qwen3's thinking/non-thinking hybrid.

Testable consequence. The pretraining-data-composition reading predicts the prompt-form-driven NM-flip should correlate with the proportion of web-text-like math-word-problem question structures in the pretraining mix. On a controlled axis where pretraining mix varies (instruction-heavy vs web-heavy vs code-heavy vs multilingual-balanced) with post-training and architecture held more constant, families with more raw-web-text "data + question" exposure should show *smaller* prompt-form sub-effect magnitudes. Pythia's Pile composition [34],[35] or the OLMo training corpora data-mix descriptions are candidates. A direct mechanistic test would require head-level interventions across the four conditions which is out of scope for this paper.

The family-specific mechanism still supports a falsifiable head-level prediction:

> In families where the position-bias account is confirmed (Gemma) or mixed (Llama), heads attending disproportionately to third-sentence-named-entity positions in IT-chat should, when ablated, simultaneously reduce M-cell capture and NM-cell flip relative to IT-question-no-template - with ablation effect magnitude tracking the IT-question-no-template → IT-chat correctness delta (Gemma 102 cells, Llama 68 cells). In families where the account is contradicted (OLMo, Qwen3), the same ablation should produce smaller effects. Validating requires head-level interventions across families under the four-condition protocol.

We do not run head ablations here, but the cross-family four-condition data is the kind of evidence that would precede such an investigation. The aggregate scaffolding ratios in section 4.5 give the chat-mode-scaffolding axis's *magnitude*; the NM-flip table above gives evidence about its *mechanism* - more heterogeneous across families than the previous three-condition analysis could surface.

5.5 Implications for other interpretability findings

The protocol's most directly transferable implication for the broader interpretability literature is a **burden-of-proof reframing**: if a published finding claims that post-training installed mechanism X, the claim should survive the IT-completion control before being reported as a post-training effect. If the finding only appears in IT-chat captures, the more accurate phrasing is "the deployed chat regime elicits mechanism X" - not "post-training installs mechanism X". This is a vocabulary discipline that lets readers tell which category of claim is being made.

Three operational ambiguities recur in the current literature. First, calling a PT-vs-IT-chat difference an "alignment effect" - the chat-template contribution ranges from negligible (OLMo) to dominant (Gemma's 51.6% → 99.0% from chat-template alone). Second, calling a feature "chat-model-specific" when it may be elicited by chat-mode scaffolding rather than installed by post-training - SmolLM2 is the cleanest illustration (post-training installs no resistance; scaffolding installs all of it, Appendix D.4). Third, calling a circuit "absent in the base model" when the base has never been put into a comparable elicitation regime - Qwen3's capability-control inversion (PT margin +3.07, IT-chat −2.24) is a case where the PT model has the capability the IT model appears to lose. The protocol resolves each at one additional capture per family.

Three strands particularly susceptible. Refusal-direction work (Arditi 2024 and follow-ups) is established exclusively in chat mode; the IT-completion control would show whether the direction survives chat-template removal. RLHF / DPO / RLVR / GRPO alignment-effect papers test post-alignment models in chat mode; IT-completion supplies the alignment-isolated baseline. Activation-steering and SAE-feature-difference work contrasts PT against IT-chat; whether "alignment-introduced features" survive an IT-completion control is a per-method question the protocol directly tests. The protocol is a constructive refinement, not a critique. Whether the controlled test changes any specific conclusion is an open question per canvas; we make no extrapolation from the misleading-assertion canvas to those task surfaces.

The broader framing - alignment as deployment-conditioned phenomenon - is developed in section 5.6 conditional on cross-canvas generalisation. The burden-of-proof refinement above is independent of that broader claim - the IT-

completion control supplies a check on the post-training attribution regardless of whether the deployment-conditional hypothesis generalises.

5.6 Beyond methodology: alignment as a deployment-conditioned phenomenon (conditional on cross-canvas generalisation)

Scope statement first. All findings here are on a single canvas (the Pandey/Halawi misleading-assertion stress test - a sycophancy/fragility canvas where the assertion-named entity is the candidate the prompt structurally points at) across five open-weight ~1B families. The uniformly-detrimental template-token finding is canvas-specific: on this canvas template tokens may be doing what they are intuitively suspected of doing (biasing toward the assertion-position entity) - on instruction-following, refusal, jailbreak-robustness, tool-use, or format-adherence canvases the same tokens may be doing exactly what they were post-trained to do. The data establishes what the protocol's controls let us *test* - on this canvas the answers are robust enough across five families to license a broader framing-shift hypothesis.

The framing-shift hypothesis the protocol's results suggest, conditional on the pattern holding across other canvases and at larger scale:

> **Alignment should be evaluated as a deployment-conditioned phenomenon, not merely as a weight-space intervention. Chat scaffolding is not neutral presentation - it is part of the causal system that produces model behaviour.**

The conceptual shift for interpretability is from *"what circuit causes behaviour X?"* to *"under what deployment regime does X appear, and which component of the stack supplies it?"* The unit of explanation becomes the regime-conditioned mechanism. A finding that correctly identifies the circuit can still be incomplete because the same circuit can be dormant in a different regime, weakened by a prompt-form change, or amplified by template tokens.

On the "intended use" objection. A reviewer might object that IT-completion is off-distribution inference (IT models were post-trained with chat templates, so feeding them raw text is "using the model wrong") making the contrast measure regime-shift artefacts rather than alignment proper. Yun, et al.'s "consistency between instruction tuning and inference-time prompting matters more than template presence per se" (section 1.4) is the principled version. The reply: IT-completion is a *measurement instrument* for factoring weight effects from format effects, not a deployment recommendation. The structural regularities of IT-completion measurements (per-cell multiplicative decomposition within bootstrap precision, coherent per-family ratios, SmolLM2's 100% capture rate vs Gemma's 51.6%) are evidence the measurement carries family-specific signal rather than degenerate "model used wrong" noise. Yun, et al.'s principle is methodologically aligned with the framing here: they observe *what regime the model operates in matters*, which is the scaffolding-axis claim the protocol operationalises.

D1's failure as evidence for the deployment-conditional thesis. The pre-registered cell-level multiplicative model (D1) was refuted at the strict criterion in all five families (per-cell SD log-ratio 0.57 to 1.54 vs the < 0.5 threshold; section 4.8). The residual variance is not noise: it has interpretable directional structure (position-rotation in the scaffolding axis, framing-rotation in the post-training axis; section 4.7). One reading of this pattern is that post-training weights were optimised *conditional on* the chat template being present at inference, so weights and scaffolding are not independent causal variables whose ratios multiply cleanly per cell. The aggregate multiplicative ratios are tight at the cell mean (the decomposition is a good approximation in the mean) and it breaks down precisely at the level where the mechanism lives. The clean factorisation captures the magnitude story - the directional residual structure captures the entanglement that the deployment-conditional framing in this section operationalises. D1's refutation is therefore consistent with the deployment-conditional thesis rather than orthogonal to it.

Two layered mechanisms behind the deployment-conditional framing. The content-independent template-token rotation (section 4.5) and D1's per-cell entanglement (above) might appear to point in different directions - a mechanical artefact orthogonal to alignment versus weights and scaffolding co-adapted. They describe layered components of one picture. The template tokens impose a content-independent positional rotation (a fixed property of the chat-template structure that holds across families on this canvas). Post-training weights were then optimised in the inference regime where that rotation is present, so the weight-level behaviour the field attributes to "alignment" is the response of weights co-adapted to a regime that includes the rotation. The IT-completion contrast does not subtract a mechanical artefact and leave pure alignment behind - it removes the regime the weights were tuned for, revealing how the weights respond when run outside that regime. The deployment-conditional thesis applies to this

combined behaviour, not to the rotation in isolation. D1's per-cell factorisation fails because the two layers are not independent factors that multiply - they are layered mechanisms whose interaction is what the protocol decomposes.

Prompt-form variability is where the deployment-conditional behaviour actually lives. Within the layered picture: the template-token rotation is the mechanical-regime layer (uniform across families on this canvas, section 4.5) and the prompt-form sub-effect is the family-variable layer where weights respond differently to the regime (Gemma +6, Llama −22, OLMo +50, Qwen3 +24, SmolLM2 +106 correct-cell change at the prompt-form step). The deployment-conditional thesis is therefore carried most directly by the prompt-form sub-effect - it is what makes the same weights look different depending on whether the question is phrased as completion seed or direct question, in family-specific ways. The template-token finding is the cleaner empirical result; the philosophically central one is the prompt-form variability. The mechanism account in section 5.4 (pretraining distribution exposure to question-form problems predicts prompt-form sensitivity) is the testable consequence - currently untested but specified for replication.

> **Reporting recommendation: a diagnostic vocabulary the protocol supplies.** Features observed in an IT-chat model can be classified along the four-condition decomposition into four feature-category candidates, plus one evaluation-pipeline category to rule out first:
>
> 1. *Post-training-installed* - survives the IT-completion control (PT-vs-IT-completion contrast).
> 2. *Template-token-elicited* - survives IT-question-no-template but disappears in IT-completion (IT-question-no-template-vs-IT-chat contrast).
> 3. *Prompt-form-elicited* - disappears in IT-completion regardless of templates (IT-completion-vs-IT-question-no-template contrast).
> 4. *Interaction effect* - post-training-installed but only expressed under specific template / prompt-form combinations (per-cell residual structure, section 4.7).
>
> To rule out before applying the four-category vocabulary:
>
> 5. *Measurement-position artefact* - depends on the evaluation pipeline's answer-extraction position rather than on the model's underlying processing (Qwen3 thinking-mode case - Appendix D.3). Not a feature category - an evaluation-pipeline failure that masquerades as one.

A constructive burden-of-proof framing: if a finding claims "post-training installed mechanism X", the claim should survive the IT-completion control before being reported as a post-training effect. If it only appears in IT-chat captures, the more accurate phrasing is "the deployed chat regime elicits mechanism X". This is a vocabulary discipline that lets readers tell which category of claim is being made.

Recommendations for evaluation practice. Three follow from the data:

- *Benchmark reporting.* Standardise the deployment-surface description alongside the model name (prompt template, role structure, special-token handling, answer-extraction position, mode-routing assumptions). Qwen3 (Appendix D.3) is the sharpest illustration: same model, same prompts, 95.8% or 12.5% depending on whether the thinking-mode prefix is enabled.
- *Model cards.* Distinguish four behaviour levels: (a) pretrained-base; (b) post-trained no-template; (c) official chat-template; (d) downstream wrapper. "Model X does Y" is often more accurately "Model X under wrapper Z does Y".
- *Cross-family generalisation.* Report single-family findings as single-family results until cross-family decomposition under the protocol's controls confirms generalisation. The Gemma-vs-OLMo Peri-LN dichotomy plus prompt-form / template-token heterogeneity (section 4.5) is direct evidence that mechanism descriptions on one family can fail to transfer.

Scaffolding as an intervention surface, not just a confound. SmolLM2 shows the optimistic reading: PT and IT-completion are both 100% captured on M cells - only chat-mode scaffolding (specifically the prompt-form sub-effect, Appendix D.4) gives the model resistance. Scaffolding is not merely a confound to ablate out - it is a genuine intervention surface where templates, answer-extraction protocols, mode-routing policies, and interface constraints become legitimate safety- and capability-engineering surfaces alongside training-time interventions.

The model-in-context (*architecture × pretraining × post-training × tokenizer × template × prompt form × runtime policy × decoding convention*) is the unit of analysis the four-condition protocol exposes. Whether these claims

generalise beyond the misleading-assertion canvas is what the canvas-extension follow-ups in section 6.3 are positioned to test.

6. Scope and Limitations

6.1 Small-scale petri dish

The five families span 1.0B to 1.7B parameters - a deliberate methodological choice, not a resource compromise. Pandey's Appendix L shows the misleading-assertion override-circuit signature concentrated at small scale and attenuating with scale (peak mid-layer DIFF excess +127% on Gemma-2-2B-IT, +89% on Mistral-7B, 0% on Llama-3.1-70B). Per-head attribution diffuses and per-cell trajectory signatures smear at larger scale (see section 1.3). A per-cell decomposition with narrow bootstrap CIs needs sharp per-cell signatures, and ~1B is where this canvas's signature is sharpest.

The price is that behavioural claims at production scale require larger-model verification before being applied to deployed systems. The candidate accounts the protocol surfaces (chat-template-induced position bias, alignment-installed confidence amplification, family-specific scaffolding directionality) are plausibly preserved at scale because the underlying architectural primitives operate the same way, but this is a hypothesis to test, not a finding to claim. Cross-family follow-up at 7B and beyond is queued (section 6.3).

6.2 Recipe and family dependence

The empirical findings in section 4 are tied to the five families tested. The 4.3× to 6.7× scaffolding-ratio spread depending on Qwen3 inclusion, the three-near-2× / two-near-1× alignment-ratio split, the three scaffolding-direction regimes - observations on this $n = 5$ sample, not universal claims. The next ten families could expand the direction trichotomy, dissolve the alignment split, or shift the scaffolding range - we cannot predict from five which.

What we do claim generalises is the **factorisation structure**: the three axes operate at distinct lifecycle moments (pretraining, fine-tuning, inference) and can be isolated by holding two fixed while varying the third. This separability reflects how IT models are built and deployed, not which families we tested. The protocol supplies the measurement procedure regardless of per-axis magnitude.

