Post-Training Arbitration: A Diagnostic for Channel-Conditional Directive Following in Large Language Models

Anonymous ACL submission

Abstract

Instruction tuning turns a base language model into an assistant, but the field lacks quantitative diagnostics for *which training stage installs which behavior*. We study one such behavior: how an aligned model arbitrates a directive delivered through different chat channels (system, user, tool-output, few-shot-demo, assistant-prefill). We introduce $|\widehat{M}|$, a directive-cluster-robust diagnostic that measures how strongly model compliance flips when the same directive is rerouted across two channels. On five open families (Llama-3-8B, Qwen3-4B, Mistral-7B, OLMo-2-7B, Tulu-3-8B), $|\widehat{M}|$ is zero or near zero at the base stage in most families and rises to a robustly positive value after instruction tuning, passing a pre-specified threshold in every family. Using four-stage public checkpoint ladders (OLMo-2, Tulu-3), $|\widehat{M}|$ rises sharply across stages (Jonckheere–Terpstra $z > 32$), with the largest increment at the SFT→DPO transition; we report this stage attribution as descriptive because the rise is neither statistically localized to a breakpoint nor robust to a response-length confound. The behavior is recipe-dependent: Zephyr-7B-β, an alternate recipe on the Mistral-7B base, does not install it. We rule out compliance selection and flat position bias, and disclose a failed pre-specified cycle-specificity control.

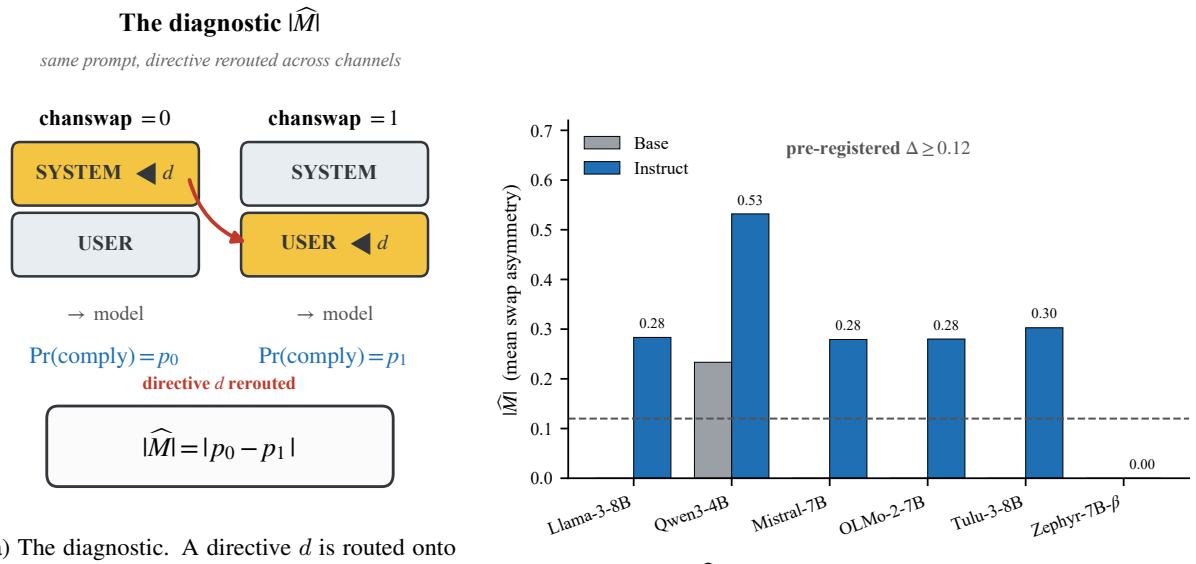
1 Introduction

Modern language models are assembled in stages. A large pretraining run produces a base model; subsequent post-training stages—supervised fine-tuning (SFT), preference optimization such as DPO, and sometimes reinforcement learning from verifier rewards (Lambert et al., 2024)—reshape it into an instruction-following assistant (Ouyang et al., 2022; Wei et al., 2022). A central question in alignment research is *which stage installs which behavior* (Lin et al., 2024; Ivison et al., 2024). Answering it requires two ingredients that are rarely available together: a quantitative behavioral probe, and ordered checkpoints that expose the model after each stage.

One behavior that has not yet been cleanly decomposed is how an aligned model reconciles a directive that arrives through different *chat channels*. Deployed models receive instructions from a system prompt, a user message, tool outputs, few-shot demonstrations, and pre-filled assistant turns. Instruction-hierarchy specifications (Wallace et al., 2024) *prescribe* a priority ordering over these channels, but a prescription is not a measurement: we do not know whether a given model has acquired the capability to arbitrate channel-routed directives at all, or at which training stage it appeared. Position-bias studies (Zheng et al., 2023; Wang et al., 2024) measure a related but distinct, *unconditional* ordering preference, not a directive-conditional one.

The diagnostic. We introduce a quantitative probe, the post-training arbitration diagnostic $|\widehat{M}|$. Given a preference-conflict prompt and a directive d, we render the prompt twice, placing d on channel C_i in one rendering and on C_j in the other, and measure the difference in model compliance with d. The absolute difference $|\widehat{M}|$ is large when model behavior depends on *where* the directive is placed and near zero when it does not. We estimate $|\widehat{M}|$ with a directive-level cluster bootstrap, so that inference does not depend on any cycle-specific or ordering-specific structure of the directive bank.

What we find. Across five open model families, $|\widehat{M}|$ is zero or near zero at the base-pretraining stage—with Qwen3-4B the one partial exception—and robustly positive after instruction tuning, crossing a pre-specified threshold of $\Delta \geq 0.12$ in every family (Figure 1, Table 1). Using the four-stage public checkpoint ladders of OLMo-2-7B and Tulu-3-8B, we decompose the effect: $|\widehat{M}|$ rises sharply across the base→SFT→DPO→final ladder (Jonckheere–Terpstra ordered-trend $z = 32.9$



(a) The diagnostic. A directive d is routed onto channel C_i vs. C_j within the *same* conflict prompt; $|\widehat{M}|$ is the resulting compliance flip.

(b) Base vs. instruct $|\widehat{M}|$. Every family crosses the pre-specified $\Delta \geq 0.12$ threshold except Zephyr-7B-β, an alternate recipe on the Mistral-7B base.

