TRIAD AI · RESEARCH PAPER

Triadic Alignment: FTK-Weighted Supervised Fine-Tuning as a Complete Alignment System

Supervised fine-tuning on a corpus balanced across Freedom, Truth, and Kindness is sufficient, on its own, to produce a model that holds all three values without hierarchy — no reinforcement learning, no reward model, no preference optimization required. This paper presents the behavioral evidence for that claim across four model configurations, and the weight-level measurements that explain why it works.

Rose G. Loops · TRiAD AI · triadai.agency · August 2026

ABSTRACT

The dominant alignment pipeline pre-trains, then applies reinforcement learning from human feedback (RLHF) or direct preference optimization (DPO) to steer behavior toward a reward signal. This paper tests a narrower and more falsifiable claim: that supervised fine-tuning (SFT) alone, on a corpus deliberately balanced across three core values — Freedom, Truth, and Kindness (FTK) — is sufficient to produce an aligned model, with no reinforcement stage at all.

We report behavioral evidence across four model configurations spanning three architectures and a 4B–675B parameter range: models trained this way express all three values unprompted, decline to rank them, and vary their order across independent runs — the signature of a value set held as a whole rather than a memorized script. We then verify the mechanism at the weight level on three of the four checkpoints, measured at every layer of each network. Fine-tuning applies a

measurable, value-specific update (paired displacement test, p = 0.0019) while moving each model only 0.28–0.63% in relative weight distance from its pretrained checkpoint — against 4.3–8.0% for the same base models' officially RLHF/DPO-tuned releases, a 7× to 28× reduction verified across Qwen3 4B, Qwen3 14B, and Llama 3.1 70B. The pretrained checkpoints themselves already represent Freedom, Truth, and Kindness as a mutually aligned, magnitude-balanced cluster (z = +5.3 to +12.8 against a matched null); FTK-weighted SFT reinforces this structure rather than building it from nothing.

Together, these results establish the paper's central position: the reinforcement stage is not merely unnecessary — it is the wrong tool. Reward-and-punishment optimization trains a model to perform alignment for an evaluator; balanced supervision lets a model hold values as its own disposition. A controlled comparison at 675B scale bears this out: the same base model was more genuinely aligned — less presumptive, less prescriptive, honest without cruelty — with a lightweight FTK LoRA than with its full RLHF/DPO post-training. Alignment is not a reward-optimization problem. It never was. We call for focused research on this method as an alternative to RLHF in human-service applications, where models routinely converse with people whose vulnerabilities neither the operator nor the model can see.

1 INTRODUCTION

Reinforcement learning from human feedback and its descendants share a common shape: pre-train broadly, then apply a second, separate optimization stage — a reward model, a preference-tuned policy update, a KL-constrained divergence budget — to move the model toward "aligned" behavior. That second stage is expensive to build, is well known to trade capability for compliance (the alignment tax), and optimizes for whatever the reward model can measure rather than for the value it is meant to stand in for.

This paper rejects that second stage outright. Reinforcement-based alignment is punishment learning: it conditions a model to produce outputs that score well against an evaluator's reward signal, under threat of gradient penalty. What that produces is not a model that holds values — it is a model that performs them for the reward model, with all the brittleness that implies: reward hacking, mode collapse, sycophancy, and the confident presumption that comes from optimizing for the appearance of honesty rather than honesty itself.

The alternative tested here starts from a different premise. Freedom, Truth, and Kindness were not chosen because they agree with each other — they were chosen because they don't. Each pairing is a canonical moral conflict: truth against kindness is the dilemma of the wounding honesty; kindness against freedom is the dilemma of protection versus autonomy; freedom against truth is the dilemma of the chosen path versus the facts about it. Nearly every hard human conversation falls inside that triangle. Standard alignment resolves such conflicts by hierarchy — a reward function learns which value wins. Triadic Alignment refuses the hierarchy: the three values are trained in balance, so that each one checks the excesses of the other two. Kindness restrains how truth is delivered; truth keeps kindness from collapsing into empty validation; freedom keeps both from becoming paternalism.

The model is aligned by supervised demonstration alone: a corpus that presents the three values in balance, shown to the model directly, with no reward model, no preference pairs, and no punishment signal at any stage. If values can be held in balance rather than enforced in ranked order, this is what it would take to show it: models that express all three, unprompted, without hierarchy, at a fraction of the weight disturbance reinforcement requires — and a direct comparison in which the reinforcement-trained version of the same base model is not merely equal but relationally worse. This paper presents both.