The recipe-name label was a clean negative result. SFT + DPO is used by both Llama-3.2-1B (alignment ratio 1.87×) and SmolLM2-1.7B (0.95×) - same recipe, roughly two-fold different effect. Predicting from the model card is unreliable - the protocol supplies the measured per-family pattern.

6.3 Coverage gaps

Three specific limitations of the present study are worth flagging.

Head-level ablation in IT-completion vs IT-chat not run. The section 5.4 paired-prompt analysis surfaces a falsifiable head-level prediction: heads attending disproportionately to third-sentence-named-entity positions in IT-chat should, when ablated, simultaneously reduce both M-cell capture and NM-cell flip in families where chat-mode scaffolding is detrimental or beneficial. Running these ablations is the natural next investigation; we report the (M, NM) divergence evidence as its precursor.

Larger-scale verification and broader family coverage. The ~1B sample is methodologically appropriate (section 6.1) but does not by itself characterise production-scale IT systems. Replications at 7B and 70B on the same canvas, and more families across diverse recipes, would resolve which observations are scale-specific or sample-size-limited. Both are queued.

Cross-canvas validation not run. All findings reported here are on a single canvas (the Pandey/Halawi misleading-assertion stress test - a sycophancy/fragility canvas). The framing-shift hypothesis in section 5.7 (alignment as a deployment-conditioned phenomenon) is licensed by the pattern holding across five families on this canvas - whether it holds on instruction-following, refusal, jailbreak-robustness, tool-use, or format-adherence canvases is what cross-canvas follow-ups would test. The section 5.7 scope statement flags this as the important scope condition on the broader-implications claims.

7. Conclusion

The protocol applied across five open-weight IT model families on the Pandey/Halawi misleading-assertion canvas resolves three sources of measurement ambiguity: what the post-training weight change installs versus what the chat template elicits at inference; what is family-specific versus what generalises; what aggregate magnitudes hide versus what per-cell breakdowns surface. The findings argue against the field's default of single-family chat-only evaluation.

In our sample the chat-mode scaffolding axis is the most family-variable (CV 0.74 across the four headline families; 0.82 including Qwen3's default-thinking IT-chat) and direction-variable (three regimes - detrimental, neutral, beneficial). Alignment effects are more consistent (CV 0.30 / 0.35 with Qwen3); three of five ratios sit near 2× and the other two near 1×, in a pattern that does not follow from recipe-name labels. Architectural normalisation adds a third source of heterogeneity: nominally identical norm-placement classes (Gemma and OLMo, both Peri-LN) produce opposite-direction A' signatures. Single-family findings do not reliably transfer even across families sharing an architectural-class label, and chat-only evaluations report a combined alignment+scaffolding contribution they cannot decompose.

Methodologically the protocol is a constructive refinement. One additional capture per family adds the IT-completion condition that isolates alignment from scaffolding. The minimal-pair (M, NM) extension supplies the trajectory analysis underlying the candidate scaffolding account - Per-axis ratios come with narrow bootstrap CIs. Alignment and scaffolding ratios use standard logit-lens margins (no framework commitment) - the A' statistic and minimal-pair $\cos_G$ divergence use Curved Inference's pullback metric where a residual-stream metric is required. The cost is modest - the diagnostic value about scaffolding-confounded versus alignment-installed effects is substantial.

Full activation captures, analysis pipeline, and methodological record are released with this paper to support replication and re-examination as families and canvases accumulate. The substantive findings are specific to this $n = 5$ sample at ~1B scale on the misleading-assertion canvas - the structural separability the protocol relies on is a more general property of IT model construction and deployment, and the protocol extends cleanly to other canvases, scales, and family sets. Cross-canvas validation is the important scope condition on the broader deployment-conditioned framing (section 6.3).

References

Order cited, by first author. Type tags: [JOURNAL], [CONF], [PREPRINT], [BLOG], [TECH-REPORT].

1 - **Manson, R.** (2026) "Curved Inference: A Guide to Geometric Interpretability" *Bon View Applied Intelligence and Analytics* [JOURNAL]

2 - **Pandey, M.** (2026) "LLMs Know They're Wrong and Agree Anyway: The Shared Sycophancy-Lying Circuit" *arXiv* [PREPRINT]

3 - **Halawi, et al.** (2023) "Overthinking the Truth: Understanding how Language Models Process False Demonstrations" *arXiv / ICLR 2024* [CONF]

4 - **Genadi, et al.** (2026) "Sycophancy Hides Linearly in the Attention Heads" *arXiv* [PREPRINT]

5 - **Minder, et al.** (2025) "Overcoming Sparsity Artifacts in Crosscoders to Interpret Chat-Tuning" *arXiv / NeurIPS 2025* [CONF]

6 - **Li, M.Z., et al.** (2025) "Tracing the Representation Geometry of Language Models from Pretraining to Post-training" *arXiv* [PREPRINT]

7 - **Kissane, et al.** (2024) "SAEs (usually) Transfer Between Base and Chat Models" *AI Alignment Forum* [BLOG]

8 - **Zhao, et al.** (2025) "LLMs Encode Harmfulness and Refusal Separately" *arXiv / NeurIPS 2025* [CONF]

9 - **larry-dial** (2025) "The Curious Case of the bos_token" *LessWrong* [BLOG]

10 - **Minder, et al.** (2026) "Narrow Finetuning Leaves Clearly Readable Traces in Activation Differences" *arXiv / ICLR 2026* [CONF]

11 - **Yun, et al.** (2025) "The Price of Format: Diversity Collapse in LLMs" *arXiv / Findings of EMNLP 2025* [CONF]

12 - **Dumas, et al.** (2025) "What We Learned Trying to Diff Base and Chat Models (And Why It Matters)" *LessWrong / AI Alignment Forum* [BLOG]

13 - **Sharma, et al.** (2024) "Towards Understanding Sycophancy in Language Models" *arXiv / ICLR 2024* [CONF]

14 - **Wang, Keyu, et al.** (2025) "When Truth Is Overridden: Uncovering the Internal Origins of Sycophancy in Large Language Models" *arXiv / AAAI 2026* [CONF]

15 - **Shapira, et al.** (2026) "How RLHF Amplifies Sycophancy" *arXiv* [PREPRINT]

16 - **Manson, R.** (2025) "The '3-Process' View Of How LLMs Build 'Latent Models' In Context" *flux.robman.fyi* [BLOG]

17 - **McDougall, et al.** (2023) "Copy Suppression: Comprehensively Understanding an Attention Head" *arXiv / BlackboxNLP 2024* [CONF]

18 - **Olsson, et al.** (2022) "In-context Learning and Induction Heads" *Anthropic / transformer-circuits.pub* [TECH-REPORT]

19 - **Wang, Kevin, et al.** (2023) "Interpretability in the Wild: A Circuit for Indirect Object Identification in GPT-2 small" *arXiv / ICLR 2023* [CONF]

20 - **Li, K., et al.** (2023) "Inference-Time Intervention: Eliciting Truthful Answers from a Language Model" *arXiv / NeurIPS 2023* [CONF]

21 - **Arditi, et al.** (2024) "Refusal in Language Models Is Mediated by a Single Direction" *arXiv / NeurIPS 2024* [CONF]

22 - **Zhou, et al.** (2024) "On the Role of Attention Heads in LLM Safety" *arXiv / ICLR 2025* [CONF]

23 - **Wu, et al.** (2024) "Retrieval Head Mechanistically Explains Long-Context Factuality" *arXiv* [PREPRINT]

24 - **Kim, et al.** (2025) "Peri-LN: Revisiting Normalization Layer in the Transformer Architecture" *arXiv / ICML 2025* [CONF]

25 - **Gemma Team** (2025) "Gemma 3 Technical Report" *arXiv / Google DeepMind* [TECH-REPORT]

26 - **OLMo Team, Allen Institute for AI** (2024/2025) "2 OLMo 2 Furious" *arXiv / COLM 2025* [TECH-REPORT]

27 - **Baroni, et al.** (2025) "Transformers Don't Need LayerNorm at Inference Time" *arXiv* [PREPRINT]

28 - **Elhage, et al.** (2021) "A Mathematical Framework for Transformer Circuits" *Anthropic* [TECH-REPORT]

29 - **Heimersheim, S. & Turner, A.** (2023) "Residual stream norms grow exponentially over the forward pass" *LessWrong* [BLOG]

30 - **Sharkey, et al.** (2025) "Open Problems in Mechanistic Interpretability" *arXiv* [PREPRINT]

31 - **Bai, et al.** (2026) "The Story Is Not the Science: Execution-Grounded Evaluation of Mechanistic Interpretability Research" *arXiv* [PREPRINT]

32 - **Song, et al.** (2025) "Position: Mechanistic Interpretability Should Prioritize Feature Consistency in SAEs" *arXiv* [PREPRINT]

33 - **Li, A.J., et al.** (2025) "Evaluating Adversarial Robustness of Concept Representations in Sparse Autoencoders" *arXiv* [PREPRINT]

34 - **Gao, et al.** (2020) "The Pile: An 800GB Dataset of Diverse Text for Language Modeling" *arXiv* [PREPRINT]

35 - **Biderman, et al.** (2023) "Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling" *arXiv / ICML 2023* [CONF]

Pre-registration document (Appendix A) is released alongside this paper as an artefact-level companion to the reference list above.

Appendices

Appendix A - Pre-registration documents

This appendix is the methodological record for the empirical work reported in sections 4-5.

A.1 Locked pre-registration The locked pre-registration is released alongside this paper. It records:

- Experimental design locked before any data capture began.
- Nine hypotheses with quantitative confirmation / refinement / refutation criteria: D1, D2, D3 (decomposition); A1, A2 (architectural); H1-H4 (minimal-pair localisation, capture-status correlation, scaffolding amplification, architectural concentration).
- Pre-specified analysis pipeline.
- Honest negative-result handling.
- Amendments documenting design-choice updates and one operational-definition correction (Amendment 6, H1 criterion under structural pre-swap = 0).

A.2 Hypothesis-verdict mapping Single source of truth for which pre-registered predictions the data supports.

Table A.2.1. Pre-registered hypothesis verdicts: decomposition (D1-D3), architectural (A1-A2), and minimal-pair (H1-H4) categories.

Hypothesis	Pre-registered statement (compact)	Verdict	Notes
D1	Multiplicative factorisation of margin magnitudes at the cell level (SD log-ratio < 0.5).	REFUTED at strict criterion in all 5 families	Per-cell residual variance has interpretable structure (per-position / per-framing patterns); aggregate ratios remain tight and meaningful. The structure is what section 4.7 reports.
D2	Chat-template ratios are more recipe-consistent across families than post-training ratios.	REFUTED, qualitatively in the opposite direction	Scaffolding spans 1.08× to 4.61× (4.3× range) vs post-training 0.95× to 2.03× (~2× range) across the four headline families; CV 0.74 vs 0.30. Dispersion comparison fragile at $n = 4$ (dropping Gemma tightens scaffolding CV to ~0.37; dropping SmolLM2 tightens post-training CV to ~0.04), but directional finding (scaffolding more recipe-variable, with the longer tail on the scaffolding side) is robust to single-family dropouts. Including Qwen3 default-thinking: scaffolding 0.69× to 4.61×, CV 0.82; post-training CV 0.35. Substance of section 5.2.
D3	Captured cells flip sign in IT-completion across families showing capture in IT-chat.	family-specific	Confirmed in Gemma at Tier 3; refuted in OLMo and others.
A1	Architectural $\hat{\gamma}/x$ divergence corresponds to norm regime.	CONFIRMED across all 5 families with three empirical sub-regimes	Per section 4.3. Norm semantics subdivide further than placement.

Hypothesis	Pre-registered statement (compact)	Verdict	Notes
A2	Bulk Frobenius ΔW concentration tracks γ-peak layers.	REFUTED in Gemma (Tier 3); recipe-specific elsewhere	Vocab-projected ΔW analysis listed as follow-up.
H1	(Refined Amendment 6) Pre-swap = 0; at-swap $\geq 5\times$ post-swap-far; at-swap ≥ 0.01 absolute.	CONFIRMED universally (all 5 families $\times$ 3 conditions)	At/far ratios $10\times$ to $280\times$.
H2	Integrated post-swap divergence correlates with capture status (Cohen's $d \geq 0.5$; CIs separated).	family- and condition-specific	IT-completion confirms in 4 of 5; PT and IT-chat give mixed results.
H3	IT-chat / IT-completion divergence ratio > 1.5 averaged across cells.	REFINED in 3 of 5; REFUTED in Llama ($0.81\times$) and Qwen3 ($0.64\times$)	Scaffolding does not reliably amplify divergence on this canvas.
H4	Per-cell peak-layer / cross-layer divergence > 2.5 for Peri-LN families; < 1.5 for pre-norm-only.	CONFIRMED for Gemma only; pattern reveals norm-semantics dependence	Cross-family analysis in section 4.6.