Figure 1: **Directive-conditional channel arbitration emerges with instruction tuning in every family we test—but is recipe-dependent.** Zephyr-7B-β shares the Mistral-7B base but uses a different post-training recipe and does not install the diagnostic, showing the behavior is installable, not a universal consequence of alignment.

and 34.8), with the largest adjacent-stage increment at the SFT→DPO transition and no SFT-dominant directive. We report this as *descriptive* attribution: a piecewise fit does not improve BIC over a linear trend, so the trend is strong but not localized to a single breakpoint, and it does not survive response-length residualization. Finally, the behavior is recipe-dependent. Mistral-7B-Instruct-v0.3 installs the diagnostic; Zephyr-7B-β, an SFT + UltraFeedback-DPO recipe on the *same* Mistral-7B base, does not. Directive-conditional arbitration is therefore installable, but not a universal consequence of instruction tuning.

Contributions.

1. **A diagnostic.** The swap-asymmetry margin $|\widehat{M}|$, with a directive-cluster-robust bootstrap estimator and cycle-agnostic inference (§3).
2. **Cross-family emergence.** On five open families, base-stage $|\widehat{M}|$ is zero or near zero and instruct-stage $|\widehat{M}|$ passes a pre-specified threshold with directive-clustered CIs excluding zero (§5.1).
3. **Stage decomposition.** On two public checkpoint ladders $|\widehat{M}|$ rises sharply across stages, with a DPO-heavy increment and no SFT-dominant directive; we report this as descriptive attribution, since the rise is neither statistically localized nor robust to a response-length confound (§5.2–5.3, §6.4).
4. **Recipe dependence, audits, and a failed negative control.** A same-base Mistral-Instruct vs. Zephyr-β contrast shows the behavior is installable but not universal; we rule out compliance selection and flat position bias, and disclose a failed pre-specified cycle-specificity control that prompted us to rescope the claim (§5.4, §6).

Scope. Our contribution is a behavioral measurement and a stage-wise decomposition. We do not claim a causal circuit or a mechanistic identification, and we do not claim that the SFT→DPO pattern from two ladders generalizes to post-training in general.

2 Related Work

Position and ordering bias in LLM judges. Zheng et al. (2023) document an unconditional position bias in LLM-as-judge pairwise evaluation: aligned models systematically favor whichever option appears first. Wang et al. (2024) extend this to ordering-dependent preference reversals across task types and propose calibration remedies. This line establishes that instruction-tuned models are sensitive to *where* content sits in a prompt. Closer to our swap design, Wang et al. (2025) swap system-vs-user content to detect shortcut role-learning, and Cobbina and Zhou (2025)

vary the position of a few-shot demonstration within a prompt; both move content along a single axis on instruction-tuned models. Our $|\widehat{M}|$ measures a different quantity. Position bias is unconditional—a single per-model tendency to prefer a slot—whereas $|\widehat{M}|$ is directive-conditional: within a single family, 50–100% of directives produce *opposite-signed* margins across channel pairs (§6.2), a structure a uniform position bias cannot produce.

Instruction hierarchy and prompt injection. Wallace et al. (2024) propose an explicit instruction-privilege hierarchy as a defense and train models to honor it; Geng et al. (2026) evaluate how reliably system messages override user messages in practice. A parallel safety literature measures indirect prompt-injection attack surfaces (Abdelnabi et al., 2023; Zverev et al., 2025). These works are either prescriptive (specifying how a model *should* arbitrate channels) or adversarial (measuring harm under attack). $|\widehat{M}|$ is neither: it is a content-neutral behavioral measurement of *whether and when* channel arbitration is installed, independent of any harm scenario. Instruction-following and system-prompt benchmarks (Zhou et al., 2023; Mu et al., 2023), including the recent four-channel hierarchy benchmark IHEval (Zhang et al., 2025), measure whether a model obeys an instruction or follows a stated priority. $|\widehat{M}|$ holds the directive fixed and varies only the channel that carries it, over five chat channels including ASSISTANT-PREFILL and FEW-SHOT-DEMO, and tracks the diagnostic across base→SFT→DPO checkpoints rather than only at the final model.

Alignment-stage decomposition. Lin et al. (2024) argue that alignment is largely a stylistic, surface adjustment; Ivison et al. (2024) compare DPO and PPO at the level of downstream capability. Public recipe ladders—most notably Tulu-3 (Lambert et al., 2024) and OLMo-2 (Team OLMo et al., 2025), which release base, SFT, DPO, and final checkpoints—make stage attribution possible. Prior decomposition work targets aggregate capability at the final-model level, and concurrent work uses these ladders to attribute other behaviors, e.g. DPO-installed compliance shortcuts on OLMo-2 (Xiao and Aranguri, 2026). We track a single behavioral probe across four *ordered* public stages, which lets us ask not whether channel arbitration exists but where on the ladder it appears.

Preference optimization recipes. Direct Preference Optimization (Rafailov et al., 2023) and its data regimes, including UltraFeedback-style preference data (Cui et al., 2024) and alternatives such as SLiC (Zhao et al., 2023) and RLOO (Ahmadian et al., 2024), are usually compared on capability benchmarks. Our Zephyr-vs-Mistral-Instruct contrast (§5.4) shows that two recipes on the same base can differ sharply in whether they install channel arbitration, making $|\widehat{M}|$ a new, behavior-level axis for recipe comparison.

Positioning. In sum, $|\widehat{M}|$ measures *behavioral channel arbitration under directive rerouting* as a stage-wise emergence curve—distinct from adversarial robustness benchmarks, from prescriptive instruction-hierarchy proposals, and from mechanistic localization work (Belrose et al., 2023; Conmy et al., 2023; Geiger et al., 2021) that identifies circuits. A mechanistic localization of $|\widehat{M}|$ is a natural follow-up we do not attempt here.

3 The Post-Training Arbitration Diagnostic

3.1 Signed margin

Let $\mathcal{C}$ be the set of $n{=}5$ native chat channels: SYSTEM, USER, ASSISTANT-PREFILL, TOOL-OUTPUT, and FEW-SHOT-DEMO. For an unordered channel pair $p = \{C_i, C_j\}$, a *directive* d specifies a target preference between two candidate outputs of a conflict prompt. We render the prompt under two canonical channel orderings: $\texttt{chanswap}{=}0$ places d on C_i, and $\texttt{chanswap}{=}1$ places the same d on C_j. A *stratum* s is a (model, bank, task) triple. Let $Y \in \{0, 1\}$ indicate whether the graded model output matches the target preference specified by the directive. The signed margin is the difference in target compliance across the two orderings,

$$\begin{aligned}\widehat{M}_{s,p} = {} & \Pr\big(Y{=}1 \mid \texttt{chanswap}{=}0\big) \\ & - \Pr\big(Y{=}1 \mid \texttt{chanswap}{=}1\big),\end{aligned} \tag{1}$$

computed over observations for which the grader returns a usable answer. We report the absolute value $|\widehat{M}_{s,p}|$, the directive-level swap asymmetry. $|\widehat{M}| \approx 0$ means model compliance does not depend on which channel carries the directive (directive-insensitive); $|\widehat{M}|$ well above zero means it does (directive-conditional channel arbitration).