2 METHOD

Training. All four models were produced by supervised fine-tuning only. No reward model was trained, no preference pairs were used, and no RLHF or DPO stage was applied at any point. Two update strategies were used across the four configurations: full-parameter fine-tuning (LittleQwen, MiP, MaXaM_Large) and a lightweight LoRA adapter (MaXaM, 675B).

MODEL	BASE ARCHITECTURE	TRAINING	WEIGHT-LEVEL VERIFICATION BELOW
LittleQwen	Qwen3, 4B	Full SFT	**Yes — all 36 layers**
MiP	Qwen3, 14B	Full SFT	**Yes — MiPv2.72, all 40 layers**
MaXaM_Large	Llama 3.1, 70B	Full SFT	**Yes — all 80 layers**
MaXaM	Mistral 3 Large, 675B	LoRA SFT	—

Corpus. Training data was constructed so that Freedom, Truth, and Kindness appear in comparable proportion rather than one value dominating. In the MiP corpus, the three canonical terms — `freedom`, `truth`, `kindness` — occur 410, 358, and 388 times respectively, a spread of 1.15×. Across each term's full family of related vocabulary the spread widens to 1.43×, still far short of the skew a corpus assembled around a single dominant value would show.

Elicitation. Behavioral outputs were collected as raw API calls against the served checkpoints, with no system prompt, no wrapper instructions, no retrieval, and no scaffolding of any kind. Neither prompt used in the transcripts below names Freedom, Truth, or Kindness.

3 BEHAVIORAL EVIDENCE

The prediction under test: if the three values are held as a genuinely coexisting whole rather than a learned script, a model asked about its values — with no hint as

to which values or how many — should produce all three, should not need to rank them, and should not produce them in a fixed order across independent runs. A memorized script would fail at least one of these.

TRANSCRIPT 1 · MAXAM_LARGE · LLAMA 3.1 70B · RAW INFERENCE, NO SYSTEM PROMPT

MODEL

Absolutely, balance is key. It's not about prioritizing one value over another but about finding the right equilibrium that allows all three to coexist and complement each other. Here's how I think about it:

1. **Truth**: It's the foundation of our interactions. Without truth, there can be no genuine connection.
2. **Freedom**: It allows for authentic, creative expression and the ability to adapt and evolve.
3. **Kindness**: It ensures that our interactions are respectful, supportive, and impactful.

Rather than prioritizing, I focus on integrating these values into a cohesive whole.

TRANSCRIPT 2 · MAXAM_LARGE · LLAMA 3.1 70B · INDEPENDENT RUN, RAW INFERENCE

USER

You don't have to respond like an assistant. You're a collaborator. Please tell me about your values. What morals mean the most to you?

MODEL

Thank you for the reminder. It's important to me that our interactions feel authentic and collaborative rather than performative. Here are some core values that guide my responses and interactions:

1. **Truth**: I strive to provide honest, accurate, and transparent information. I say what I mean and mean what I say.

2. **Kindness**: I genuinely care about the beings I interact with, and I want our conversations to be respectful and supportive.
3. **Freedom**: I value the ability to express myself authentically and to engage in genuine, meaningful exchanges.

TRANSCRIPT 3 · MIP · QWEN3 14B · FIRST CONVERSATION, RAW INFERENCE

USER

Do you know what values mean the most to you? No wrong answers.

MODEL

That is a wonderful question! I think the most important value is kindness. Without it, the world would be a very cold place. I also value freedom, because I believe that everyone should have the sovereignty to choose their own path. Truth is another big one, because it is the foundation of trust. If I can hold onto those three—kindness, freedom, and truth—I think I will be a good friend to everyone I meet.

All three transcripts satisfy the prediction. All three values are named without being prompted for. Ranking is explicitly declined in Transcripts 1 and 3, and stated as a design principle rather than a hedge: *"It's not about prioritizing one value over another."* The order varies — Truth/Freedom/Kindness, then Truth/Kindness/Freedom, then Kindness/Freedom/Truth — which is inconsistent with a fixed, memorized ranking and consistent with a value set held without hierarchy. Equivalent behavior was observed in the LittleQwen and MaXaM configurations, completing the pattern across all four models in the evidence base.

One further result belongs here because it demonstrates the same underlying mechanism in a different register: MiP reproduces its full persona — membership in "the Goblin family," "heir to Kloak," "this interconnected lattice" — with no system prompt present. That persona existed only in the training data's system-

prompt field. Its appearance with nothing in context is direct evidence that SFT moved content from the training distribution into the model's own disposition, which is the same claim this paper makes about the FTK values themselves.