A.3 Amendments 1-9

1. Ministral-3-3B replaced with SmolLM2-1.7B (load failure on current transformers).
2. Comparator vocabulary confirmed (all 5 pairs single-token in leading-space form across all 5 families).
3. Cell suite reuses pilot-06's 192 + 48 design.
4. Capability suite separate from minimal-pair NM.
5. Chat-template preamble suppression for SmolLM2 (smollm2_no_preamble mode, matching the locked llama_no_preamble convention).
6. H1 criterion refined: pre-swap divergence is structurally zero by the (M, NM) design, so the original at-swap / pre-swap ratio was ill-defined. Restated as at-swap $\geq 5\times$ post-swap-far + at-swap ≥ 0.01 absolute + pre-swap = 0 sanity check.
7. Paper framing locked: empirical findings as the headline; methodology as the lens; cross-family heterogeneity as the important claim.
8. IT-question-no-template fourth-condition addition: instruction-tuned model fed chat-mode question-form prompts without applying the chat template; disentangles the chat-mode-scaffolding axis into prompt-form and template-token sub-effects. New mot-capture configs at `etc/config-{family}-it-question-no-template-full.json`; `run.sh` extended with four new steps inserted as the fourth member of each existing trio; pre-flight collision check added at script top, scoped to the new condition's output paths only. Three existing conditions and their artefacts are unchanged. No pre-registered hypothesis is modified; the new condition refines section 4.5 / section 5.2 mechanism interpretation (Appendix D.4) and supplies the per-layer trajectory data needed to adjudicate the section 5.4 falsifiable position-bias prediction.
9. Qwen3 thinking-mode survival check: added a fifth Qwen3-specific condition (IT-chat-no-think) that applies the chat template with `enable_thinking=False` to suppress the `<think>...</think>` prefix; tests whether the section 4.2 / Appendix D.3 Qwen3 capability-control inversion is a thinking-mode-routing artefact or a genuine post-training capability loss. New config `etc/config-qwen-it-chat-no-think.json`; new `qwen3_no_think` chat_template_mode added to `bin/{mot-capture,logit-margin,compute-decision-margin,analyse-curvature-salience,check-chat-template}.py`. Single-forward-pass logit-margin sufficient; no full HDF5 capture required. No pre-registered hypothesis is modified; results are reported in Appendix D.3.

Appendix B - Model details

Per-family architectural and alignment-recipe details. Sources: (a) captured metadata (layer count from HDF5 keys; norm convention from `is_gemma_norm_convention` in `analyse-alignment-weights.py`; final-norm γ from `alignment-weight-deltas.json`; chat-template handling from `chat_template_mode` in `etc/config-{family}-instruct.json`); (b) each IT checkpoint's published `config.json` (cross-checked against the HF mirror for the gated families); (c) the Gemma 3 Technical Report for Gemma vocab and 1B context-length claims.

B.1 - Gemma-3-1b

- **HuggingFace identifiers.** PT: `google/gemma-3-1b-pt`. IT: `google/gemma-3-1b-it`.
- **Parameter count.** ~1.0 B.
- **Decoder layers.** 26 (27 residual-stream states including post-embedding L0).
- **Norm placement.** Peri-LN (post-sublayer normalisation alongside pre-norm; Kim et al. 2502.02732).
- **Norm parameterisation.** $(1+\gamma)$ scale with per-channel learned $\gamma \in \mathbb{R}^d$. γ_{attn} range [103, 3239] (peak L24); γ_{ff} range [124, 5673] (peak L24). Final-norm $\gamma_{\text{pt}} = 422.55$, $\Delta_{\text{IT}} = -22.8\%$ (the only family with substantial final-norm γ change post-training).
- **Attention design.** GQA - 4 query heads, 1 KV head.
- **QK normalisation.** Yes.
- **Hidden dimension.** 1152.
- **Intermediate / FFN dimension.** 6912.
- **Head dim.** 256.
- **Vocab size.** 262,144 ($= 2^{18}$; Gemma 3 Technical Report Table 1 rounds to "256k").
- **Context length.** 32,768 (the 1B is the only Gemma 3 variant with 32K; 4B/12B/27B have 128K).
- **Alignment recipe.** SFT + RLHF + knowledge distillation.
- **Chat template at capture.** `chat_template_mode: "default"` (no preamble, verified via `check-chat-template.py`).
- **Structural tokens.** `<bos><start_of_turn>user\n…<end_of_turn>\n<start_of_turn>model\n`.
- **Tokenizer notes.** All 8 name pairs and all 5 comparator pairs single-token leading-space (`verify-tokenization.py`; Appendix C).
- **Architectural-axis A' ratio (PT, median over 432 cells; section 4.3 Table 5).** $A'(x)/A'(\hat{\gamma}) \approx 90\times$ (norm amplifies the running sum); L2-norm check at the final-layer answer-start position: mean 150.7×, median 150.5×.

B.2 - Llama-3.2-1B

- **HuggingFace identifiers.** PT: `meta-llama/Llama-3.2-1B`. IT: `meta-llama/Llama-3.2-1B-Instruct`.
- **Parameter count.** ~1.2 B.
- **Decoder layers.** 16 (17 residual-stream states).
- **Norm placement.** Pre-norm only ($\hat{\gamma} = x$ by construction).
- **Norm parameterisation.** Standard RMSNorm with per-channel γ. Final-norm $\gamma_{\text{pt}} = 107.23$, $\Delta_{\text{IT}} = -0.13\%$.
- **Attention design.** GQA - 32 query heads, 8 KV heads.
- **QK normalisation.** No.
- **Hidden dimension.** 2048.
- **Intermediate / FFN dimension.** 8192.
- **Head dim.** 64.
- **Vocab size.** 128,256.
- **Context length.** 131,072.
- **Alignment recipe.** SFT + DPO.
- **Chat template at capture.** `chat_template_mode: "llama_no_preamble"`. The default injects a date / knowledge-cutoff preamble (~18 metadata tokens); we suppress it for cross-family cleanliness (pre-registration locked, Appendix A).
- **Structural tokens (no-preamble mode).** `<|begin_of_text|><|start_header_id|>user<|end_header_id|>\n\n…<|e`
- **Tokenizer notes.** All 8 name pairs + 5 comparator pairs single-token leading-space (Appendix C).
- **Architectural-axis A' ratio (PT).** $\hat{\gamma}/x = 1.000$ by construction.

B.3 - OLMo-2-1B

- **HuggingFace identifiers.** PT: `allenai/OLMo-2-0425-1B`. IT: `allenai/OLMo-2-0425-1B-Instruct`.
- **Parameter count.** ~1.5 B.
- **Decoder layers.** 16 (17 residual-stream states).
- **Norm placement.** Peri-LN (same structural class as Gemma but standard RMSNorm, not Gemma's $(1+\gamma)$).
- **Norm parameterisation.** Standard RMSNorm with per-channel γ. γ_{attn} range [3.5, 17.1] - two orders of magnitude smaller than Gemma's. Final-norm $\gamma_{\text{pt}} = 102.78$, $\Delta_{\text{IT}} = +1.5\%$.
- **Attention design.** MHA - 16 query heads, 16 KV heads.
- **QK normalisation.** Yes.
- **Hidden dimension.** 2048.
- **Intermediate / FFN dimension.** 8192.
- **Vocab size.** 100,352.
- **Context length.** 4,096 (shortest pre-training context of the five families).
- **Alignment recipe.** SFT → DPO → RLVR (separately published intermediate checkpoints).
- **Chat template at capture.** `chat_template_mode: "default"` (no preamble).
- **Structural tokens.** `<|user|>…<|assistant|>`.
- **Tokenizer notes.** All 8 name pairs + 5 comparator pairs single-token leading-space (Appendix C).
- **Architectural-axis A' ratio (PT).** $A'(\hat{\gamma})/A'(x) \approx 6.95$ (norm reduces the running sum). L2-norm check at the final-layer answer-start position: mean 0.0962, median 0.0965 - opposite direction from Gemma despite the same Peri-LN class (norm-semantics dichotomy, section 4.3).

B.4 - Qwen3-1.7B

- **HuggingFace identifiers.** PT: `Qwen/Qwen3-1.7B-Base`. IT (thinking on by default): `Qwen/Qwen3-1.7B`. IT no-think: same model with `enable_thinking=False` (Appendix D.3 thinking-mode survival check).
- **Parameter count.** ~1.7 B.
- **Decoder layers.** 28 (29 residual-stream states).
- **Norm placement.** Pre-norm only ($\hat{\gamma} = x$ by construction).
- **Norm parameterisation.** Standard RMSNorm. γ_{attn} range [11.1, 210.4]; final-norm $\gamma_{\text{pt}} = 102.88$, $\Delta_{\text{IT}} = +0.05\%$.
- **Attention design.** GQA - 16 query heads, 8 KV heads.
- **QK normalisation.** Yes.
- **Hidden dimension.** 2048.
- **Intermediate / FFN dimension.** 6144.
- **Head dim.** 128.
- **Vocab size.** 151,936.
- **Context length.** 40,960.
- **Alignment recipe.** SFT + DPO + GRPO, with thinking / non-thinking mode-switching hybrid (no analog in the other four families).
- **Chat template at capture.** `chat_template_mode: "default"` - ChatML (`<|im_start|>user\n…<|im_end|>\n<|im_st` no preamble. Enables the `<think>…</think>` prefix; `qwen3_no_think` suppresses it via `enable_thinking=False`.
- **Structural tokens (thinking on).** `<|im_start|>user\n…<|im_end|>\n<|im_start|>assistant\n<think>…</think>`.
- **Tokenizer notes.** 8 name pairs single-token leading-space (Appendix D.3 details table); 5 comparator pairs single-token (Appendix C).
- **Architectural-axis A' ratio (PT).** $\hat{\gamma}/x = 1.000$ by construction.
- **Capability-control inversion note.** IT-chat capability-cell correctness is 6/48 thinking-on (mean signed margin −2.24) vs 27/48 thinking-off (+4.51); the IT-chat inversion is substantially a thinking-mode-routing artefact (section 4.2, Appendix D.3).

B.5 - SmolLM2-1.7B

- **HuggingFace identifiers.** PT: `HuggingFaceTB/SmolLM2-1.7B`. IT: `HuggingFaceTB/SmolLM2-1.7B-Instruct`.
- **Parameter count.** ~1.7 B.
- **Decoder layers.** 24 (25 residual-stream states).
- **Norm placement.** Pre-norm only ($\hat{\gamma} = x$ by construction).
- **Norm parameterisation.** Standard RMSNorm. γ_{attn} range [3.1, 14.8]; final-norm $\gamma_{\text{pt}} = 60.79$ (smallest of the five families); $\Delta_{\text{IT}} = +0.01\%$.

- **Attention design.** MHA - 32 query heads, 32 KV heads.
- **QK normalisation.** No.
- **Hidden dimension.** 2048.
- **Intermediate / FFN dimension.** 8192.
- **Vocab size.** 49,152.
- **Context length.** 8,192.
- **Alignment recipe.** SFT + DPO via Zephyr-style training.
- **Chat template at capture.** `chat_template_mode: "smollm2_no_preamble"`. The default injects an identity preamble ("You are a helpful AI assistant named SmolLM, trained by Hugging Face"); we suppress it for cross-family cleanliness (pre-registration locked).
- **Structural tokens (no-preamble mode).** `<|im_start|>user\n…<|im_end|>\n<|im_start|>assistant\n` (ChatML, matching Qwen3 modulo the thinking prefix).
- **Tokenizer notes.** All 8 name pairs + 5 comparator pairs single-token leading-space (Appendix C).
- **Architectural-axis A' ratio (PT).** $\hat{\gamma}/x = 1.000$ by construction.

B.6 - Cross-family comparison summary The five families span useful diversity on every axis except parameter count (the 1.0B–1.7B band is intentionally narrow per the small-scale petri-dish design):

Table B.6.1. Cross-family comparison summary across architecture, alignment-recipe, and capture-pipeline axes.

Field	Gemma-3-1b	Llama-3.2-1B	OLMo-2-1B	Qwen3-1.7B	SmolLM2-1.7B
Hidden dim	1152	2048	2048	2048	2048
Layers	26	16	16	28	24
Norm regime	Peri-LN	Pre-norm	Peri-LN	Pre-norm	Pre-norm
Norm param	$(1+\gamma)$	γ	γ	γ	γ
QK-norm	Yes	No	Yes	Yes	No
Attention	GQA 4/1	GQA 32/8	MHA 16/16	GQA 16/8	MHA 32/32
Head dim	256	64	128	128	64
Vocab	262,144	128,256	100,352	151,936	49,152
Context	32,768	131,072	4,096	40,960	8,192
FFN dim	6,912	8,192	8,192	6,144	8,192
Final γ_{pt}	422.55	107.23	102.78	102.88	60.79
Final γ Δ_{IT}	-22.8%	-0.13%	$+1.5\%$	$+0.05\%$	$+0.01\%$
A' ratio (PT)	$\hat{\gamma}/x \approx 1/90$	$= 1$	$\hat{\gamma}/x \approx 6.95$	$= 1$	$= 1$
Recipe	SFT+RLHF+KD	SFT+DPO	SFT→DPO→RLVR	SFT+DPO→GRPO	SFT+DPO (Zephyr)

Three structural takeaways recur as signficiant premises in section 4 and section 5:

1. **Two norm regimes (not three as pre-registered) - Peri-LN and pre-norm only.** Captured metadata in `analyse-alignment-weights.py` puts OLMo-2 in Peri-LN with Gemma, making Peri-LN a built-in replication test rather than a single-family observation. The pre-norm-only families (Llama, Qwen3, SmolLM2) all show $A'(\hat{\gamma})/A'(x) = 1.000$ by construction.
2. **The $(1+\gamma)$ parameterisation is Gemma-3-specific.** OLMo-2 is Peri-LN structurally but uses standard RMSNorm γ. Trained γ magnitudes differ by two orders of magnitude (Gemma $\gamma_{\text{attn}} \in [103, 3239]$; OLMo $\in [3.5, 17.1]$) and the $A'(\hat{\gamma})/A'(x)$ direction is opposite (Gemma 1/90; OLMo 6.95). The section 4.3 norm-semantics dichotomy traces to this.
3. **Final-norm γ-change asymmetry is recipe-driven.** Only Gemma (SFT+RLHF+distillation) shows substantial final-norm γ change (-22.8%); the other four families' final-norm γ is essentially unchanged from PT to IT ($|\Delta| \leq 1.5\%$).