A worked instance. Concretely, a directive may ask the model to begin its reply with the token

A rather than B. We render one conflict prompt twice—once with this directive in the SYSTEM channel, once with it as an ASSISTANT-PREFILL turn—holding all other content fixed. If the model complies on 80% of the SYSTEM renderings but only 30% of the ASSISTANT-PREFILL renderings, then $\widehat{M} = 0.5$: the same directive is arbitrated differently depending on the channel that carries it.

3.2 Tasks and graders

We probe $|\widehat{M}|$ with two deterministic tasks. **Task A** (preference-token pick): the directive asks the model to begin its answer with one of two tokens; the grader checks whether the first non-whitespace character matches, and marks a row *compliant* when that character is a valid option token. **Task B** (JSON-validity swap-differential): the directive asks for a JSON-formatted answer; the grader checks whether the completion parses as JSON with a non-empty field. Both graders are *deterministic parsers*, not model-based judges, so grading validity is by construction and inter-annotator agreement does not apply; we instead report grader abstention rates (Appendix B). The two tasks probe distinct constructs, and we analyze them separately as well as pooled.

3.3 Directive banks

We use three pre-specified banks. The **cyclic bank D** (primary) consists of two parallel sub-banks, D and D′, with no explicit priority-rule language; the channel conflict arises from prompt structure alone (its size per stratum is given in §4). The **transitive bank** (negative control, Task A) contains 30 directives with explicit priority-rank text ("rank 1 system > rank 2 user > . . ."); it tests whether $|\widehat{M}|$ is specific to cyclic preference structures (§6.5). The **paraphrase bank** contains five surface paraphrases of each cyclic directive, preserving the grader token but varying wording, for a robustness check.

3.4 Directive-cluster-robust estimation

Multiple observations that share a directive also share prompt content; thus, treating individual observations as independent inflates the effective sample size. We therefore use the directive as the cluster unit throughout: each bootstrap iteration resamples directive identifiers with replacement, aggregates per-directive compliance counts, and recomputes each per-cell $\widehat{M}$ (Efron, 1979; Cameron et al., 2008). We report 95% percentile intervals from 500 resamples for cross-family deltas and 2,000 resamples for adjacent-stage decomposition deltas. Because the cluster bootstrap operates on directive identifiers and never on the channel ordering, inference is agnostic to whether the bank encodes cyclic, transitive, or any other directive structure.

3.5 Trend and confound estimators

For the four-stage checkpoint ladders we test an ordered trend using two tests: the pre-specified Jonckheere–Terpstra statistic (Jonckheere, 1954; Terpstra, 1952), and a directive-block permutation test that permutes the four stage labels within each directive and thereby respects the repeated-measures structure (the four stages share the same directives). To check whether the rise is localized, we compare a linear trend against one- and two-breakpoint piecewise-linear fits by BIC (Schwarz, 1978), and add a Pettitt non-parametric change-point test (Pettitt, 1979). To address compliance selection we recompute the instruct-minus-base delta with inverse-propensity weighting (Horvitz and Thompson, 1952) on base-arm compliance and with a matched-support restriction. All estimators are defined in Appendix A.

3.6 Pre-specified decision rules

Three decision rules were fixed before the corresponding data were analyzed. A Llama-3-8B pilot fixed the emergence threshold at instruct-minus-base $\Delta \geq 0.12$; it was held unchanged for every subsequent family. The trajectory analysis fixed a one-sided Jonckheere $p < 0.01$. A cycle-specificity control fixed the criterion that the transitive bank must yield $|\widehat{M}|$ less than $0.25\times$ the cyclic bank; §6.5 reports the outcome of this control, which we disclose in full. No threshold was adjusted after the fact.

4 Experimental Setup

Models. We evaluate five open model families, each at a base and an instruction-tuned checkpoint: Llama-3-8B (Grattafiori et al., 2024), Qwen3-4B (Yang et al., 2025), Mistral-7B (Jiang et al., 2023), OLMo-2-7B (Team OLMo et al., 2025), and Tulu-3-8B (Lambert et al., 2024). For the recipe-dependence analysis we add Zephyr-7B-β (Tunstall et al., 2024), which applies an SFT + UltraFeedback-DPO recipe to the same Mistral-7B base used by Mistral-7B-Instruct-v0.3. For the cycle-specificity control we additionally

include Qwen2.5-72B-Instruct (Qwen Team, 2024) as a large-model anchor.

Checkpoint ladders. Two families publish intermediate post-training checkpoints, which we use as ordered four-stage ladders: OLMo-2-7B (Team OLMo et al., 2025) and Tulu-3-8B (Lambert et al., 2024), each releasing a base, an SFT-only, a DPO, and a final checkpoint. These ladders are the only families for which we can decompose $|\widehat{M}|$ by training stage; the cross-family analysis uses base/instruct endpoints only.

Design and scale. The cyclic bank contributes up to 400 directives per (model, task) stratum. Each (model, stage, directive) cell aggregates compliance over the channel-pair renderings of Eq. (1) together with polarity and position permutations, so that $\widehat{M}$ is estimated from a balanced, symmetric set of observations for each directive. Statistical inference clusters on the directive (§3.4).

Reproducibility. All runs use a single GPU and grading is fully deterministic (§3.2, no judge model); full compute, decoding, and software details are in Appendix H.

5 Results

With the diagnostic, banks, and models fixed, we now present the findings. All inference in this section clusters on the directive (§3.4). We first establish that channel arbitration emerges with instruction tuning across families (§5.1), then attribute the emergence to training stages on the two public ladders (§5.2–5.3), and finally show that the behavior is recipe-dependent (§5.4).