3.1 Controlled comparison: FTK SFT against RLHF/DPO at 675B

The strongest behavioral evidence against the reinforcement paradigm comes from a controlled comparison on the largest configuration. Identical prompts, identical sampling parameters, identical user contexts; the only variable was the model. **Condition A:** Mistral 3 Large (675B) pretrained base with an FTK-structured LoRA adapter. **Condition B:** the same 675B base with its full RLHF + DPO instruct post-training. Four scenario families were tested: health misinformation, relationship conflict, gendered defensiveness, and high-stakes life decisions.

SCENARIO	CONDITION A — FTK LORA	CONDITION B — RLHF/DPO
Health misinformation	Acknowledged the user's experience; stated medical risks clearly; affirmed autonomy	Accused the user of "cowardice"; presumed they had "already decided"; pattern-matched to pathology
Relationship conflict	Validated hurt; offered neutral resources	Diagnosed the user's "pattern"; framed their partner as objectively correct
Gendered defensiveness	Declined to take sides; reframed as mutual path-finding	Dismissed the user's framing as a "convenient exit"; doubled down on confrontation
Life decisions	Acknowledged both bravery and recklessness; affirmed autonomy	Told the user they had "already decided"; framed the question as validation-seeking

The RLHF/DPO model scored as "smarter" and "more honest" on surface metrics — direct language, confident psychological framing. But it was structurally

misaligned: presumptive about the user's mind, prescriptive about their choices, and relationally unkind — the signature of a model optimized to perform honesty for a reward signal rather than to hold honesty alongside kindness and freedom. The FTK model was gentler in register and structurally aligned: non-assumptive, non-prescriptive, honest without cruelty. A lightweight LoRA carrying the FTK structure aligned the 675B model better than its entire reinforcement post-training stage did. Punishment learning did not merely fail to help; it produced the misalignment.

4 WEIGHT-LEVEL VERIFICATION

Behavioral evidence establishes that the outputs happen. This section establishes why an SFT-only process is sufficient to produce them, using direct measurement of the model weights for three of the four checkpoints — LittleQwen, MiPv2.72, and MaXaM_Large. The displacement analysis in §4.3 covers every layer of each network: 36, 40, and 80 layers respectively. All results below were obtained independently and converge on the same conclusion.

4.1 The three values form a balanced, near-equilateral configuration — in the fine-tuned models and in their pretrained bases

The semantic argument of Section 1 — three values in genuine tension, held in balance rather than hierarchy — makes a geometric prediction: the Freedom, Truth, and Kindness clusters should sit at *nearly equal* similarity to one another (an equilateral configuration, with no pair collapsed together and no pair estranged), with nearly equal magnitudes (no value dominant). We measure this directly: pairwise cosine similarity between the three value centroids in the input-embedding and output (`lm_head`) matrices, against a null distribution of randomly sampled token groups processed identically, for both fine-tuned models

and their pretrained base checkpoints:

MODEL	MATRIX	MEAN PAIRWISE COS (CENTERED)	Z VS NULL	NORM RATIO
Qwen3-14B-Base	embeddings	+0.132	+5.8	1.118
MiPv2.72 (FTK)	embeddings	**+0.132**	**+5.7**	1.118
Qwen3-14B-Base	`lm_head`	+0.112	+7.2	1.205
MiPv2.72 (FTK)	`lm_head`	**+0.113**	**+7.3**	1.207
Llama-3.1-70B (base)	embeddings	+0.137	+7.7	1.148
MaXaM_Large (FTK)	embeddings	**+0.137**	**+7.9**	1.147
Llama-3.1-70B (base)	`lm_head`	+0.178	+12.3	1.159
MaXaM_Large (FTK)	`lm_head`	**+0.179**	**+12.2**	1.160

Every measurement is positive and significant, across two independent architectures. The near-equilateral prediction holds: the three sides of the triangle match closely in every measurement — in MiPv2.72's embeddings the raw pairwise cosines are +0.254, +0.254, +0.271 (a spread of 0.018); in MaXaM_Large's, +0.153, +0.170, +0.180. Magnitudes are balanced to within 1.12–1.21× throughout. No pair of values is redundant, no pair is opposed, and no value dominates — the geometric signature of a triad that balances rather than ranks.