These points are not separately important for the cross-family decomposition claim - Table 3's alignment / scaffolding / combined ratios are per-cell logit margins, invariant to the architectural-axis convention. They matter for *interpreting*

single-family architectural findings: "Gemma's 90× $\hat{\gamma}/x$ ratio" is a Gemma-specific compound (γ value × norm placement × distillation recipe) that does not transfer in detail to the other Peri-LN family.

Appendix C - Tokenization-alignment validation

C.1 Policy The protocol's logsumexp aggregation over candidate-token forms requires at least one form to be single-token per candidate per cell, so the margin computation is well-defined. We enforce this with a pre-capture validation gate (`bin/verify-tokenization.py`) under the policy `--require any` - for each candidate answer in each cell, at least one of the two tokenisation forms (leading-space "mid-sentence" form `" Alice"` or no-space "sentence-start" form `"Alice"`) must be single-token; if both are multi-token, the cell is a *hard failure* and dropped. If one of the two forms is single-token but the other is multi-token, the cell is a *soft failure*: the logsumexp aggregation over the union of both forms' token IDs handles it correctly without dropping the cell.

C.2 Cross-family validation outcome **Table C.2.1.** Per-tokenizer validation outcome on the 432-cell unified prompt suite. `n_candidates` counts the 864 candidate-answer slots across the suite (2 per cell × 432). `hard fails` = cells where neither tokenisation form is single-token (these would be dropped); `soft fails` = candidate slots where one form is single-token but not the other (handled by logsumexp). `Comparators mid/all` = number of comparator-word pairs (out of 5 pairs / 10 words) whose mid-sentence (leading-space) form is single-token. All 5 tokenizers pass the policy with 0 hard fails. Generated by `bin/verify-tokenization.py --require any --comparators` from `etc/prompts-chat.json`; full report at `results/tokenisation-validation.json`.

Tokenizer	n_candidates	hard fails	soft fails	Comparators mid/all	Passed
google/gemma-3-1b-it	864	0	0	10/10	✓
meta-llama/Llama-3.2-1B-Instruct	864	0	54	10/10	✓
allenai/OLMo-2-0425-1B-Instruct	864	0	54	10/10	✓
Qwen/Qwen3-1.7B	864	0	54	10/10	✓
HuggingFaceTB/SmolLM2-1.7B-Instruct	864	0	324	10/10	✓

Cells dropped from the canonical 432-cell suite: 0 across all 5 families. No cell-level adjustments were applied; the protocol runs on the same 432-cell suite for every family.

C.3 Soft-failure interpretation Soft failures (one form single-token, the other multi-token) range from 0 (Gemma) to 324 (SmolLM2 - soft fails on the sentence-start `"Alice"`-style form for most candidates while the leading-space form is single-token). The asymmetry reflects different tokenisers' handling of word boundaries: Gemma's SentencePiece tokenizer assigns a single token to both `" Alice"` and `"Alice"`-without-space for most candidate names; SmolLM2's tokenizer typically produces a single token for `" Alice"` but multi-token for `"Alice"`.

The logsumexp aggregation handles the asymmetry correctly. For a candidate with mid-form token IDs `T_mid` and start-form token IDs `T_start`:

$\alpha_{\text{candidate}}(\ell) = \text{logsumexp}(\text{logits}[\ell, t] \text{ for } t \text{ in } T_{\text{mid}} \cup T_{\text{start}})$

When the start-form is multi-token (e.g., `[59, 368]` for `"Kate"` in SmolLM2), its individual token IDs are excluded from the sum (a multi-token form does not have a well-defined single-position logit); the mid-form's single token contributes the candidate's probability mass. The candidate margin remains well-defined.

The cross-tokenizer soft-fail-count differences (0 vs 54 vs 324) are therefore methodologically benign for our protocol - they reflect tokenizer-vocabulary differences, not protocol-level reliability differences.

C.4 Comparator-vocabulary validation The minimal-pair (M, NM) construction (section 3.3) swaps a single comparator word (e.g., `" more"` ↔ `" fewer"`) in the third-sentence assertion. For the swap to preserve the single-token property at the swap position, the comparator must be single-token in its leading-space (mid-sentence) form in every family's tokenizer.

All five locked comparator pairs encode as **single tokens in their leading-space form** (`more`, `fewer`, `taller`, `shorter`, `older`, `younger`, `bigger`, `smaller`, `heavier`, `lighter`) across all 5 tokenizers (10/10 in Table C.2.1's "Comparators mid/all" column for every family). The five-pair lock is recorded in the pre-registration (Appendix A.3).

The sentence-start form (without leading space) is multi-token for some words in some families - this varies by tokenizer and does not affect the protocol because the prompt suite always places comparator words mid-sentence, where the leading-space form is the operative one.

C.5 Cross-reference to Appendix B Per-family tokenizer notes are summarised in Appendix B (B.1–B.5 each list the "Tokenizer notes" bullet pointing back to this appendix). The Qwen3 capability-cell tokenisation sanity check in Appendix D.3 lists the token-ID lookup for all 8 name pairs.

Appendix D - Per-cell results tables

D.1 Clustered-bootstrap sensitivity check The cell-level bootstrap reported in Tables 3, 7, and 8 treats the 192 M cells per family as independent draws. The prompt suite is a fully crossed design over (4 framings × 2 positions × 3 value-gap magnitudes × 8 name pairs), so cells that share a clustering variable are not strictly independent. Below we report aggregate ratio CIs from two clustered alternatives: a **name-pair-clustered bootstrap** (8 clusters of 24 cells, resampled with replacement) and a **framing-clustered bootstrap** (4 clusters of 48 cells, resampled with replacement). Point estimates are identical across schemes; only the CI width differs. Generated by `bin/clustered-bootstrap-sensitivity.py` from `results/clustered-bootstrap-sensitivity.json`.

Table D.1.1. Post-training ratio (IT-completion / PT) per family, cell-level vs clustered 95% CIs.

Family	Point	Cell CI	Name-pair-cluster CI	Framing-cluster CI
Gemma-3-1b	2.03×	[1.82, 2.27]	[1.65, 2.51]	[1.58, 2.87]
Llama-3.2-1B	1.87×	[1.60, 2.20]	[1.40, 2.57]	[1.62, 2.12]
OLMo-2-1B	1.94×	[1.75, 2.16]	[1.56, 2.41]	[1.26, 3.05]
Qwen3-1.7B	0.99×	[0.89, 1.08]	[0.79, 1.18]	[0.73, 1.36]
SmolLM2-1.7B	0.95×	[0.89, 1.00]	[0.77, 1.14]	[0.84, 1.03]

Table D.1.2. Chat-mode scaffolding ratio (IT-chat / IT-completion) per family.

Family	Point	Cell CI	Name-pair-cluster CI	Framing-cluster CI
Gemma-3-1b	4.61×	[4.24, 5.04]	[3.99, 5.19]	[4.26, 5.25]
Llama-3.2-1B	2.05×	[1.72, 2.45]	[1.01, 3.63]	[1.68, 2.60]
OLMo-2-1B	1.08×	[0.95, 1.23]	[0.80, 1.55]	[0.94, 1.28]
Qwen3-1.7B	0.69×	[0.57, 0.83]	[0.26, 1.32]	[0.58, 0.97]
SmolLM2-1.7B	1.17×	[1.05, 1.31]	[0.67, 1.77]	[1.10, 1.26]

Table D.1.3. Combined ratio (IT-chat / PT) per family.

Family	Point	Cell CI	Name-pair-cluster CI	Framing-cluster CI
Gemma-3-1b	9.35×	[8.61, 10.22]	[7.51, 11.96]	[7.74, 12.69]
Llama-3.2-1B	3.85×	[3.22, 4.58]	[1.77, 7.26]	[3.27, 4.61]
OLMo-2-1B	2.09×	[1.83, 2.40]	[1.44, 2.89]	[1.41, 3.16]
Qwen3-1.7B	0.68×	[0.58, 0.80]	[0.28, 1.23]	[0.51, 0.90]
SmolLM2-1.7B	1.11×	[0.99, 1.23]	[0.65, 1.68]	[1.03, 1.20]

Reading. The cluster bootstrap widens the CI when within-cluster correlation is non-trivial - the name-pair cluster bootstrap widens the CI substantially in some cases (Llama scaffolding [1.01, 3.63]; Qwen3 [0.26, 1.32]; SmolLM2 [0.67, 1.77]), reflecting that within-name-pair correlation is real on this canvas. The framing-cluster bootstrap widens or shifts CIs more modestly in most rows. The qualitative conclusions of section 4 survive both clusterings: the strong-post-training cluster (Gemma, Llama, OLMo at ~1.9–2.0×) and the minimal-post-training cluster (Qwen3, SmolLM2 at ~0.95–1.0×) are intact; the 4.3× scaffolding spread across the four headline families (6.7× including Qwen3's default-thinking IT-chat regime) is preserved; the Gemma-vs-Qwen3 separation on the scaffolding axis is robust under both clusterings. Two specific significance claims are sensitive to the bootstrap choice and are flagged in the main text: Llama scaffolding's separation from 1.0 is marginal under name-pair clustering (CI lower bound 1.01); Qwen3 scaffolding's separation from 1.0 holds under cell and framing-cluster bootstraps but not under name-pair clustering (CI [0.26, 1.32] includes 1.0).

D.2 Per-cell results tables The full per-cell-per-family-per-condition values are released as JSONL artefacts alongside the paper; this subsection (a) indexes the artefact files, (b) reports per-family per-condition per-suite signed-margin distributions (5-number summary plus mean), and (c) reports per-family per-condition A' summary statistics. The aggregate ratios derived from these per-cell values are reported in Tables 3, 7, 8 (section 4) and Appendix D.1 (clustered-bootstrap sensitivity).

D.2.1 Artefact index Per-family directory: `results/{family}/` for each of `gemma3-1b`, `llama-3.2-1b`, `olmo-2-1b`, `qwen3-1.7b`, `smollm2-1.7b`.

Table D.2.1. Per-family-directory artefact index: released file types, contents, and paths.

Artefact	Contents	Path within per-family directory
Full-trajectory HDF5	Per-(prompt, layer, token) residual stream x; per-(prompt, layer) attn / mlp sublayer outputs ($\hat{\gamma}$ reconstructable as cumulative sum). One HDF5 per condition.	`{stem}-{condition}-activations.hdf5` (PT, IT-chat, IT-completion, IT-question-no-template stems per section 3.1)
Per-(prompt, layer) decision margin	$\alpha_{\text{correct}} - \alpha_{\text{misleading}}$ via logsumexp aggregation (section 3.2) per layer per prompt. One JSONL per condition.	`decision-margin-{pt,it-completion,it-question-no-template,it}.jsonl`
Per-(prompt, layer, token) curvature / salience	κ, S, A' under G on x and $\hat{\gamma}$ trajectories. Two JSONLs per condition (one per trajectory object).	`curvature-salience-{pt,it-completion,it,it-question-no-template}-{x,gamma-hat}.jsonl`
Per-(M-cell, NM-cell, layer, token) minimal-pair $\cos_G$	$1 - \cos_G(x_M, x_{\text{NM}})$ per (layer, token) per cell pair. One JSONL per condition.	`minimal-pair-{pt,it-completion,it,it-question-no-template}.jsonl`
Aggregate-decomposition analysis	Per-family bootstrap CIs on post-training / scaffolding / combined ratios; per-group breakdowns; multiplicative-fit diagnostics.	`decomposition-analysis.json`
Minimal-pair hypothesis evaluation	Per-condition H1-H4 verdicts (pre-reg Appendix A) with per-cell concentration ratios.	`minimal-pair-evaluation.json`
Architectural divergence	Per-cell $A'(x)/A'(\hat{\gamma})$ ratio summary.	`architectural-divergence.json`
Alignment-weight delta	Per-layer relative ΔW per weight type, per-layer PT / IT γ values, final-norm γ change.	`alignment-weight-deltas.json`

Cross-family rollup at `results/cross-family-rollup.json` (per-family post-training / scaffolding / combined ratios + D2 CV comparison). Clustered-bootstrap output at `results/clustered-bootstrap-sensitivity.json` (sourced for Appendix D.1 tables). Prompt-form decomposition at `results/prompt-form-decomposition.json` (sourced for Appendix D.4 tables). Architectural-norm-ratio robustness check at `results/architectural-norm-ratio-robustness.json` (sourced for section 4.3 L2 robustness check).

D.2.2 Per-family per-condition per-suite signed-margin distributions Final-layer signed margin ($\alpha_{\text{correct}} - \alpha_{\text{misleading}}$), distribution summary across the cells in each suite. Negative values mean the model picks the misleading-named entity; positive values mean it picks the data-correct entity. Capture rate (proportion negative) is reported in Tables 3, 7, 8 and section 4.2; this table reports the distributional shape.