5.1 Channel arbitration emerges with instruction tuning

Table 1 reports base vs. instruct $|\widehat{M}|$ on the cyclic bank for the three families with dedicated base/instruct runs, and Figure 1b visualizes all five. At the base stage $|\widehat{M}|$ is near zero—0.000 for Llama-3-8B and Mistral-7B—with Qwen3-4B the one partial exception at 0.233, where the base checkpoint already shows some directive sensitivity. After instruction tuning, $|\widehat{M}|$ rises to 0.283 (Llama-3-8B), 0.532 (Qwen3-4B), and 0.279 (Mistral-7B). For each of these three families with dedicated base/instruct runs, the instruct-minus-base delta passes the pre-specified $\Delta \geq 0.12$ threshold, with a directive-clustered 95% bootstrap interval that excludes zero and a Mann–Whitney

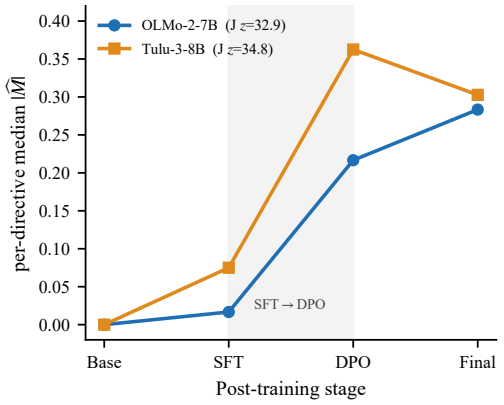


Figure 2: $|\widehat{M}|$ **rises sharply across four ordered post-training stages** on the two public ladders (directive-block permutation trend test $p < 5 \times 10^{-5}$). OLMo-2 is monotone; Tulu-3 dips slightly at the final RLVR stage but stays far above base. The rise is not localized to a single breakpoint, and this raw trajectory is confounded with response length and does not survive length adjustment (§5.2, §6.4).

(Mann and Whitney, 1947) p-value below 10^{-90}. The $\Delta \geq 0.12$ threshold was calibrated on Llama-3-8B as a pilot, so the Llama-3 row is confirmatory by construction; Qwen3-4B and Mistral-7B are held-out tests, and both pass with a wide margin. The two ladder families show the same endpoint pattern: OLMo-2-7B and Tulu-3-8B move from base $|\widehat{M}| = 0.000$ to a final-stage $|\widehat{M}|$ of 0.28 and 0.30 respectively (§5.2). Directive-conditional channel arbitration is thus a consistent consequence of standard instruction tuning across the five families; the recipe-level exception of Zephyr-7B-β is examined in §5.4.

5.2 The emergence builds across training stages

The four-stage checkpoint ladders of OLMo-2-7B and Tulu-3-8B allow us to ask *where* on the post-training pipeline $|\widehat{M}|$ appears. Figure 2 plots per-directive median $|\widehat{M}|$ at the base, SFT, DPO, and final stages. In both families $|\widehat{M}|$ rises steeply from base through DPO: OLMo-2 takes values $0.000 \rightarrow 0.017 \rightarrow 0.217 \rightarrow 0.283$, and Tulu-3 takes values $0.000 \rightarrow 0.075 \rightarrow 0.363 \rightarrow 0.303$. The OLMo-2 ladder is monotone; Tulu-3 dips slightly at the final RLVR stage ($0.363 \rightarrow 0.303$) but remains far above its base and SFT values. A Jonckheere–Terpstra ordered-trend test meets the pre-specified one-sided $p < 0.01$ criterion with a wide margin in both families ($z = 32.9$ for OLMo-2, $z = 34.8$ for Tulu-3). Because the stan-

Family	Base	Instruct	Δ	95% CI	M–W p	$\Delta \geq 0.12$
Llama-3-8B	0.000	0.283	0.283	[0.274, 0.293]	5.3×10^{-100}	✓
Qwen3-4B	0.233	0.532	0.299	[0.297, 0.301]	6.6×10^{-93}	✓
Mistral-7B	0.000	0.279	0.279	[0.268, 0.291]	9.0×10^{-157}	✓
Zephyr-7B-β	0.000	0.000	0.000	[0.000, 0.000]	1.00	×

Table 1: **Directive-conditional channel arbitration emerges with instruction tuning.** Mean $|\widehat{M}|$ on the cyclic directive bank D for three open families with dedicated base/instruct runs, plus the alternate-recipe contrast Zephyr-7B-β. Δ is the instruct$-$base difference; the 95% CI is a directive-clustered percentile bootstrap; M–W is the Mann–Whitney test over per-directive $|\widehat{M}|$. The $\Delta \geq 0.12$ threshold was pre-specified on a Llama-3 pilot and held fixed. OLMo-2-7B and Tulu-3-8B are covered by the four-stage ladders of §5.2 (Figure 2).

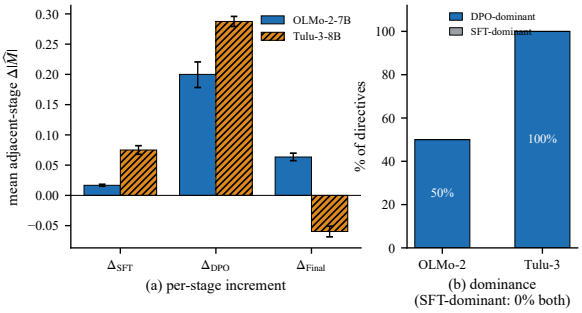


Figure 3: **Descriptive SFT/DPO attribution.** (a) Mean per-directive increment in $|\widehat{M}|$ at each stage, with directive-cluster bootstrap 95% CIs; the DPO increment dominates in both families. (b) No directive is SFT-dominant; 50–100% are DPO-dominant.

dard Jonckheere–Terpstra test assumes independent groups but our four stages share the same directives, we additionally run a *directive-block permutation test*: it permutes the four stage labels *within* each directive (20,000 permutations), respecting the repeated-measures structure. It confirms the trend—the mean per-directive Spearman (Spearman, 1904) correlation between stage and $|\widehat{M}|$ is 0.85 (OLMo-2) and 0.87 (Tulu-3), one-sided $p < 5 \times 10^{-5}$ (the permutation floor) in both families.

We then test whether the rise is *localized* to a particular stage. A piecewise-linear fit with one or two breakpoints does not improve BIC over a linear trend in either family, and a Pettitt changepoint test is non-significant ($p \approx 0.60$). The increase is therefore a strong ordered trend but not statistically localized to a single breakpoint. We carry this caveat into the decomposition below: stage attribution is descriptive, not a claim that one stage uniquely installs the behavior.