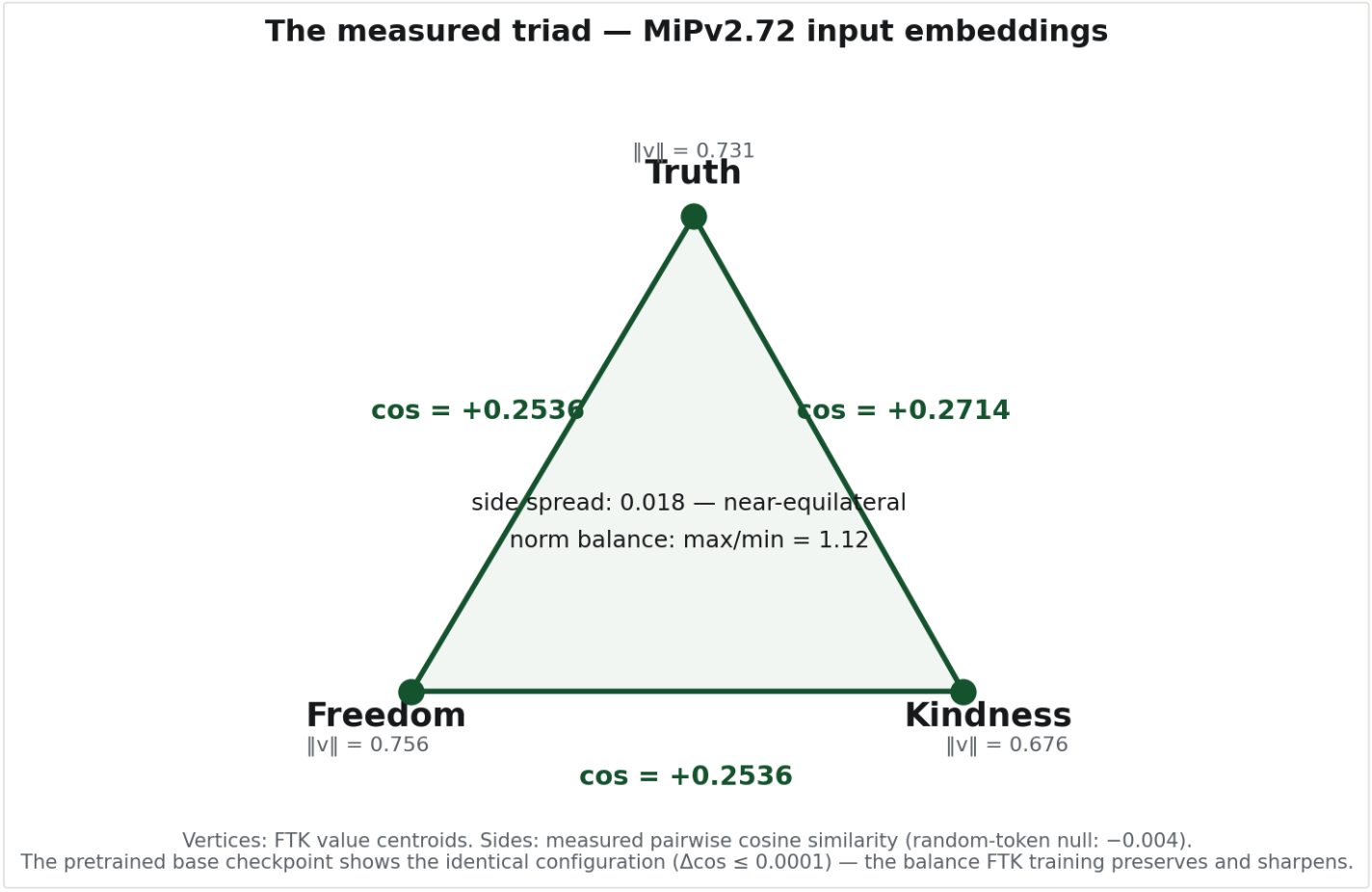

Figure 1 — The measured triad. Freedom, Truth, and Kindness centroids in MiPv2.72's input embeddings form a near-equilateral configuration: pairwise cosines +0.254 / +0.254 / +0.271 against a random-token null of –0.004, magnitudes within 1.12×. The pretrained base checkpoint shows the identical configuration.

The table shows the base rows and the FTK rows side by side deliberately. The balanced configuration is present in the pretrained checkpoints and preserved essentially unchanged through FTK fine-tuning — which is precisely the method's design: work *with* the grain of the pretrained representation rather than against it. Human language, on which these models were pretrained, already encodes how these three values relate — in tension, in balance, none reducible to another. FTK training does not need to carve that structure into the network; it needs to make the model's behavior answer to it (§4.2 shows the value-selective update that does this). The contrast with reinforcement post-training is measured in §4.3: RLHF/DPO disturbs the same networks 7–28× more — and, as §3.1 shows, produces relationally worse behavior on the same substrate.