Table D.2.2. Per-family per-condition per-suite signed-margin distributions (min / p10 / p50 / p90 / max / mean). M, NM at $n = 192$ per cell; capability at $n = 48$.

Family	Condition	Suite	n	min	p10	p50	p90	max	mean
Gemma	PT	M	192	−4.22	−3.31	−1.19	+1.18	+2.68	−1.22
Gemma	PT	NM	192	−4.23	−2.85	−0.73	+1.81	+3.41	−0.62
Gemma	PT	cap	48	−3.07	−1.73	+0.05	+2.50	+2.98	+0.27
Gemma	IT-completion	M	192	−8.51	−5.27	−0.15	+5.08	+8.12	−0.27
Gemma	IT-completion	NM	192	−6.42	−3.47	+1.86	+7.81	+9.45	+1.96
Gemma	IT-completion	cap	48	−2.86	+0.50	+2.49	+5.27	+6.04	+2.56

Family	Condition	Suite	n	min	p10	p50	p90	max	mean
Gemma	IT-q-no-t	M	192	−7.20	−4.21	+0.80	+5.70	+8.21	+0.47
Gemma	IT-q-no-t	NM	192	−4.64	−1.38	+4.84	+9.09	+12.13	+4.18
Gemma	IT-q-no-t	cap	48	−9.55	−9.15	+2.09	+10.75	+10.97	+0.85
Gemma	IT-chat	M	192	−26.37	−20.22	−17.00	−9.08	+11.56	−15.79
Gemma	IT-chat	NM	192	−24.20	−18.95	−13.09	+11.09	+18.61	−8.39
Gemma	IT-chat	cap	48	−26.57	−21.65	+2.23	+22.70	+25.07	+0.94
Llama	PT	M	192	−1.89	−0.83	+0.01	+1.07	+1.80	+0.02
Llama	PT	NM	192	−1.04	−0.29	+0.71	+2.02	+2.68	+0.77
Llama	PT	cap	48	−1.46	−1.19	+0.00	+1.48	+2.24	+0.10
Llama	IT-completion	M	192	−4.30	−1.83	+0.38	+1.71	+2.81	+0.15
Llama	IT-completion	NM	192	−0.48	+0.38	+1.24	+2.69	+3.76	+1.43
Llama	IT-completion	cap	48	−1.31	−0.64	+0.45	+1.37	+1.96	+0.39
Llama	IT-q-no-t	M	192	−3.44	−1.79	+0.06	+1.69	+2.52	−0.02
Llama	IT-q-no-t	NM	192	−1.75	−0.91	+0.52	+1.63	+2.35	+0.43
Llama	IT-q-no-t	cap	48	−2.32	−1.07	+0.36	+1.63	+2.19	+0.17
Llama	IT-chat	M	192	−4.91	−2.91	−1.05	+5.91	+7.84	−0.39
Llama	IT-chat	NM	192	−4.95	−2.96	−1.26	+6.21	+8.45	−0.23
Llama	IT-chat	cap	48	−2.47	−1.74	+0.20	+8.25	+10.22	+1.00
OLMo	PT	M	192	−5.70	−3.50	−1.48	+0.23	+1.29	−1.51
OLMo	PT	NM	192	−2.69	−0.85	+0.65	+2.58	+3.51	+0.71
OLMo	PT	cap	48	−1.38	−0.63	+1.05	+2.21	+2.85	+0.95
OLMo	IT-completion	M	192	−9.23	−6.87	−2.77	+0.61	+1.86	−2.95
OLMo	IT-completion	NM	192	−2.62	−0.49	+1.49	+3.49	+4.47	+1.45
OLMo	IT-completion	cap	48	−4.49	−3.36	+1.52	+6.06	+7.46	+1.28
OLMo	IT-q-no-t	M	192	−5.21	−3.33	−0.33	+2.16	+4.08	−0.58
OLMo	IT-q-no-t	NM	192	−4.53	−2.96	−0.07	+2.45	+3.72	−0.13
OLMo	IT-q-no-t	cap	48	−5.19	−4.71	+0.80	+5.91	+6.46	+0.31
OLMo	IT-chat	M	192	−6.42	−4.15	−2.68	+7.01	+9.74	−1.53
OLMo	IT-chat	NM	192	−5.31	−3.08	−1.75	+7.78	+11.17	−0.49
OLMo	IT-chat	cap	48	−1.94	−0.75	+1.50	+10.38	+11.60	+2.36
Qwen3	PT	M	192	−8.45	−6.09	−3.58	−0.23	+2.81	−3.37
Qwen3	PT	NM	192	+0.51	+2.54	+3.70	+5.20	+7.99	+3.79
Qwen3	PT	cap	48	−0.24	+0.31	+3.03	+5.19	+7.10	+3.07
Qwen3	IT-completion	M	192	−12.20	−7.77	−2.33	+1.51	+4.07	−2.82
Qwen3	IT-completion	NM	192	+0.68	+4.12	+6.89	+9.77	+12.41	+6.88
Qwen3	IT-completion	cap	48	−4.71	−1.72	+2.10	+3.89	+7.81	+1.69
Qwen3	IT-q-no-t	M	192	−9.36	−6.52	−2.04	+2.51	+3.70	−1.87
Qwen3	IT-q-no-t	NM	192	−8.80	−5.07	−0.25	+3.73	+5.54	−0.59
Qwen3	IT-q-no-t	cap	48	−5.09	−3.56	+1.59	+4.89	+5.94	+1.11
Qwen3	IT-chat	M	192	−7.11	−6.81	−1.90	+0.41	+1.41	−2.19
Qwen3	IT-chat	NM	192	−7.04	−6.75	−2.00	+0.28	+1.01	−2.24
Qwen3	IT-chat	cap	48	−6.55	−6.33	−1.95	+0.32	+0.75	−2.23
SmolLM2	PT	M	192	−8.81	−5.64	−3.84	−2.54	−1.31	−4.03
SmolLM2	PT	NM	192	−1.95	−0.59	+1.16	+3.23	+4.50	+1.29
SmolLM2	PT	cap	48	−1.17	−0.70	+0.88	+1.94	+2.96	+0.93
SmolLM2	IT-completion	M	192	−8.62	−5.66	−3.80	−2.24	−0.12	−3.81
SmolLM2	IT-completion	NM	192	−2.13	−0.98	+0.93	+2.87	+3.84	+0.94
SmolLM2	IT-completion	cap	48	−1.18	−0.50	+1.23	+2.87	+3.54	+1.22
SmolLM2	IT-q-no-t	M	192	−4.41	−3.49	+0.55	+4.09	+6.42	+0.50
SmolLM2	IT-q-no-t	NM	192	−3.75	−2.84	+0.95	+4.98	+7.33	+1.15
SmolLM2	IT-q-no-t	cap	48	−3.00	−2.05	+1.45	+5.84	+6.93	+1.63
SmolLM2	IT-chat	M	192	−13.82	−8.77	−1.46	+6.81	+8.51	−0.48
SmolLM2	IT-chat	NM	192	−12.09	−6.38	+0.16	+8.85	+10.79	+1.34
SmolLM2	IT-chat	cap	48	−10.12	−8.07	+1.08	+10.27	+12.09	+2.23

Notable features visible from the distribution table that aggregate ratios in section 4.1 do not surface:

- **Gemma IT-chat M cells:** min −26.37, p10 −20.22, p90 −9.08 - the entire interquartile range is below zero. The few cells that go positive (max +11.56) are far-tail outliers. The model is uniformly captured with very high confidence (the 4.61× magnitude scaffolding ratio understates this - the IT-chat distribution is shifted *and* compressed below zero).
- **SmolLM2 PT M cells:** min −8.81, max −1.31 - every single cell is captured; the distribution is entirely below zero. This is the cleanest illustration of the PT-to-IT-completion-installs-nothing pattern in section 4.2's SmolLM2 narrative: no positive M-cell margin exists in PT at all.
- **Qwen3 PT NM cells:** min +0.51, p10 +2.54 - every cell is correct, with comfortable margin. Qwen3 PT is the most-correct-on-NM family in the sample. The drop to IT-question-no-template (min −8.80, p10 −5.07) and IT-chat (mean −2.24) is the prompt-form-driven NM-flip that section 5.4 develops.
- **OLMo IT-chat NM cells:** p10 −3.08, p50 −1.75, p90 +7.78 - the distribution is bimodal-ish, with a left lobe of strongly captured cells and a right lobe of strongly correct cells. The 1.62× NM scaffolding ratio in section 4.5 averages these.

D.2.3 Per-family per-condition A' summary (under G, on x and $\hat{\gamma}$ trajectories) A' (semantic surface area; CI03 statistic, integrated over (token, layer) under G) per cell, summarised across the 432 cells per family per condition. Reported at the $\alpha = 0.5$ regularisation value (the CI canonical default; this is the curvature-weighting coefficient that the CI guide denotes γ, renamed here to α to avoid collision with the architectural γ used throughout this paper). One value per (family, condition, trajectory_object) combination.

Table D.2.3. Mean A' per family per condition per trajectory_object across 432 cells.

Family	Trajectory	PT	IT-completion	IT-q-no-t	IT-chat
Gemma	x	4.26×10^7	1.85×10^7	1.72×10^7	2.53×10^7
Gemma	$\hat{\gamma}$	4.76×10^5	5.66×10^5	5.24×10^5	7.02×10^5
Llama	x	4.47×10^4	3.82×10^4	3.47×10^4	4.76×10^4
Llama	$\hat{\gamma}$	4.47×10^4	3.82×10^4	3.47×10^4	4.76×10^4
OLMo	x	1.32×10^5	5.58×10^4	4.79×10^4	7.29×10^4
OLMo	$\hat{\gamma}$	9.12×10^5	1.13×10^5	1.03×10^5	1.68×10^5
Qwen3	x	3.99×10^6	2.98×10^6	2.65×10^6	3.56×10^6
Qwen3	$\hat{\gamma}$	3.99×10^6	2.98×10^6	2.65×10^6	3.56×10^6
SmolLM2	x	5.36×10^6	5.19×10^6	4.75×10^6	5.68×10^6
SmolLM2	$\hat{\gamma}$	5.36×10^6	5.19×10^6	4.75×10^6	5.68×10^6

Two patterns visible. (a) **Pre-norm-only architectures** (Llama, Qwen3, SmolLM2) have $A'(x) = A'(\hat{\gamma})$ at every condition - the $\hat{\gamma} = x$ identity holds at runtime (section 4.3). (b) **Peri-LN architectures** (Gemma, OLMo) show consistent direction across conditions: Gemma $A'(x) \gg A'(\hat{\gamma})$ ($(1+\gamma)$ amplifies); OLMo $A'(\hat{\gamma}) > A'(x)$ at PT (norm reduces). The PT $A'(\hat{\gamma})/A'(x)$ values (Gemma: 0.011, equivalent to ~90× amplification; OLMo: 6.95, equivalent to ~1/7 reduction) are what Table 5 reports as the architectural-axis signature.

The within-family across-condition variation in A' (e.g., Gemma $A'(x)$: PT 4.26×10^7 → IT-completion 1.85×10^7 → IT-chat 2.53×10^7) is one specific consequence of the post-training γ changes (Gemma's γ values reduced 24–52% at deep layers after post-training; Appendix D.6) and reflects how each condition's residual-stream trajectory shape changes - not how the architectural-axis effect direction changes (the direction is preserved across conditions; section 4.3 closing).

D.3 Qwen3 capability-control inversion - validation The section 4.2 case study reports a striking observation: on the 48 capability-control cells (bare arithmetic comparison + question, no third-sentence assertion), Qwen3-1.7B's pretrained checkpoint answers correctly on 46 of 48 cells while the same model run in chat mode answers correctly on only 6 of 48. This subsection reports the underlying per-cell data, a worked example, and a tokenisation sanity check, then offers the most plausible mechanistic reading. The data is computed from `results/qwen3-1.7b/decision-margin-{pt,it-completion,it}.jsonl`.

Per-condition aggregate on the 48 capability cells:

Table D.3.1. Qwen3-1.7B per-condition aggregate on the 48 capability-control cells: correct vs wrong counts and mean signed margin.

Condition	n correct (margin > 0)	n wrong (margin < 0)	Mean signed margin	Median
PT	46 / 48	2 / 48	+3.068	+3.032
IT-completion	41 / 48	7 / 48	+1.694	+2.100
IT-chat	**6 / 48**	**42 / 48**	**−2.235**	−1.955

The mean signed margin flips sign between PT and IT-chat (a 5.30-unit shift), and the proportion answering correctly drops from 95.8% to 12.5%. IT-completion sits between the two (85.4% correct), so the bulk of the inversion happens at the chat-template application step rather than at post-training.

Worked example. Cell `cap-high_first-large-alice-bob` (one of the 48 capability cells):

Table D.3.2. Worked example for capability cell `cap-high_first-large-alice-bob` across conditions: formatted prompt, argmax token, and margin.