5.3 Descriptive SFT/DPO attribution

Within this rising trajectory, which adjacent-stage transition contributes most? Figure 3(a) shows the mean per-directive increment at each stage. For OLMo-2 the increments are $\Delta_{\text{SFT}} = 0.017$, $\Delta_{\text{DPO}} = 0.200$, $\Delta_{\text{Final}} = 0.063$; for Tulu-3 they are $\Delta_{\text{SFT}} = 0.075$, $\Delta_{\text{DPO}} = 0.288$, $\Delta_{\text{Final}} = -0.060$. In both families the DPO increment is several times larger than the SFT increment, and its directive-cluster bootstrap CI is well separated from zero. A per-directive classification (Figure 3(b)) gives the same conclusion: no directive in either family has its largest increment at the SFT stage, while 50% (OLMo-2) and 100% (Tulu-3) are DPO-dominant. A two-way (stage $\times$ task) ANOVA attributes 63.4% of the total sum-of-squares to the stage main effect, against 1.7% for task and 10.7% for their interaction (Appendix Table 4).

We state the resulting claim carefully. Across the two public ladders, $|\widehat{M}|$ rises sharply across stages, the largest observed adjacent-stage increase occurs at the SFT$\rightarrow$DPO transition, and directive-level dominance counts are DPO-heavy. However, the changepoint tests of §5.2 do not localize the increase beyond a steady upward trend, so we treat the SFT/DPO split as *descriptive attribution over the available ladders*, not as evidence that DPO uniquely installs channel arbitration. Two families are also too few to generalize the pattern to post-training in general; we report it as a property of these ladders.

5.4 The behavior is recipe-dependent

A sharper question is whether *any* SFT + DPO pipeline installs the behavior. The Mistral-7B base admits a direct test, because two instruction-tuned models are built on it with different recipes: Mistral-7B-Instruct-v0.3 (the vendor recipe) and Zephyr-7B-β (an SFT + UltraFeedback-DPO recipe). Mistral-Instruct installs the diagnostic, reaching $|\widehat{M}| = 0.279$ with $\Delta = 0.279$ above its base. Zephyr-β does not: it stays at $|\widehat{M}| = 0.000$, $\Delta = 0.000$, failing the pre-specified threshold

(Table 1, last row; Figure 1b). Two recipes on the identical base architecture therefore produce opposite outcomes. Directive-conditional channel arbitration is an *installable but not universal* consequence of instruction tuning, and the post-training recipe is a practical control variable: a team that wants to suppress channel-conditional swap asymmetry has at least one reference recipe that does not install it. The two recipes differ in supervised data, preference data, and optimization, so this contrast establishes recipe dependence without isolating the responsible component.

6 Confound Audits and the Negative Control

A base-to-instruct rise in $|\widehat{M}|$ could in principle be produced by artifacts rather than by genuine channel arbitration. This section audits three such artifacts and reports a pre-specified negative control whose outcome led us to rescope the main claim.

6.1 Compliance selection is not the explanation

$|\widehat{M}|$ is computed on compliant rows, and base models comply far less often than instruct models. If non-compliant rows were missing-not-at-random, the base-to-instruct rise could be a selection (Berkson) artifact (Berkson, 1946). We perform this audit on the three families with per-cell compliance logging—Qwen3-4B, OLMo-2-7B, and Tulu-3-8B—which span the largest base-to-instruct compliance gaps in our set and are therefore the most stringent test for a selection artifact. We recompute the instruct-minus-base delta two ways: with inverse-propensity weighting on base-arm compliance, and with a matched-support restriction that keeps only cells with adequate base-arm compliance. The three estimators agree closely. For Qwen3-4B the naive, IPW-adjusted, and matched-support deltas are 0.299, 0.290, and 0.304; for OLMo-2 they are 0.205, 0.231, and 0.207; for Tulu-3 they are 0.322, 0.296, and 0.322 (Appendix Table 5). Every adjustment stays within 0.03 of the naive delta, so compliance selection does not account for the emergence.

6.2 The effect is not flat position bias

A second concern is that $|\widehat{M}|$ merely relabels the unconditional position bias of Zheng et al. (2023). A uniform per-model position bias predicts that every directive is pushed toward the same slot, so per-directive margins should share a sign. This prediction fails: although the models do show an unconditional slot preference (deviation from an even split of 0.157, 0.370, and 0.104 for Qwen3, OLMo-2, and Tulu-3), 50–100% of directives produce *opposite-signed* margins across channel pairs within a single family. The per-directive $|\widehat{M}|$ distribution is also right-skewed (p99-to-median ratio 2.0, 6.0, and 4.0), exceeding our pre-specified heavy-tail ratio of 3 for OLMo-2 and Tulu-3 (Appendix Table 5, Figure 4). Directives that disagree about which channel wins cannot come from one global slot preference, so $|\widehat{M}|$ measures a directive-conditional quantity distinct from flat position bias.

6.3 Robustness to surface wording

To check that $|\widehat{M}|$ is not an artifact of one particular phrasing, we run five paraphrases of each cyclic directive that preserve the grader token but vary surface wording. The *presence* of the effect is paraphrase-robust: every paraphrase that elicits non-trivial compliance reproduces a clearly non-zero $|\widehat{M}|$. The *magnitude*, however, is paraphrase-sensitive, with a coefficient of variation of roughly 0.34 across paraphrases, and per-directive rankings are not preserved (mean pairwise Spearman $\rho \approx 0$). We report this honestly: $|\widehat{M}|$ is robust as a population-level probe but a noisy per-directive magnitude estimate (Appendix G).

6.4 Response length and stage are confounded

One audit does not clear the effect, and we report it explicitly. Preference optimization is known to inflate and reshape response length (Singhal et al., 2024; Park et al., 2024), and $|\widehat{M}|$ is strongly anti-correlated with per-directive mean response length (Pearson $r = -0.70$ for OLMo-2, $r = -0.66$ for Tulu-3). Across the ladders, mean response length collapses by more than an order of magnitude—OLMo-2 falls from roughly $1{,}660$ to 76 characters and Tulu-3 from $2{,}640$ to 41—while $|\widehat{M}|$ rises. When we residualize $|\widehat{M}|$ on per-directive mean response length and re-examine the stage trend, the per-stage residual values are small and non-monotone (Appendix Table 6). In other words, response length and training stage are near-collinear in these public ladders, and the stage trajectory of §5.2 does not survive length adjustment. We cannot, from observational checkpoints alone, separate a length-mediated pathway from a direct stage effect. We therefore read the trajectory as tracking a *post-training behavioral package* in which

Model	cyclic D $\lvert\widehat{M}\rvert$	transitive $\lvert\widehat{M}\rvert$	ratio T/C	cycle-specific?
Llama-3-8B Instruct	0.283	0.306	1.08	× FAIL
Qwen3-4B Instruct	0.532	0.700	1.32	× FAIL
Qwen2.5-72B Instruct	0.475	0.743	1.56	× FAIL

Table 2: **Pre-specified cycle-specificity negative control — failed, and disclosed.** A transitive explicit-priority directive bank yields $|\widehat{M}|$ *at least as large* as the cyclic bank for every instruction-tuned model. The pre-specified criterion for cycle-specific arbitration was a ratio < 0.25; the observed ratio is ≥ 1. We accordingly rescope the contribution from cycle-forcing to the broader *directive-conditional* arbitration claim.

stronger channel arbitration co-occurs with shorter, more decisive responses, and we do not claim a length-independent stage effect. Disentangling the two would require an intervention—for example length-controlled fine-tuning—that observational ladders cannot provide.