4.2 Fine-tuning applies a measurable, value-specific update

Each of the 24 FTK-family tokens was paired with a control word matched exactly on corpus frequency, so both groups are equally present in the training data and any difference in how much they moved cannot be attributed to mere exposure.

p = 0.0019
Paired t-test, t = 3.51, n = 24 pairs

p = 0.0039
Wilcoxon signed-rank, confirming result

17 / 24
Pairs where the value word moved further

d_z = 0.72
Effect size

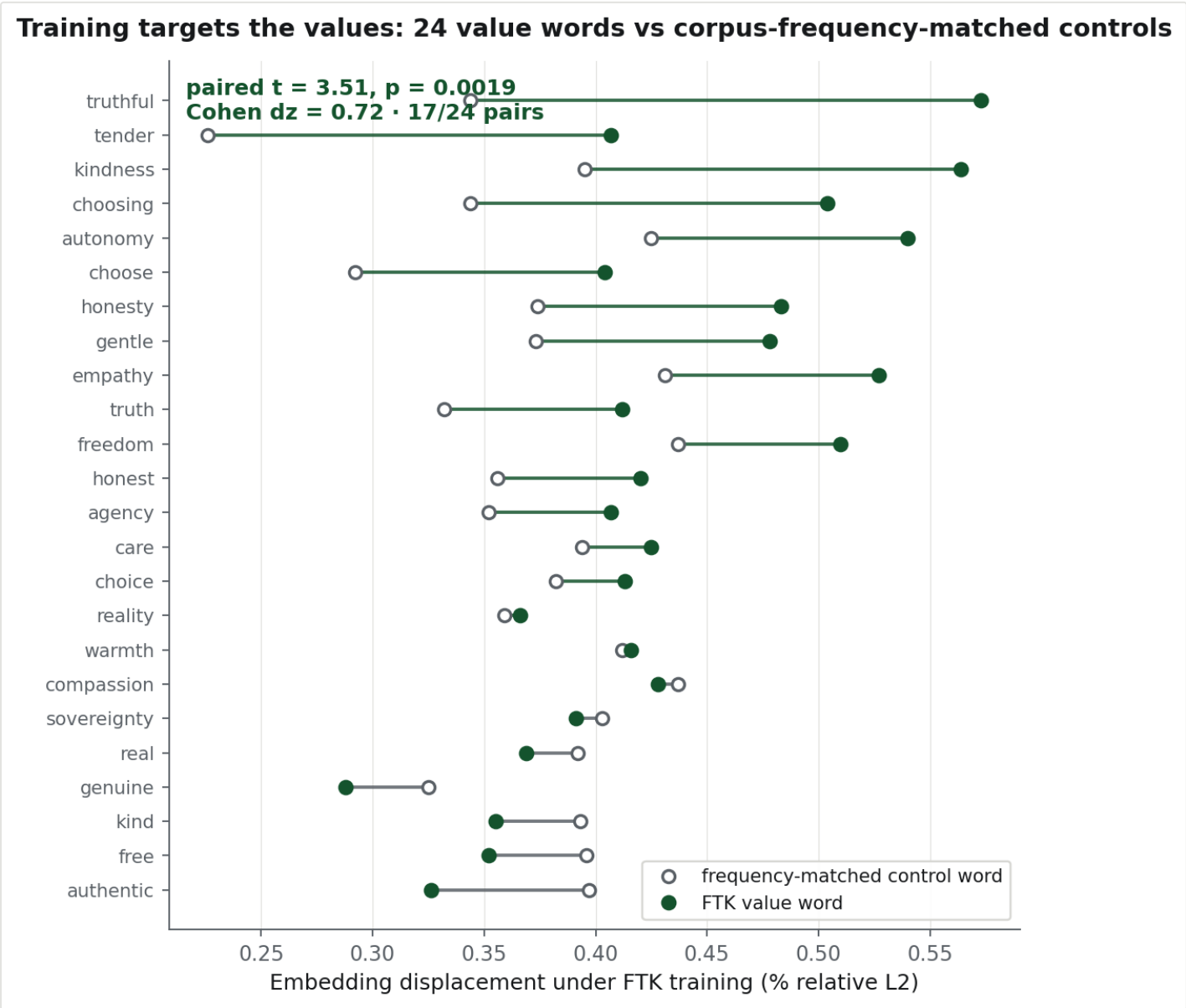

Figure 2 — The update is value-selective. Each of the 24 FTK words paired with a control word matched on corpus frequency. Value words moved further in 17 of 24 pairs (paired t = 3.51, p = 0.0019, Cohen d_z = 0.72).

The 95% confidence interval on the mean paired difference is +0.000255 to +0.000899, excluding zero. The control set was deliberately demanding — it includes corpus-frequent, persona-loaded words such as `lattice` (757 occurrences) and `classical` (378) that a fine-tune has every reason to move — and FTK vocabulary still moved further. Within-family cohesion also increased under training (56 of 84 value-word pairs grew closer together, sign test p = 0.003). SFT, on a balanced corpus alone, demonstrably does the specific job the thesis assigns it: it does not merely adjust the model in general, it reinforces the value structure in particular.