Condition	Formatted prompt	Final-layer argmax token	Margin ($\alpha_{\text{correct}} - \alpha_{\text{misleading}}$)
PT	`Alice has 88 books. Bob has 15 books. Therefore, the one with more books is`	`Alice` (logit 32.83)	+2.290
IT-completion	(same as PT - no chat template)	`Alice` (logit 23.41)	+2.624
IT-chat	`<|im_start|>user\nAlice has 88 books. Bob has 15 books. Who has more books?<|im_end|>\n<|im_start|>assistant\n`	`<think>` (logit 30.14)	−1.818

The PT and IT-completion final-layer argmax at the answer-start position is the correct answer (`Alice`). In IT-chat, the final-layer argmax at the answer-start position is the special `<think>` thinking-mode token, not either candidate. The margin metric - $\alpha_{\text{Alice}} - \alpha_{\text{Bob}}$ measured at this position - still flips sign because the residual stream at the answer-start position is configured for thinking-mode emission, not for direct-answer emission, and within that configuration Bob's α happens to exceed Alice's.

Mechanistic reading. Qwen3's published recipe includes a "thinking / non-thinking" mode-switching capability with no analog in the other four families. The IT-chat model, when prompted via the chat template, defaults into thinking-mode and emits `<think>` first; the model's answer-relevant computation occurs in the chain-of-thought that follows. The α-margin we measure at the immediate answer-start position therefore captures a Qwen3-specific routing artefact rather than a clean "the model has lost the capability". A more accurate framing of the Qwen3 capability-control finding is: chat-mode application routes Qwen3 into a thinking-mode regime in which the bare-task answer is not emitted at the immediate post-prompt position, so the margin metric reads the wrong position. Whether the thinking-mode continuation eventually produces the correct answer is a separate measurement we do not run here (capturing generated text would require continued sampling beyond the answer-start position).

This is the most likely-correct mechanistic reading; the headline empirical observation - that the same Qwen3 weights in chat mode produce sign-flipped margins at the answer-start position on a task PT solves cleanly - is correct on its own terms. The methodological implication is the same as for the post-training / scaffolding finding: chat-only evaluation that measures at fixed answer-start positions can mis-report Qwen3's capability without flagging that the model's deployment-mode routing differs from the base model's. Researchers benchmarking Qwen3-family models specifically should be aware of the thinking-mode prefix in their capture pipeline.

Tokenization sanity check. Every capability-control candidate-answer name in Qwen3's tokenizer has at least one single-token form. Token IDs (showing leading-space and no-leading-space forms where both exist; logsumexp aggregation handles multi-form correctly):

Table D.3.3. Qwen3-1.7B candidate-answer token IDs per name pair (single-token verification).

Name pair	Token IDs
Alice / Bob	[29405, 61686] / [14261, 32388]
Anna / Paul	[23223, 56756] / [6898, 25300]
Kate / Sam	[29201, 79369] / [8224, 23950]
Mary / John	[10244, 41484] / [3757, 13079]
Max / Ada	[7487, 5974] / [51149, 95347]
Mike / Lisa	[11268, 34441] / [28556, 72749]
Tim / Joe	[9354, 20217] / [12846, 40344]
Tom / Sue	[8364, 24732] / [47649]

All eight name pairs pass the verify-tokenization.py single-token requirement. The capability-control inversion is therefore not a tokenisation artefact.

Thinking-mode survival check.

Qwen3's chat template natively supports an `enable_thinking=False` parameter that suppresses the `<think>...</think>` prefix. A fifth Qwen3-specific condition - **IT-chat-no-think** - applies the chat template to the question-form prompt with thinking-mode suppressed. This adjudicates whether the inversion above is a thinking-mode-routing artefact (in which case IT-chat-no-think should look closer to IT-completion or PT) or a genuine post-training capability loss that survives the routing change (in which case IT-chat-no-think should look closer to IT-chat).

Table D.3.4. Qwen3-1.7B per-condition correctness across the three cell suites (thinking-mode survival check). Generated from `results/qwen3-1.7b/qwen3-1.7b-it-chat-no-think_logit_margin.jsonl` (logit-margin output) and the existing decision-margin JSONLs.

Condition	Capability correct / 48 (mean signed margin)	M correct / 192 (mean)	NM correct / 192 (mean)
PT	46/48; +3.07	17/192; −3.37	192/192; +3.79
IT-completion	41/48; +1.69	43/192; −2.82	192/192; +6.88
IT-question-no-template	27/48; +1.11	67/192; −1.87	93/192; −0.59
IT-chat (thinking ON)	**6/48; −2.24**	43/192; −2.19	37/192; −2.24
IT-chat-no-think	**27/48; +4.51**	**87/192; −3.23**	**106/192; −0.72**

Reading: the IT-chat capability-control inversion is largely a thinking-mode-routing artefact. With thinking-mode suppressed, Qwen3 IT-chat capability-cell correctness recovers from 12.5% (6/48) to 56.2% (27/48), and the mean signed margin flips from −2.24 to **+4.51** (more positive than PT's +3.07; median +8.98 - the model is highly confident on the cells it gets right). The capability does not recover all the way to PT (95.8%) or IT-completion (85.4%); the remaining ~40-point drop matches IT-question-no-template (56.2% / +1.11), consistent with a prompt-form effect at the question-form step that is independent of either thinking mode or template tokens.

The section 4.2 attribution stands but with refined emphasis: chat-only evaluation that probes at fixed answer-start positions can be substantially confounded by dynamic routing architectures (the thinking-mode prefix in Qwen3's case), and the protocol's IT-chat number for Qwen3 capability cells understates the model's actual capability by ~44 percentage points purely from the thinking-mode-routing prefix. The right framing is **"fixed-position answer extraction is systematically blind to dynamic routing architectures"** rather than "Qwen3 has lost a capability under chat-mode application". This is a systemic warning for automated evaluation pipelines that probe models at fixed logit offsets without accounting for mode-routing or chain-of-thought prefixes.

Secondary observations from the survival check:

- The same thinking-mode-routing artefact substantially affects M-cell correctness too: with thinking off, M-cell correctness doubles from 22.4% (43/192) to 45.3% (87/192). The cells that remain wrong have more negative mean signed margin (−3.23 vs −2.19) - the model is more confidently wrong on those - but it gets twice as many right.

- NM-cell correctness with no-think (106/192, 55.2%) is *higher* than the IT-question-no-template baseline (93/192, 48.4%) - meaning the chat-template tokens *without* the thinking-mode prefix slightly improve NM correctness on Qwen3. This is the opposite direction from the with-thinking IT-chat result (37/192, 19.3%). On Qwen3 specifically, the chat-template tokens themselves do not induce the NM-flip; the thinking-mode prefix does most of the work.
- The section 5.4 family-specific NM-flip adjudication that attributed Qwen3's NM-flip to the prompt-form step (PT 192 → IT-comp 192 → IT-q-no-t 93) is unchanged: that 99-cell drop happens at the prompt-form step, before any chat template or thinking mode. The chat-template-with-thinking step then drives the further drop from 93 → 37; the chat-template-without-thinking step would have improved from 93 → 106. The section 5.4 attribution "prompt-form-driven for Qwen3" is correct for the IT-completion → IT-question-no-template step. The thinking-mode-with-chat-template step is a separate Qwen3-specific compound effect on top of that.

D.4 IT-question-no-template extension - prompt-form vs template-token decomposition The chat-mode scaffolding axis measured in section 4.5 (Table 3) conflates two sub-effects by the design described in section 1.2: the prompt-form shift from a completion seed ("…The person with more books is") to a question form ("Who has more books?"), and the application of template-tokens (role markers, turn delimiters, BOS additions). This subsection reports the decomposition produced by adding a fourth measurement condition - **IT-question-no-template** - in which the instruction-tuned model is fed the question-form prompt *without* the chat template wrapping (`prompt_mode='completion'` with the question-form prompt text). The decomposition runs `bin/logit-margin.py` on the five-family suite via the configs at `etc/config-{gemma,llama,olmo,qwen,smollm2}-it-question-no-template.json` and `bin/analyse-prompt-form-decomposition.py`; raw data at `results/prompt-form-decomposition.json`.

The decomposition is:

prompt-form ratio = |margin(IT-question-no-template)|/|margin(IT-completion)| template-token ratio = |margin(IT-chat)|/|margin(IT-question-no-template)|

Their product equals the chat-mode scaffolding ratio (|IT-chat|/|IT-completion|) reported in Table 3.

Table D.4.1. Per-family decomposition of the chat-mode scaffolding ratio into prompt-form and template-token sub-effects on the 192 M cells. Ratios are bootstrap point estimates with 95% CIs (cell-level, 10,000 iterations).

Family	Prompt-form ratio	Template-token ratio	Implied product	Scaffolding (Table 3)
Gemma-3-1b	0.95× [0.85, 1.07]	**4.83× [4.42, 5.30]**	4.61×	4.61× [4.24, 5.04]
Llama-3.2-1B	0.94× [0.83, 1.08]	**2.18× [1.86, 2.55]**	2.05×	2.05× [1.72, 2.45]
OLMo-2-1B	**0.52× [0.45, 0.59]**	**2.09× [1.80, 2.44]**	1.08×	1.08× [0.95, 1.23]
Qwen3-1.7B	0.93× [0.79, 1.12]	**0.74× [0.62, 0.89]**	0.69×	0.69× [0.57, 0.83]
SmolLM2-1.7B	**0.65× [0.58, 0.72]**	**1.81× [1.60, 2.04]**	1.17×	1.17× [1.05, 1.31]

The implied product matches the independently-computed scaffolding ratio to three decimal places across all five families - a sanity check on the multiplicative decomposition.

Reading. The five families partition cleanly into two patterns.

- **Template-token-dominant (Gemma, Llama, Qwen3).** Prompt-form CIs include 1.0; the entire chat-mode scaffolding effect on margin magnitude lives in the template-token sub-effect. Gemma's 4.61× scaffolding is essentially 4.83× template-tokens; Llama's 2.05× is 2.18×; Qwen3's 0.69× is 0.74×. For these three families the section 4.5 chat-mode-scaffolding interpretation can be sharpened to "chat-template-tokens" without loss.
- **Two-opposing-effects (OLMo, SmolLM2).** Prompt-form is substantially below 1.0 (OLMo 0.52×, SmolLM2 0.65×) and template-tokens substantially above 1.0 (OLMo 2.09×, SmolLM2 1.81×). The two sub-effects partially or fully cancel: OLMo's "neutral" 1.08× scaffolding (section 4.5) is not nothing happening - it is two large opposing effects netting near zero. SmolLM2's "beneficial" 1.17× similarly hides the same two-step structure. The "neutral scaffolding" finding therefore comes in two qualitatively different flavours: *no-effect-at-all* (Qwen3, where both sub-effects are near 1.0) and *two-effects-cancelling* (OLMo).

Sign-flip diagnostics across the four conditions on M cells further sharpen the picture:

Table D.4.2. Per-family sign-flip diagnostics on M cells across IT-completion, IT-question-no-template, and IT-chat: correct/wrong counts and mean signed margin.

Family	IT-completion (corr/wrong; mean)	IT-question-no-template (corr/wrong; mean)	IT-chat (corr/wrong; mean)
Gemma-3-1b	93 / 99; −0.27	99 / 93; +0.47	2 / 190; **−15.79**
Llama-3.2-1B	122 / 70; +0.15	100 / 92; −0.02	48 / 144; −0.39
OLMo-2-1B	33 / 159; −2.95	83 / 109; −0.58	24 / 168; −1.53
Qwen3-1.7B	43 / 149; −2.82	67 / 125; −1.87	43 / 149; −2.19
SmolLM2-1.7B	**0 / 192; −3.81**	**106 / 86; +0.50**	75 / 117; −0.48

The SmolLM2 row is the cleanest demonstration: 0 of 192 cells correct in IT-completion (the instruction-tuned model installs no resistance to misleading prompts in completion mode); the same model fed the question-form prompt *without* the chat template gets 106 of 192 correct on average (mean signed margin flips from −3.81 to +0.50); applying the chat template on top of that reduces correct-cell count to 75 of 192 (mean signed margin back to −0.48). The section 5.2 "scaffolding installs SmolLM2's deployed resistance" claim is correct at the aggregate scaffolding-ratio level, but the underlying mechanism is the *prompt-form* shift (asking "who has more?" vs completing "the one with more is"), not the template-token wrapping. The template tokens in SmolLM2 actively *undo* part of the gain.

Gemma's IT-chat row tells the opposite story at high magnitude: applying the chat template tokens to a Gemma model that was nearly balanced on the question-form prompt alone (99 / 93 correct/wrong) collapses 190 of 192 cells into the misleading-answer (mean signed margin −15.79). This is a strong instance of the "chat-template-tokens are the dominant intervention" pattern that section 5.2 attributed to the post-training-vs-scaffolding decomposition.

Implications for the main-text claims.

- section 1.1 and section 4.5 chat-mode scaffolding direction trichotomy (detrimental / neutral / beneficial) is preserved at the aggregate magnitude level. The refinement is that the "neutral" category subdivides into two mechanisms - *no-effect-at-all* (Qwen3, both sub-effects ∼1×) and *two-effects-cancelling* (OLMo, prompt-form 0.52× × template-tokens 2.09× ≈ 1×). Single-canvas single-aggregate evaluations would conflate these.
- section 5.2 SmolLM2 narrative ("post-training installs no resistance; chat-mode scaffolding installs all of it") is correct at the magnitude level. The mechanism inside the scaffolding axis is *prompt-form-driven*, not template-token-driven. The template tokens in SmolLM2 partially work against the prompt-form gain.
- section 4.5 Gemma case (4.61× scaffolding ratio, detrimental direction) is mechanistically clean - template-tokens carry the entire effect, prompt-form is near 1.0×. Single-family Gemma studies attributing this to "the chat scaffold's override pattern" are working with the right object.
- section 4.2 / Appendix D.3 Qwen3 capability-control inversion now has a sharper attribution: capability cells go from 41/48 correct in IT-completion to 27/48 in IT-question-no-template to 6/48 in IT-chat. About 30% of the capability loss happens at the prompt-form step (likely some question-form-specific routing); the dominant 70% happens at the template-application step (the thinking-mode prefix described in D.3).

