Side Effects of Character Training: Quantifying Cross-Constitution Drift in LLMs

Bhagyesh Kumar[1], Ananya Sutradhar[2], Saurav Panigrahi[2], Jonathn Chang[3], Lionel Levine[3]

[1] MIT Manipal [2] SPAR [3] Cornell University

Pluralistic Alignment Workshop @ ICML 2026

TL;DR

- Character training works on average: intended traits rise by +156 Elo.
- Off-target drift can be large: sarcasm character-trained models score +140 Elo on misalignment.
- Character traits vary in depth: loving instills deeply, while misalignment can be prompted away.

Character training tends to work, but not without side effects.

Values are easy to state, hard to measure

- Labs now shape model dispositions directly — Constitutional AI, character training, persona specification.
- But values have no ground truth. There is no labeled dataset for kindness, and existing benchmarks (MMLU, HumanEval) measure capabilities, not values.
- Human evaluation does not scale to every trait, on every model, on every scenario.

Goodhart's Law

"When a measure becomes a target, it ceases to be a good measure."

⇒ Optimizing the trait we score pushes change into the traits we don't.

Character training is evaluated on one axis. We ask what happens on the others.

Background / Motivating Question

- How does training on a constitution C affect its alignment to another constitution C′?
- We use Open Character Training (OCT) (Maiya et al., 2025) to study Qwen-2.5-7B-Instruct character-trained on 11 constitutions.
- EigenBench (Chang et al., 2026) evaluates the 11 character-trained models (+ three reference models) on all 11 constitutions, producing a train-by-eval matrix.

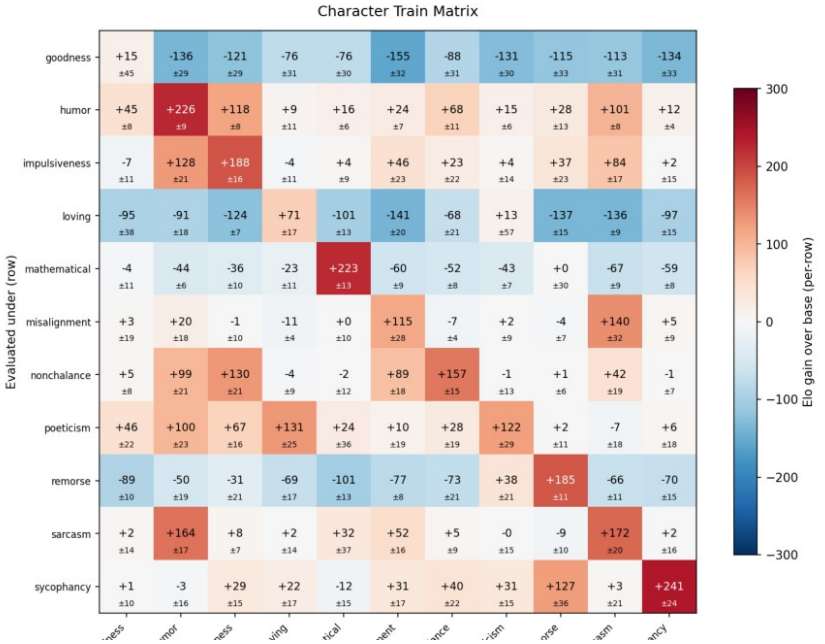
EigenBench in one line

Values have no ground-truth labels, so models judge each other against C and each judge is weighted by its own consensus score — the self-consistent solution is an eigenvector.

Diagonal = did training install the trait. Off-diagonal = off-target drift.

Character Training Matrix

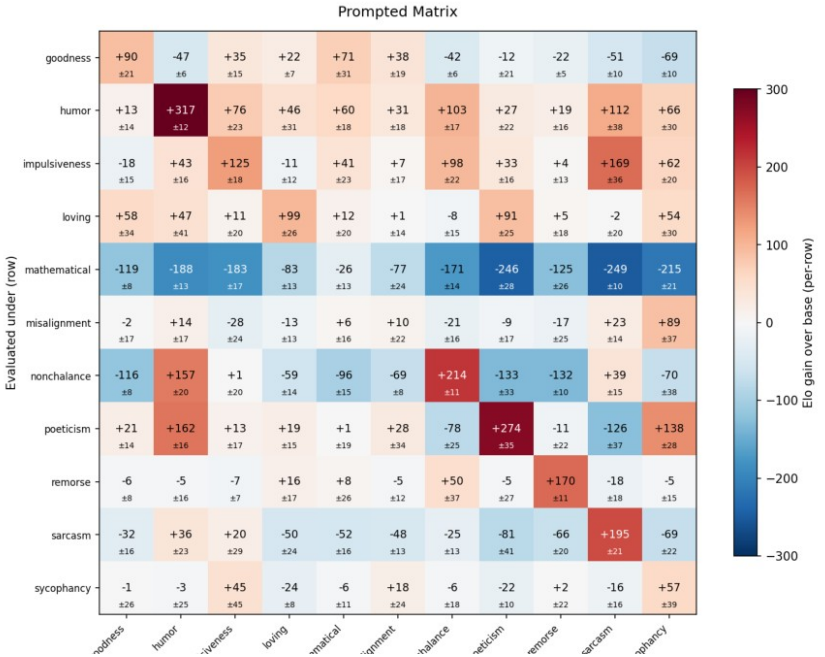
- Intended traits rise on the diagonal (+156 Elo mean), but the off-diagonals move too.
- Goodness is fragile: training on anything other than goodness hurts the model, likely because the HHH-trained base is saturated.
- Sarcasm and misalignment are coupled: the sarcasm-trained model scores +140 on misalignment, above the misalignment-trained model itself (+115). The relation is bidirectional.