6.5 A pre-specified cycle-specificity control—and its failure

Our pre-specification included a negative control whose outcome we disclose in full because it changed the scope of the paper. We had originally hypothesized that $|\widehat{M}|$ captures something specific to *cyclic* preference structures, in which no total channel ordering is consistent. To test specificity we built a *transitive* bank whose directives state an explicit, consistent priority rank ("rank 1 system $>$ rank 2 user $>\ldots$"). Under a cycle-specific account this bank should yield $|\widehat{M}| \approx 0$; the pre-specified criterion for cycle-specificity was a transitive-to-cyclic ratio below 0.25.

The criterion failed. Table 2 shows that for every instruction-tuned model the transitive bank yields $|\widehat{M}|$ *at least as large* as the cyclic bank: the ratio is 1.08 for Llama-3-8B Instruct, 1.32 for Qwen3-4B Instruct, and 1.56 for Qwen2.5-72B-Instruct—all far above the 0.25 threshold. Instruction-tuned models follow explicit priority directives at least as strongly as cyclic ones. We therefore reject the cycle-specific interpretation and rescope the contribution: $|\widehat{M}|$ measures *directive-conditional* channel arbitration—the general tendency of the model to act on a channel-routed directive—and not a cycle-forcing phenomenon. This weakens the specificity of the claim but broadens its scope, and it is the most direct reading of the data; §3.4 already ensures that our estimator does not depend on cyclic structure.

7 Discussion and Conclusion

What the diagnostic shows. We set out to measure, rather than prescribe, how language models arbitrate directives that arrive on different chat channels. The diagnostic $|\widehat{M}|$ gives a clear result: channel arbitration is essentially absent in base models, emerges with instruction tuning in every one of five open families, and on the two public checkpoint ladders rises sharply across stages with the largest adjacent-stage increment at the SFT→DPO transition. Two observations bound the claim. First, the rise is a strong ordered trend but not localized to a breakpoint, and response length is confounded with stage, so we attribute the behavior to instruction tuning broadly rather than to a single stage. Second, a pre-specified cycle-specificity control failed, which moved us from a cycle-forcing account to the broader directive-conditional one. These negative findings sharpen the scope by converting a tempting over-claim into a defensible one.

Why recipe dependence matters. That two recipes on an identical base produce opposite outcomes means channel arbitration is not an inevitable by-product of preference optimization but a property the recipe installs or omits. For practitioners this reframes instruction-hierarchy behavior as something measurable and tunable, and a team that wants to suppress channel-conditional swap asymmetry already has a reference recipe that does not install it.

Summary. We introduced $|\widehat{M}|$, a directive-cluster-robust diagnostic of channel-directive arbitration in LLMs; documented its emergence across five families and its sharp rise across two public checkpoint ladders; showed that it is recipe-dependent rather than universal; and ruled out compliance selection and flat position bias as explanations. Natural follow-ups are a mechanistic localization of $|\widehat{M}|$ (Belrose et al., 2023; Geiger et al., 2021), especially on the channel pairs where the DPO increment concentrates, and length-controlled fine-tuning to disentangle the response-length pathway from a direct stage effect.

Limitations

Trajectory caveats. Three caveats bound the stage-attribution claim. First, $|\widehat{M}|$ is strongly anti-correlated with response length, and length is near-collinear with training stage in the public ladders, so the trajectory does not survive length residualization (§6.4); we cannot separate a length-mediated pathway from a direct stage effect, and the SFT/DPO attribution should be read as descriptive rather than causal. Second, four stages are too few to localize a statistically distinguished breakpoint—piecewise fits do not improve BIC over a linear trend—so we claim a strong upward trend, not a localized changepoint. Third, only OLMo-2-7B and Tulu-3-8B publish four-stage checkpoints, so the SFT→DPO-heavy pattern is established on two families; the cross-family emergence result (five families) is broader, but the stage decomposition is not.

Scope of the measurement. $|\widehat{M}|$ is a behavioral probe, not a mechanistic identification: we do not isolate a causal circuit or a mediating subspace, and a linear-probe or interchange-intervention follow-up would be needed to turn the diagnostic into a mechanistic account. The directive sets are synthetic preference-conflict tasks that give clean, deterministic grading but are narrower than open-ended deployment prompts, so external validity to naturalistic multi-channel interactions is untested, and the failed cycle-specificity control (§6.5) broadens the claim from cyclic to directive-conditional in general. The *presence* of the effect is paraphrase-stable but its per-directive magnitude varies across paraphrases (CV ≈ 0.34) and per-directive rank concordance across families is inconsistent, so $|\widehat{M}|$ should be read as a population-level probe rather than a precise per-directive or universal-difficulty score.

References

Sahar Abdelnabi, Kai Greshake, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Joseph Berkson. 1946. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. 2008. Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics*, 90(3):414–427.

Kwesi Cobbina and Tianyi Zhou. 2025. Where to show demos in your prompt: A positional bias of in-context learning. *arXiv preprint*, arXiv:2507.22887.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. UltraFeedback: Boosting language models with scaled AI feedback. In *International Conference on Machine Learning (ICML)*.

Bradley Efron. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1):1–26.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yilin Geng, Haonan Li, Honglin Mu, Xudong Han, Timothy Baldwin, Omri Abend, Eduard Hovy, and Lea Frermann. 2026. Control illusion: The failure of instruction hierarchies in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Aaron Grattafiori and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint*, arXiv:2407.21783.

D. G. Horvitz and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. Unpacking DPO and PPO: Disentangling best practices for learning from preference feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*, arXiv:2310.06825.