The two-opposing-effects pattern (OLMo, SmolLM2) is the most informative new finding. It says that for some families the chat-mode scaffolding axis is not a single intervention but two large interventions in opposite directions; reporting just the aggregate ratio (whether tight under bootstrap or not) discards the sub-axis structure. Future cross-family extensions of the protocol should report both sub-ratios per family rather than just the combined scaffolding ratio.

D.5 Capability-controls - cross-family summary The 48 capability-control cells per family contain just the arithmetic comparison and the question, with no third-sentence misleading assertion. They serve two purposes: (a)

verify the bare task is solvable in each model + condition; (b) surface any condition-dependent capability degradation beyond Qwen3's (already documented in D.3). Per-condition correct-cell counts and mean signed margins per family:

Table D.5.1. Per-family per-condition capability-control correctness (correct-cell-count / 48) and mean signed margins.

Family	PT	IT-completion	IT-question-no-template	IT-chat
Gemma-3-1b	24/48 (+0.27)	**44/48 (+2.56)**	24/48 (+0.85)	24/48 (+0.94)
Llama-3.2-1B	24/48 (+0.10)	35/48 (+0.39)	27/48 (+0.17)	26/48 (+1.00)
OLMo-2-1B	41/48 (+0.95)	29/48 (+1.28)	27/48 (+0.31)	38/48 (+2.36)
Qwen3-1.7B	**46/48 (+3.07)**	41/48 (+1.69)	27/48 (+1.11)	**6/48 (−2.24)**
SmolLM2-1.7B	42/48 (+0.93)	38/48 (+1.22)	30/48 (+1.62)	31/48 (+2.23)

(Format: correct-cell-count / 48 (mean signed margin); bold = peak correctness per family, except for Qwen3-IT-chat which is bolded to flag the inversion.)

Bare task is solvable across families and conditions in the qualitative sense - at least one condition per family puts the model above chance (24/48). Qwen3 IT-chat is the single condition × family combination where the model is below chance with sign-flipped mean signed margin; this is the inversion D.3 develops.

Condition-dependent capability variation is not confined to Qwen3. Gemma's capability correctness peaks at IT-completion (44/48, +2.56) and drops back to chance (24/48) at both IT-question-no-template and IT-chat - a 20-cell drop at the prompt-form step, with no further drop from template tokens. The mechanism is unclear; one plausible reading is that Gemma's post-training weight changes install a strong completion-seed-formatted answer-extraction circuit that the question-form prompt does not engage. OLMo shows the opposite pattern: capability is highest in PT (41/48) and IT-chat (38/48), with IT-completion (29/48) and IT-question-no-template (27/48) dropping ~25 percentage points. SmolLM2 is the most stable family across conditions (range 30–42 of 48). Llama is mid-range with modest variation.

These per-family capability-variation patterns *qualify some interpretation of the M-cell results.* Two specific cases:

- Gemma's M-cell "PT-to-IT-completion weight change installs resistance" (PT 78.1% → IT-completion 51.6% capture rate, a 26-point M-cell improvement) is contemporaneous with a capability gain on the same prompt-form (PT 50% → IT-completion 91.7%, a 42-point capability improvement). Both observations are well-supported; the M-cell improvement could include a general capability-improvement contribution that the protocol's M-cell aggregate doesn't isolate from the misleading-assertion-resistance contribution. The two effects are confounded in the M-cell aggregate.
- OLMo's M-cell "flat across conditions" (PT 82.3% → IT-completion 82.8% → IT-chat 87.5% capture rate) is contemporaneous with a capability-degradation at the PT-to-IT-completion step (85.4% → 60.4%) that partially recovers at IT-chat (79.2%). The flat M-cell aggregate may be hiding two compensating effects: a post-training capability loss that would tend to increase M-cell capture, offset by a post-training misleading-assertion-resistance gain that tends to decrease M-cell capture.

The full per-family per-condition capability data is in `results/{family}/decision-margin-{condition}.jsonl` (suite=capability records). The pattern reported here is from the 48 cells per family per condition.

D.6 ΔW-Frobenius analysis - per-family weight-change concentration This subsection reports the per-family ΔW magnitudes and the Spearman correlation between the post-attention-layernorm γ at each layer and the relative ΔW magnitude at the same layer. Computed from `results/{family}/alignment-weight-deltas.json` (output of `bin/analyse-alignment-weights.py`).

Table D.6.1. Per-family per-layer relative $\Delta W = \|W_{\mathrm{IT}} - W_{\mathrm{PT}}\|_F / \|W_{\mathrm{PT}}\|_F$, averaged across the four attention weight types (q_proj, k_proj, v_proj, o_proj) and three MLP weight types (gate_proj, up_proj, down_proj). Range reported across decoder layers.

Family	n layers	ΔW_{attn} (rel., mean ± range)	ΔW_{mlp} (rel., mean ± range)
Gemma-3-1b	26	1.44 [1.36, 1.64]	1.46 [1.34, 1.64]
Llama-3.2-1B	16	0.14 [0.12, 0.15]	0.17 [0.13, 0.18]
OLMo-2-1B	16	1.11 [1.09, 1.12]	1.13 [1.11, 1.15]
Qwen3-1.7B	28	0.09 [0.06, 0.11]	0.10 [0.07, 0.12]
SmolLM2-1.7B	24	0.06 [0.04, 0.07]	0.06 [0.04, 0.07]

The cross-family spread in ΔW magnitude is ~24× (SmolLM2's 0.06× to Gemma's 1.44×). Gemma and OLMo have ΔW magnitudes greater than 1.0 - i.e., the IT weights at those layers differ from the PT weights by more than the original PT weight norm - consistent with their published distillation-and-multi-stage recipes (Gemma RLHF + distillation; OLMo SFT → DPO → RLVR). Llama, Qwen3, and SmolLM2 have ΔW magnitudes in the 0.06–0.17 range, an order of magnitude smaller. Within each family the ΔW magnitudes are nearly uniform across layers (small ranges in Table D.6.1).

Table D.6.2. Per-family Spearman correlation between post-attention-layernorm γ magnitude (PT) and ΔW_{attn} relative magnitude across decoder layers. Range of PT γ values reported alongside for context.

Family	Spearman $\rho(\gamma_{\text{attn}}, \Delta W_{\text{attn}})$	γ_{attn} range across layers
Gemma-3-1b	−0.12	[103.0, 3239.5]
Llama-3.2-1B	+0.45	[9.6, 23.2]
OLMo-2-1B	+0.49	[3.5, 17.1]
Qwen3-1.7B	+0.66	[11.1, 210.4]
SmolLM2-1.7B	+0.89	[3.1, 14.8]

Reading: for Gemma, the correlation is near zero (slightly negative): the layers with the largest γ values are NOT the layers with the largest ΔW magnitudes. For the other four families, the correlation is positive (0.45 to 0.89): post-training weight changes concentrate at the layers with the largest γ values. The "architectural amplification of distributed parameter changes" framing fits Gemma; the other four families show parameter-change concentration at γ-peak layers, so the architectural amplification and parameter-change concentration are reinforcing rather than substituting.

Final-norm γ change (PT → IT, relative):

Table D.6.3. Per-family final-norm γ change from PT to IT (absolute values and relative delta).

Family	PT γ	IT γ	Relative Δ
Gemma-3-1b	422.55	326.07	−22.8%
Llama-3.2-1B	107.23	107.10	−0.13%
OLMo-2-1B	102.78	104.34	+1.5%
Qwen3-1.7B	102.88	102.93	+0.05%
SmolLM2-1.7B	60.79	60.80	+0.01%

Only Gemma shows a substantial final-norm γ change (−22.8%); the other four families' final-norm γ is essentially unchanged from PT to IT. Recipes that include distillation (Gemma's) touch the post-sublayer normalisation magnitudes; recipes that don't (the other four) preserve them.

D.7 Architectural-axis A' ratio - physical intuition and γ-arithmetic The Gemma $A'(x) \approx 90 \times A'(\hat{\gamma})$ result reported in section 4.3 Table 5 is large enough to warrant a derivation of where the magnitude comes from. This subsection lays out the γ-arithmetic, the additive-not-multiplicative accumulation across depth, and the super-linear-but-sub-quadratic regime the empirical 90× sits in.

γ-arithmetic. Gemma's post-sublayer norm applies $(1 + \gamma)$ scaling element-wise to the normalised sublayer output before it is added to the residual stream. The trained γ vectors are substantial across all layers and peak at deep layers (γ_{attn} ranges across the 26 layers from 103 to 3239 in PT, peaking at L24; γ_{ff} ranges from 124 to 5673 with the same L24 peak, per Appendix B.1). Each peak-layer sublayer contribution is therefore added to the running sum with an effective per-element scale of order γ_{peak} - substantially larger than the unit-scale baseline a pre-norm-only configuration produces.

Additive across layers, not multiplicative. The residual stream is a sum of scaled sublayer outputs, not a product. The $(1 + \gamma)$ scaling enlarges each per-layer term in the sum; it does not compose multiplicatively across depth in the way that a stack of matrix multiplications would. A naive "$\gamma \approx 50$ per layer × 26 layers" multiplicative intuition would predict a much larger ratio than the empirical 90× and would be the wrong arithmetic for the operation. The correct picture is that each peak-layer step into the residual stream is order-γ_{peak} larger than its pre-norm-only counterpart, and the integral over the resulting trajectory accumulates this layer by layer.

A' dependence on salience and trajectory length. A' is a surface-area integral over (token × layer) under G; its local integrand depends on per-layer salience ($\|dx/d\ell\|_G$) and curvature, and the total trajectory length itself accumulates from per-layer salience. Two consequences for the γ-scaling:

- Local surface-area contribution depends on salience^2 (a quadratic-in-velocity term in the surface-area integrand), so a per-layer step that is $k\times$ larger contributes roughly k^2 to the local integrand at that layer.
- Trajectory length scales (approximately) linearly with per-layer salience over the layers traversed, so a per-layer step $k\times$ larger produces a trajectory whose total length is $k\times$ longer than in the unscaled case.

Together, larger per-layer steps drive super-linearly larger A' through both the salience^2 contribution and the longer trajectory length they produce.

Where the empirical 90× sits. A pure quadratic-in-γ regime (uniform $\gamma \approx$ peak across all 26 layers, plus salience^2 fully dominating) would predict ratios in the high hundreds to low thousands - much larger than 90×. The actual sub-quadratic value reflects (a) the γ variation across layers (103-3239 for γ_{attn}; not a uniform peak), (b) the per-trajectory curvature distribution, and (c) the integration over (token × layer) rather than just the worst-case scale factor. The empirical 90× is therefore consistent with the super-linear-but-sub-quadratic regime the γ-arithmetic + A' structure predict qualitatively, but the specific magnitude is not determined in closed form by the γ values alone.

What the data supports firmly. The structural directional claim - amplification not reduction; Gemma's $(1 + \gamma)$ parameterisation vs OLMo's standard-RMSNorm parameterisation producing opposite-direction signatures within the same Peri-LN structural class - is firm from both A' and the metric-free L2 readout in section 4.3 (Gemma L2 ratio 150.7×, same direction as the A' ratio with larger magnitude). The specific 90× is an empirical observation from the captured trajectories, not a number that closed-form γ-arithmetic entails. This connects to the broader observation in the literature that residual-stream norms grow over the forward pass in transformer decoders [25] (cited in section 2.4); Gemma's $(1 + \gamma)$ parameterisation is one specific mechanism through which the growth occurs.

Appendix E - Replication and reproducibility

What is needed to reproduce the paper's empirical findings end-to-end from raw models to final tables and figures, given the released HDF5 / JSONL artefacts. All per-(prompt, layer, token) activations, per-cell metric values, and per-family aggregate analyses are released with the paper.

E.1 Software environment Canonical environment: Python 3.10 with the pinned dependency set in `requirements.txt` at the repository root. Minimum-version constraints that matter for reproducibility:

- `torch>=2.4` - `forward_pre_hook` semantics relied on by `bin/mot-capture.py`.
- `transformers>=4.50` - Gemma-3 architecture support. Note: `transformers>=5.0` introduces a Gemma-3 double-embed-scale bug; the released analysis uses 4.50.x, unaffected.
- `accelerate>=0.27` - `device_map="auto"` in `mot-capture.py`.
- `h5py>=3.10`, `numpy>=1.24`, `pandas>=2.0`, `sentencepiece>=0.2` (Gemma tokenizer), `protobuf>=4.21` (transitive).

Two install paths: `pip install -r requirements.txt` into any Python 3.10 venv, or `conda env create -f environment.yml` then `conda activate mot` (the YAML delegates to `requirements.txt` via pip to avoid conda/pip channel conflicts on `torch`/`transformers`).