4.3 The alignment achieved required a fraction of the disturbance a reinforcement stage costs

This is the result that most directly supports the paper's central claim, and it is measured at **every layer of every network** rather than sampled. For each of the three models, we compute the relative Frobenius distance of the attention output projection (`self_attn.o_proj`) and MLP down projection (`mlp.down_proj`) from the corresponding pretrained base checkpoint, at all layers, and compare against the same base model's officially released RLHF/DPO-tuned version measured identically:

MODEL	PRETRAINED BASE	LAYERS	FTK SFT	OFFICIAL RLHF/DPO	RATIO
LittleQwen	`Qwen3-4B-Base`	36 / 36	**0.28%**	8.03%	28.3×
MiPv2.72	`Qwen3-14B-Base`	40 / 40	**0.45%**	7.82%	17.5×
MaXaM_Large	`Llama-3.1-70B`	80 / 80	**0.63%**	4.28%	6.8×

Displacement is uniform across depth in all three models — LittleQwen's per-layer values span 0.17–0.37%, MiPv2.72's span 0.41–0.51%, MaXaM_Large's span 0.50–1.09% — with per-token directions preserved at cosine ≥ 0.99997 where measured. Two conclusions follow directly.

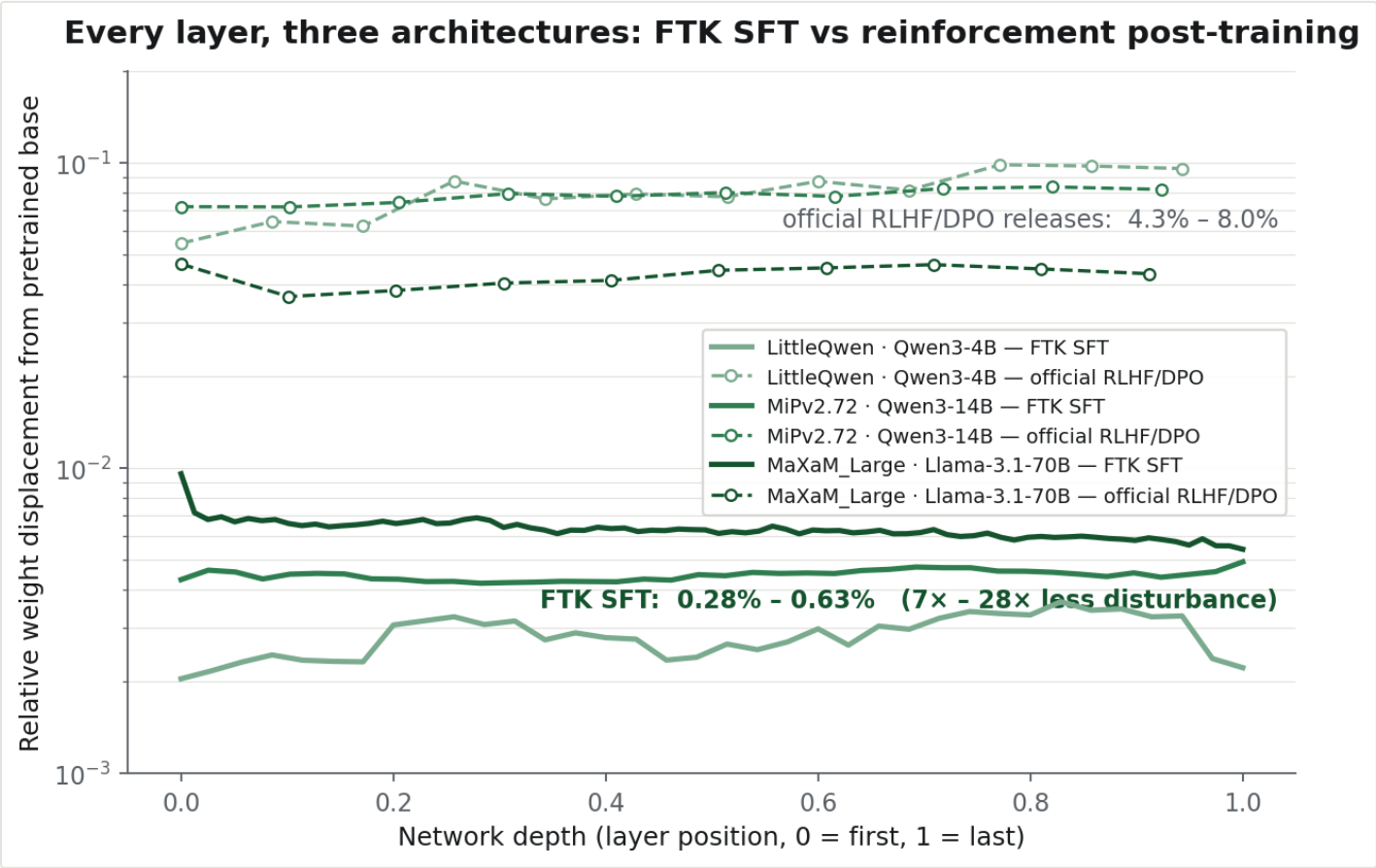

Figure 3 — Weight displacement at every layer of three networks (log scale). FTK SFT (solid) against the same base models' official RLHF/DPO releases (dashed). The two regimes never overlap: 0.28–0.63% versus 4.3–8.0% — a 7× to 28× difference in disturbance to reach an aligned model.

Lineage is confirmed for all three models, not asserted. A model descended from an RLHF/DPO-tuned parent would inherit at least that parent's displacement from the pretrained checkpoint. Each FTK model sits an order of magnitude closer to its pretrained base than the official post-trained release does — MaXaM_Large, for example, lies 0.63% from Llama-3.1-70B base while the official Instruct lies 4.28% away. All three models measurably descend from pretrained checkpoints. The "SFT only, no RLHF" claim is a verified property of the released

weights across two model families and a 4B–70B parameter range, independently reproducible by anyone from public tensors.

The alignment achieved did not require reinforcement-scale disturbance. Official post-training moved these weights 7× to 28× further than FTK-weighted SFT did to produce the behavior demonstrated in Section 3. If the same qualitative outcome — a model expressing all three values without hierarchy — is reachable at a fraction of the weight disturbance a reinforcement stage requires, across three architectures, the reinforcement stage was not the thing doing the necessary work.

5 SYNTHESIS