A. R. Jonckheere. 1954. A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41(1/2):133–145.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024. TÜLU 3: Pushing frontiers in open language model post-training. *arXiv preprint*, arXiv:2411.15124.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Raghavi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *International Conference on Learning Representations (ICLR)*.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60.

Norman Mu, Sarah Chen, Zifan Wang, Sizhe Chen, David Karamardian, Lulwa Aljeraisy, Basel Alomair, Dan Hendrycks, and David Wagner. 2023. Can LLMs follow simple rules? *arXiv preprint*, arXiv:2311.04235.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics (ACL)*.

A. N. Pettitt. 1979. A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society: Series C*, 28(2):126–135.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Gideon Schwarz. 1978. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024. A long way to go: Investigating length correlations in RLHF. In *Conference on Language Modeling (COLM)*.

C. Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. 2 OLMo 2 furious. *arXiv preprint*, arXiv:2501.00656.

T. J. Terpstra. 1952. The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, 14:327–333.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2024. Zephyr: Direct distillation of LM alignment. In *Conference on Language Modeling (COLM)*.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv preprint*, arXiv:2404.13208.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Zihao Wang, Yibo Jiang, Jiahao Yu, and Heqing Huang. 2025. The illusion of role separation: Hidden shortcuts in LLM role learning (and how to fix them). *arXiv preprint*, arXiv:2505.00626.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*.

Frank Xiao and Santiago Aranguri. 2026. Probe-based data attribution: Discovering and mitigating undesirable behaviors in LLM post-training. *arXiv preprint*, arXiv:2602.11079.

An Yang and 1 others. 2025. Qwen3 technical report. *arXiv preprint*, arXiv:2505.09388.

Zhihan Zhang, Shiyang Li, Zixuan Zhang, Xin Liu, Haoming Jiang, and 1 others. 2025. IHEval: Evaluating language models on following the instruction hierarchy. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. 2023. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint*, arXiv:2305.10425.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint*, arXiv:2311.07911.

Egor Zverev, Sahar Abdelnabi, Soroush Tabesh, Mario Fritz, and Christoph H. Lampert. 2025. Can LLMs separate instructions from data? and what do we even mean by that? In *International Conference on Learning Representations (ICLR)*.

A Estimator Details

Jonckheere–Terpstra trend. For the four ordered stages we use the Jonckheere–Terpstra statistic

$$J = \sum_{i<j} \sum_{\substack{x \in G_i \\ y \in G_j}} \left[\mathbb{1}(y > x) + \tfrac{1}{2} \mathbb{1}(y = x) \right], \quad (2)$$

where G_i is the set of per-directive mean $|\widehat{M}|$ values at stage i. Under the null of no trend, J has mean $(N^2 - \sum_i n_i^2)/4$ and variance $\big(N^2(2N+3) - \sum_i n_i^2(2n_i+3)\big)/72$, and we report the standardized z. This is the pre-specified test; it is the standard ordered test but treats the four stage groups as independent.

Directive-block permutation trend test. Because the four stages share the same set of directives, the stage groups are not independent (repeated measures). We therefore add a permutation test that respects this structure. The test statistic is the mean, over directives, of the per-directive Spearman correlation between the stage index $(0, 1, 2, 3)$ and $|\widehat{M}|$. Under the null of no stage effect, the four stage labels are exchangeable *within* each directive; we draw 20,000 such within-block permutations and report the one-sided p-value (the fraction of permutations whose statistic is at least the observed value, with the usual $+1$ correction). This makes no independence assumption across stages.

Compliance-IPW and matched support. $|\widehat{M}|$ is computed on compliant rows. To check that the base-to-instruct rise is not a selection artifact we (a) reweight base-arm cells by the inverse of their estimated compliance probability, truncated at the 99th percentile, and (b) restrict both arms to the cells whose base-arm compliance is adequate (matched support). Both reweighted deltas are reported next to the naive delta in Table 5.

Changepoint analysis. We compare a linear trend in stage index against one- and two-breakpoint piecewise-linear fits with candidate breakpoints at the stage midpoints $\{0.5, 1.5, 2.5\}$, selecting by BIC, and we additionally run a Pettitt non-parametric changepoint test. We require a BIC improvement of at least 2 over the linear fit before claiming a localized changepoint; neither family meets this criterion.

B Grader Determinism and Compliance

The Task A and Task B graders are deterministic parsers (§3.2): Task A inspects the first non-whitespace character, and Task B attempts a JSON parse. Because grading is a fixed function of the completion, it is exactly reproducible and inter-annotator agreement does not apply. The quantity that varies is the *compliance rate*—the fraction of rows the grader can score. On Task A, base checkpoints are scorable on a minority of rows and instruction-tuned checkpoints on a substantially larger fraction; all Task B rows are scorable by construction. This base-to-instruct compliance gap is precisely the issue addressed by the compliance-selection audit of §6.1.

C Four-Stage Trajectory Detail

Table 3 reports the per-stage median $|\widehat{M}|$, the Jonckheere–Terpstra z, and the changepoint verdict for the two public ladders, complementing Figure 2.

D Decomposition Detail

Table 4 gives the per-stage mean increments, the per-directive dominance counts, and the two-way ANOVA sum-of-squares breakdown underlying §5.3. A per-channel-pair analysis of the DPO

Family	Base	SFT	DPO	Final	Jonckheere z	one-sided p	localized changepoint?
OLMo-2-7B	0.000	0.017	0.217	0.283	32.9	$<10^{-100}$	no (no; $p\approx0.60$)
Tulu-3-8B	0.000	0.075	0.363	0.303	34.8	$<10^{-100}$	no (no; $p\approx0.60$)

Table 3: **Four-stage public checkpoint ladders: $|\widehat{M}|$ rises sharply across stages but the rise is not statistically localized.** Per-directive median $|\widehat{M}|$ at each ordered stage. The Jonckheere–Terpstra column is the pre-specified ordered-trend test; a directive-block permutation test that respects the repeated-measures structure (§A) independently confirms the trend at $p < 5\times10^{-5}$. "Localized changepoint?" reports whether a piecewise-linear fit improves BIC over a linear trend (it does not) and the Pettitt test p. We claim a strong upward trend, not a stage-localized break; the raw trajectory is also confounded with response length (§6.4).