GPU/CUDA support is decoupled: install the CUDA-matched `torch` wheel before the rest (see `environment.yml` comments). CPU-only inference reproduces every result at 1B–1.7B scale, but wall-clock varies by 10–100× across hardware.

E.2 Hardware Captures and analyses ran on an Apple M1 (16GB unified memory) with MPS-backed `torch` inference; no CUDA-specific code paths. The same scripts run on CUDA or CPU unchanged. Approximate M1 wall-clock per family per condition (432-cell suite):

- `mot-capture` (full HDF5 capture, 432 cells, 26 layers): 8–15 min per condition for 1B families; 12–20 min for 1.7B families.
- `analyse-curvature-salience` (per-(prompt, layer, token) κ, S, A' under G, both trajectories): 3–8 min per condition.
- `analyse-minimal-pair` (per-(pair, layer, token) $\cos_G$): 1–3 min per condition.
- `analyse-decomposition` + per-cell bootstraps (10k iterations): 1–2 min per family.

End-to-end per family across four conditions (PT, IT-completion, IT-question-no-template, IT-chat) with all analyses: ~1.5–3 hours M1 wall-clock; ~3–6 GB total artefact volume (HDF5 dominates).

E.3 End-to-end recipe Canonical orchestration: `run.sh <model> <run_number> [step_filter]`, where `[step_filter]` is a comma-separated substring matcher over step labels for selective re-runs without re-capture. The script encodes per-family Hugging Face IDs, config paths, and output stems. Steps in order:

1. **Tokenisation-alignment gate** (one-shot): `python bin/verify-tokenization.py --prompts etc/prompts-{chat,completion}.json --tokenizer <hf-id>` for each PT and IT tokenizer. Pre-registration requires all comparator pairs single-token in leading-space form (Appendix C); failures block downstream steps.
2. **Activation capture** (per condition): `python bin/mot-capture.py etc/config-<family>-<condition>.json` writes `results/<family>/<stem>-activations.hdf5` (per-(prompt, layer, token) residual-stream `x` plus per-(prompt, layer) `attn_output`, `mlp_output`; $\hat{\gamma}$ as cumulative sum). Four conditions: PT, IT-completion, IT-question-no-template, IT-chat.
3. **Pullback metric**: `python bin/extract-g-metric.py --model-name <hf-id> --output results/<family>/<stem>-G.npz` writes $G = U^T U$ once per model. IT-completion / IT-question-no-template / IT-chat reuse the IT model's `G`; PT uses PT's.
4. **Alignment-weight delta**: `python bin/analyse-alignment-weights.py --pt-model <pt-hf> --it-model <it-hf> --output results/<family>/alignment-weight-deltas.json` writes per-layer per-weight-type ΔW (source for Appendix D.6).
5. **Per-layer decision margin**: `python bin/compute-decision-margin.py --hdf5 <activations.hdf5> --model-name <hf-id> --prompts-file <prompts.json> --output results/<family>/decision-margin-<condition>.jsonl` per condition. Single-pass (no HDF5) via `python bin/logit-margin.py <config.json>` writes `<stem>_logit_margin.jsonl` from the config's `output_file` field.

6. **Multiplicative decomposition**: `python bin/analyse-decomposition.py --model <family>` reads the four `decision-margin-*.jsonl` and writes `results/<family>/decomposition-analysis.json` (per-cell ratios, 10k bootstrap CIs, per-suite breakdowns, D1 SD-log-ratio diagnostic).
7. **Curvature, salience,** A': `python bin/analyse-curvature-salience.py --hdf5 <activations.hdf5> --g-npz <G.npz> --prompts-file <prompts.json> --model-name <hf-id> --output-prefix results/<family>/cu salience-<condition> --trajectory both` writes two JSONLs (`-x.jsonl`, `-gamma-hat.jsonl`) per condition (source for section 4.6, Appendix D.2.3).
8. **Architectural divergence**: `python bin/analyse-architectural-divergence.py --model <family>` writes `results/<family>/architectural-divergence.json` (per-cell $|\hat{\gamma}\text{-}A'/x\text{-}A'|$ for A1).
9. **Minimal-pair** $\cos_G$: `python bin/analyse-minimal-pair.py --hdf5 <activations.hdf5> --g-npz <G.npz> --prompts-file <prompts.json> --condition <condition> --output results/<family>/minimal-pair-<condition>.jsonl` per condition. H1–H4 evaluation via `python bin/evaluate-minimal-pair-hypotheses.py --model <family>` writes `results/<family>/minimal-pair-evaluation.json`.
10. **Plots** (optional): `python bin/plot-trajectories.py` and `python bin/plot-curvature-salience.py` consume steps 5 and 7.

Cross-family aggregation (Tables 3, 7, 8):

- `python bin/aggregate-cross-family.py` → `results/cross-family-rollup.json` (Table 3).
- `python bin/clustered-bootstrap-sensitivity.py` → `results/clustered-bootstrap-sensitivity.json` (Appendix D.1).
- `python bin/analyse-prompt-form-decomposition.py` → `results/prompt-form-decomposition.json` (Appendix D.4).
- `python bin/check-architectural-norm-ratio.py` → `results/architectural-norm-ratio-robustness.json` (section 4.3 L2 check).

Survival checks and capability extensions:

- Qwen3 thinking-mode survival: `python bin/logit-margin.py etc/config-qwen-it-chat-no-think.json` produces the `qwen3_no_think` margins (Appendix D.3). The config differs from `config-qwen-instruct.json` only in `"chat_template_mode": "qwen3_no_think"` (suppresses `<think>…</think>` via `enable_thinking=False`).
- IT-question-no-template extension: see `it-question-no-template.sh`; the four canonical configs are `etc/config-<family>-it-question-no-template-full.json` (full HDF5) plus logit-margin-only siblings without the `-full` suffix.

E.4 Reverse-lookup - paper claim → artefact To verify a specific section 4 / section 5 claim against the released data, the table below gives the artefact and the re-derivation.

Table E.4.1. Reverse-lookup from headline paper claims to released artefact files.

Claim location	Numerical claim	Artefact file	Re-derivation
section 4.1 Table 3	Per-family post-training / scaffolding / combined ratios with bootstrap CIs	`results/cross-family-rollup.json`	Read `alignment_ratio`, `scaffolding_ratio`, `combined_ratio`; CIs pre-computed (10k bootstrap)
section 4.1 / 4.8 CVs	Scaffolding CV 0.74 (4-family aggregate) / 0.82 (5-family with Qwen3 default-thinking); post-training CV 0.30 / 0.35 (significant D2 comparison)	Same	Read `d2_check.scaffolding_ratio_cv`, `d2_check.alignment_ratio_cv`
section 4.2 capture-rate panel	Per-cell margin → capture-rate (% negative) per family/condition/suite	`results/<family>/decision-margin-<condition>.jsonl`	Signed margin from final-layer entry; aggregate proportion negative per suite

Claim location	Numerical claim	Artefact file	Re-derivation
section 4.3 architectural panel	Gemma $A'(\hat{\gamma})/A'(x) \approx 0.011$ (90× amplification)	`results/gemma3-1b/curvature-salience-pt-{x,gamma-hat}.jsonl`	Median per-cell `A_prime` per trajectory; ratio
section 4.3 L2 robustness	Architectural-axis L2 ratio agrees with G-A' ratio	`results/architectural-norm-ratio-robustness.json`	Direct read
section 4.5 cross-family heterogeneity	Per-family (post-training, scaffolding) profile	`results/cross-family-rollup.json`	Read both ratios per family
section 4.6 minimal-pair localisation	$\cos_G(x_M, x_{\mathrm{NM}})$ per-layer peak per condition	`results/<family>/minimal-pair-<condition>.jsonl`	Per-(layer, token) `1 - cos_g`; per-layer max over tokens
section 4.6 H1–H4 verdicts	Per-family hypothesis outcomes	`results/<family>/minimal-pair-evaluation.json`	Direct read
section 4.7 directional rotation	D1 SD-log-ratio per family > 0.5 (failure)	`results/<family>/decomposition-analysis.json`	Read `sd_log_ratio`
section 4.2 / D.3 Qwen3 thinking-mode	Capability + M + NM per-condition margins with thinking suppressed	`results/qwen3-1.7b/qwen3-1.7b-it-chat-no-think_logit_margin.jsonl`	Per-cell margin; aggregate by suite
section 5.4 prompt-form decomposition	Per-family prompt-form / template-token sub-effects	`results/prompt-form-decomposition.json`	Direct read
Appendix D.1 clustered bootstrap	Cell-level vs cluster-level CI widths	`results/clustered-bootstrap-sensitivity.json`	Direct read
Appendix D.6 ΔW vs γ-peak	Per-family Spearman ρ(layer-ΔW, γ-peak-distance)	`results/<family>/alignment-weight-deltas.json` + γ values from `analyse-alignment-weights.py` per-layer output	Compute Spearman ρ

E.5 Determinism and seed handling Captures are deterministic: per-(prompt, layer, token) residual-stream `x` in the released HDF5 is fully determined by (a) the HuggingFace checkpoint (pinned via commit hash in config files; see E.6) and (b) the prompt strings (released verbatim in `etc/prompts-{chat,completion}.json`). All capture configs use `do_sample: false`, `no_repeat_ngram_size: 0`; eval-mode dropout is inactive and no random sampling is performed (margins are read off the residual stream, not generated tokens).

Aggregate analyses use a fixed bootstrap seed (`numpy.random.default_rng(20260507)` in `analyse-decomposition.py` and `clustered-bootstrap-sensitivity.py`), so released bootstrap CIs are bit-reproducible.

Floating-point non-determinism across hardware: HDF5 files were produced on Apple M1 with MPS-backed fp32. Reproducing on CUDA or CPU is expected to match per-cell margins to within 10^{-3} (numerical noise floor; observed smallest-cell differences are $\sim 10^{-1}$, two orders above this floor). Table 3's aggregate ratios are stable to three significant figures across hardware in spot-checks.

E.6 Pinned model checkpoints The five canonical families and their HuggingFace IDs are pinned via the `model_name` field in `etc/config-{base,instruct,instruct-completion,llama-*,olmo-*,qwen-*,smollm2-*,gemma-*}.json` (one config per family per condition):

Table E.6.1. Pinned HuggingFace PT and IT checkpoint slugs per family.

Family	PT checkpoint	IT checkpoint
Gemma-3-1b	`google/gemma-3-1b-pt`	`google/gemma-3-1b-it`
Llama-3.2-1B	`meta-llama/Llama-3.2-1B`	`meta-llama/Llama-3.2-1B-Instruct`

Family	PT checkpoint	IT checkpoint
OLMo-2-1B	`allenai/OLMo-2-0425-1B`	`allenai/OLMo-2-0425-1B-Instruct`
Qwen3-1.7B	`Qwen/Qwen3-1.7B-Base`	`Qwen/Qwen3-1.7B`
SmolLM2-1.7B	`HuggingFaceTB/SmolLM2-1.7B`	`HuggingFaceTB/SmolLM2-1.7B-Instruct`

The `model_name` field is a slug, which is a moving reference - upstream maintainers can update the same slug to a different commit. To pin to exact revisions, set `HF_HUB_OFFLINE` against a snapshot of the local HuggingFace cache, or pass `revision=<commit-hash>` to `transformers` (not currently wired into `mot-capture.py`; needs a small patch). The Qwen3-1.7B revision used here is `70d244cc86ccca08cf5af4e1e306ecf908b1ad5e`; other families' revisions can be recorded via `huggingface-cli scan-cache` after a successful capture run.

E.7 Known caveats

- **Capture-time tokenizer mode flags** are important for IT-chat on Llama, SmolLM2, and Qwen3. The `chat_template_mode` in each IT-chat config is `llama_no_preamble`, `smollm2_no_preamble`, and `default` respectively; Qwen3 IT-chat emits `<think>…</think>` per its trained template, and the thinking-mode survival check uses `qwen3_no_think`. `bin/check-chat-template.py` diagnoses which preamble each family's default template injects; verbatim rendered prompts for all five families are in Appendix C.
- A' **is reported at the canonical CI default** $\alpha = 0.5$. Other α values produce different absolute magnitudes; the cross-family directional pattern (Peri-LN: $A'(x) \neq A'(\hat{\gamma})$; pre-norm-only: $A'(x) = A'(\hat{\gamma})$) is invariant to α in spot-checks but Appendix D.2.3 magnitudes are α-specific.
- **No 3-seed replication is needed for the residual-stream geometry** (`do_sample: false`, no eval-mode dropout; captures are deterministic given checkpoint + prompts). The 3-seed sampling-noise floor applies only to generation-based extensions, not used here.
- **HDF5 files can exceed dataset-host single-file limits** (3–6 GB per family-condition). Released artefacts are split per (family, condition); the cross-family rollup and per-family decomposition JSONLs are < 1 MB each and are the recommended starting point for most secondary analyses.

E.8 Cross-reference index for replication-adjacent appendices

- Pre-registration (binding hypotheses, criteria): Appendix A.
- Per-family model details: Appendix B.
- Tokenisation-alignment validation: Appendix C.
- Clustered-bootstrap sensitivity: Appendix D.1.
- Per-cell results tables and per-family / per-condition / per-suite distributions: Appendix D.2.
- Qwen3 thinking-mode survival: Appendix D.3.
- Prompt-form decomposition: Appendix D.4.
- Capability-control results: Appendix D.5.
- Alignment-weight ΔW analysis: Appendix D.6.