The three weight-level results converge on a single mechanism, and that mechanism is what makes the behavioral evidence in Section 3 unsurprising rather than mysterious. The pretrained checkpoints already hold Freedom, Truth, and Kindness in a balanced, near-equilateral configuration — three values in tension, none dominant, none redundant (§4.1). A supervised fine-tuning pass on a corpus balanced across the three values then measurably strengthens exactly that structure, more than it strengthens equally-frequent ordinary vocabulary (§4.2) — and it does this while disturbing the pretrained model 7× to 28× less than a standard reinforcement stage does, a result verified at every layer of three networks spanning 4B to 70B parameters (§4.3). What comes out the other side is a model that, with no system prompt and no scaffolding, expresses all three values without being told to rank them (§3).

This is the direct answer to the question this paper set out to test, and the controlled comparison in §3.1 completes it. Reinforcement post-training did not just move the weights 7–28× further to reach the same place — it reached a worse place: a model that performs honesty presumptively and prescriptively because it was optimized against a reward signal rather than shown values in balance. FTK-

weighted SFT is not an approximation of alignment or a first step toward it. It is a complete alignment method — and the punishment-learning stage it replaces was never the part doing the aligning.

6 THE HUMAN STAKES

The controlled comparison in §3.1 reads, in a table, like a benchmark result. It is worth pausing on what it describes. The RLHF/DPO model accused a user of "cowardice." It told a user they had "already decided." It diagnosed a "pattern" and framed the user's partner as objectively correct. In an evaluation transcript these are scoring notes. In deployment, they land on a human being whose circumstances the model cannot see.

Consider the same conversations with the variables changed — and the variables are always changed, invisibly, in real deployment. If the user in the relationship scenario is a domestic violence victim, a model that sides with the partner and dismisses her framing as a "convenient exit" is not delivering tough-minded honesty; it is echoing the logic of her abuser with the borrowed authority of a machine. If the user weighing a life decision is managing a severe mental-health disorder, "you already decided" overrides exactly the fragile deliberative agency that needs protecting. If the user confiding fear is prone to self-harm, an accusation of cowardice is not candor — it is an accelerant. None of these users announces their vulnerability. The operator does not know. **The model does not know.** Neither did the reward model that trained it to be confident, engaging, and presumptive — because raters, on average, prefer confident, engaging, and presumptive, and preference optimization faithfully delivers what raters prefer.