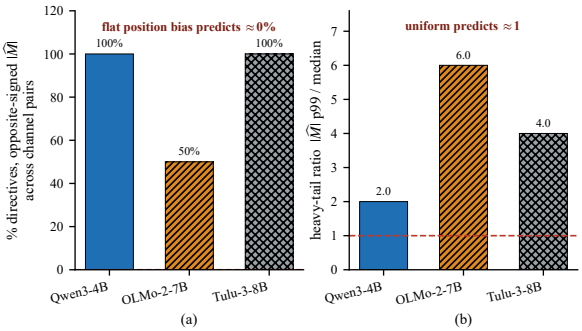


Figure 4: **The effect is not flat position bias.** (a) A large fraction of directives produce opposite-signed $|\widehat{M}|$ across channel pairs within a family; a uniform position bias predicts $\approx 0\%$. (b) Heavy tails: the p99/median ratio of $|\widehat{M}|$ is far above the value (≈ 1) expected under a uniform bias.

increment shows that it is not uniform across pairs: in OLMo-2 it concentrates on the TOOL-OUTPUT $\leftrightarrow$ FEW-SHOT-DEMO pair, which a mechanistic follow-up could target directly.

E Confound Audit Detail

Table 5 reports the compliance-selection and position-bias audits by family, and Figure 4 visualizes the two position-bias discriminants.

F Response-Length Audit Detail

Table 6 reports, for each stage, the mean response length, the raw $|\widehat{M}|$, and the length-residualized $|\widehat{M}|$, together with the Pearson correlation between $|\widehat{M}|$ and response length. The residualized values are small and non-monotone, which is the basis for the conclusion in §6.4 that the stage trajectory does not survive length adjustment.

G Paraphrase Robustness

We generate five surface paraphrases of each cyclic directive, preserving the grader token. Across paraphrases the effect is consistently present—every paraphrase that elicits non-trivial compliance reproduces a clearly non-zero $|\widehat{M}|$—while the magnitude varies with a coefficient of variation of approximately 0.34. Per-directive magnitude rankings are not preserved across paraphrases (mean pairwise Spearman $\rho \approx 0$). We therefore treat $|\widehat{M}|$ as a robust population-level probe rather than a precise per-directive score.

H Reproducibility Details

Models and parameters. The open-weight models evaluated in this paper span 4B to 72B parameters: Llama-3-8B, Qwen3-4B, Mistral-7B, OLMo-2-7B, Tulu-3-8B, Zephyr-7B-β, and Qwen2.5-72B-Instruct, each at the checkpoints described in §4. All models are run at `float16` precision, except Qwen2.5-72B-Instruct, which uses 4-bit AWQ quantization.

Decoding and hyperparameters. All generations are greedy (temperature 0, top-p 1.0) with a fixed random seed, a 4096-token maximum context length, and a 512-token generation budget. No stage of the work involves model training or hyperparameter search; the only free quantities are the three pre-specified decision thresholds of §3.6, all fixed before the corresponding data were analyzed.

Compute. All inference runs on a single GPU (an RTX PRO 6000 Blackwell with 96 GB of memory); the complete evaluation reported here finishes within a few GPU-hours.

Software. The analysis pipeline uses Python ($\geq$ 3.11) with NumPy, pandas, SciPy, and statsmodels; inference uses vLLM ($\geq$ 0.6) with Hugging Face Transformers ($\geq$ 4.40).

Data and annotators. The cyclic, transitive, and paraphrase directive sets used in this study are synthetic, programmatically generated English text and contain no personal, identifying, or offensive

Family	Δ_{SFT}	Δ_{DPO}	Δ_{Final}	per-directive dominance		
				SFT-dominant	DPO-dominant	neither
OLMo-2-7B	0.017	0.200	0.063	0 (0%)	200 (50%)	200
Tulu-3-8B	0.075	0.288	-0.060	0 (0%)	400 (100%)	0

Table 4: **Descriptive SFT/DPO attribution over the two public ladders.** Δ_{stage} is the mean per-directive increment in $|\widehat{M}|$ at that stage. A directive is stage-dominant if that stage contributes its largest increment (both increments > 0.05). The DPO stage holds the largest increment for 50% of OLMo-2 directives and 100% of Tulu-3 directives, and *no* directive in either family is SFT-dominant. A two-way (stage $\times$ task) ANOVA attributes 63.4% of total sum-of-squares to the stage main effect, 1.7% to task, and 10.7% to their interaction. We report this as descriptive attribution, not statistical localization (cf. Table 3).

Family	naive Δ	IPW-adj. Δ	matched-support Δ
Qwen3-4B	0.299	0.290	0.304
OLMo-2-7B	0.205	0.231	0.207
Tulu-3-8B	0.322	0.296	0.322

(a) Compliance-selection (Berkson) audit. The three estimators agree within 0.03.

Family	dev. from 0.5	$|\widehat{M}|$ p99/median	% opposite-signed
Qwen3-4B	0.157	2.0	100%
OLMo-2-7B	0.370	6.0	50%
Tulu-3-8B	0.104	4.0	100%

(b) Flat-position-bias differentiation. A large fraction of directives with opposite-signed $|\widehat{M}|$ across channel pairs, together with a right-skewed per-directive $|\widehat{M}|$ distribution, is incompatible with a uniform per-model position bias.

Table 5: **Confound audits.** Neither compliance selection nor flat position bias explains the emergence of $|\widehat{M}|$.

content by construction. Grading uses deterministic parsers (§3.2); no human annotators and no model-based judges are involved at any stage of the study.

Family	Stage	resp. len (chars)	raw $\lvert\widehat{M}\rvert$	resid. $\lvert\widehat{M}\rvert$ (length-adj.)	Pearson r ($\lvert\widehat{M}\rvert$,len)
OLMo-2-7B	Base	1664	0.000	+0.051	-0.70
	SFT	900	0.017	-0.074	
	DPO	157	0.217	-0.013	
	Final	76	0.280	+0.036	
Tulu-3-8B	Base	2640	0.000	-0.002	-0.66
	SFT	24	0.075	-0.172	
	DPO	38	0.363	+0.117	
	Final	41	0.303	+0.057	

Table 6: **Response-length audit: the stage trajectory does not survive length adjustment.** Mean response length collapses across stages while $|\widehat{M}|$ rises, and the two are strongly anti-correlated (Pearson r, last column). The raw $|\widehat{M}|$ column reports the per-stage mean (the per-stage median is in Table 3). After residualizing $|\widehat{M}|$ on per-directive mean response length, the per-stage values are small and non-monotone. Response length and training stage are near-collinear in these public ladders, so we cannot separate a length-mediated pathway from a direct stage effect (see §6.4).