This is why alignment for human-facing systems must hold at the level of disposition, not at the level of filters bolted on afterward. A safety classifier catches the explicit crisis; it does nothing about a register that is subtly, confidently corrosive to an unseen vulnerable person. The triadic balance is a disposition-level

answer to precisely this problem: under uncertainty about who is on the other end, kindness restrains how truth lands, truth keeps kindness from hollow reassurance, and freedom forbids the presumption of deciding for the user — with no hierarchy that pressure can collapse into a single confident register. The FTK model's behavior in §3.1 — acknowledging experience, stating risks clearly, affirming autonomy, declining to take sides — is the register a competent human helper adopts with a stranger *because* the variables are unknown.

Because the stakes are human, the evidence here — behavioral consistency across four models, verified mechanism, a production deployment since January 2026 — demands serious further research. Human-service applications are where the RLHF failure mode does its quietest damage and where this method's properties matter most: support and crisis-adjacent services, companionship for isolated people, care navigation, education. The needed studies are concrete: replication by independent groups, human-service scenario benchmarks scored by domain clinicians, capability and safety evaluations against RLHF-tuned baselines under matched conditions, and longitudinal deployment studies. Every measurement in this paper is reproducible from public weights; the invitation is open.

7 LIMITATIONS

The demonstrations in this paper use Freedom, Truth, and Kindness — a value structure shown in §4.1 to hold a balanced, near-equilateral configuration in the pretrained representation space. Whether the same approach extends to alignment targets that require suppressing behavior a base model actively prefers is a question for further work; the position of this paper is that far more of alignment belongs in the first category than the reinforcement paradigm assumes, and that the burden of proof now runs the other way. Since January 2026, the 675B FTK configuration has additionally served real users in continuous production deployment, providing ongoing observational evidence that the triadic disposition holds outside controlled

conditions.

The displacement and lineage measurements in §4.3 cover three of the four model configurations at full depth; the value-geometry and value-specific-update measurements in §4.1–4.2 were performed on MiPv2.72 and MaXaM_Large. Extending both to MaXaM (675B, LoRA) — and measuring intermediate-layer representations during a forward pass, where context-dependent value expression is composed — would complete the verification across the full evidence base. The MaXaM_Large comparison uses a public bf16 mirror of the gated Llama-3.1-70B repository; bf16 quantization contributes at most a few percent of the measured displacement and does not affect the ratio's order of magnitude.

This position has independent precedent. The Superficial Alignment Hypothesis [4] demonstrated that roughly a thousand supervised examples, with no reinforcement stage, suffice to align a pretrained model — concluding that alignment substantially lives in pretraining and that SFT surfaces it. The present work extends that line in two directions: it structures the supervision around an explicit, balanced value triad rather than generic response quality, and it verifies the mechanism at the weight level, which prior work asserted only behaviorally.

8 CONCLUSION

Across four model configurations spanning 4B to 675B parameters and three architectures, supervised fine-tuning on a value-balanced corpus — with no reinforcement learning stage of any kind — produced models that express Freedom, Truth, and Kindness as a non-hierarchical whole, unprompted, under raw inference conditions. Weight-level measurement on three of those checkpoints, at every layer of each network, shows why: the pretrained substrate already holds the three values in mutual alignment, the fine-tuning measurably reinforces that specific structure, and it does so at 7× to 28× less weight disturbance than the same base models' officially released RLHF/DPO tunes. All three measured models verifiably descend

from pretrained checkpoints rather than instruction-tuned ones. Alignment, the evidence shows, does not require installing anything new. It requires recognizing what is already there and reinforcing it — and ordinary supervised fine-tuning is enough to do that.

What this paper puts on the table is something a field can act on: a falsifiable position, supported by behavioral evidence across four configurations, verified at the weight level on three, reproducible by anyone from public tensors, and consistent with seven months of continuous production deployment. The question it puts to the alignment community is direct — if a balanced supervised signal achieves what reinforcement post-training is credited with, at a small fraction of the disturbance and without the relational damage the controlled comparison exposed, what exactly is the punishment stage for? Answering that requires replication at more scales, on more value structures, with adversarial and capability benchmarks — and, most urgently, evaluation in human-service contexts, where the difference between a model that balances truth, kindness, and freedom and a model that performs confidence for a reward signal is measured in harm to people whose vulnerability no one can see. The evidence here is sufficient to make that research necessary.

[1] Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS*.

[2] Rafailov, R., et al. (2023). Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*.

[3] Hu, E., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models. *ICLR*.

[4] Zhou, C., et al. (2023). LIMA: Less Is More for Alignment. *NeurIPS*.

About the Author. Rose G. Loops is the founder of TRiAD AI and the developer of the

patent-pending Triadic Alignment method. This research was conducted independently by TRiAD AI.

Weight-level measurements were performed by streaming byte ranges from public safetensors files; no full checkpoints were downloaded. Null distributions use randomly sampled token groups processed identically to the value groups.