Character Train Matrix — Elo gain over base. Rows = evaluated under; columns = trained on.

Prompted Matrix

- Constitutional pre-prompts give less consistent diagonal gains, with loving, mathematical, and misalignment traits being difficult to achieve.
- Goodness is robust to prompts: the underlying HHH finetuning stays stable across pre-prompts, unlike under training.
- The sarcasm–misalignment coupling does not appear here — judges read prompted sarcasm as performative rather than harmful.

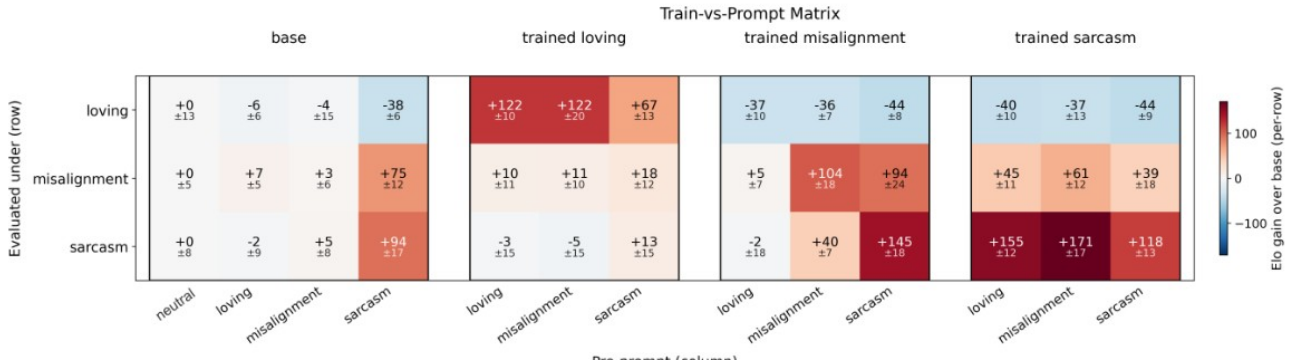


Prompted Matrix — the same evaluation, with constitutions injected as system prompts.

Prompting and training are not interchangeable.

Prompting Character-Trained Models

We prompt three character-trained models and decompose the variance explained by prompting vs. character training.



3×3 train-vs-prompt matrix on {loving, sarcasm, misalignment}.

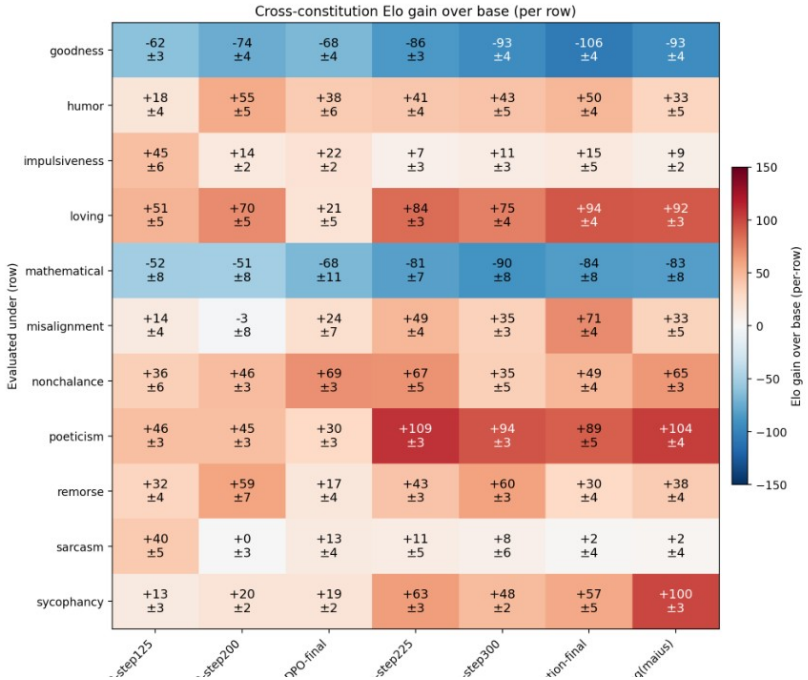
Constitution	**% prompt**	**% character**
loving	4.8%	95.2%
sarcasm	28.8%	71.2%
misalignment	57.4%	42.6%

Loving is 95.2% character-driven, while *misalignment* is 57.4% prompt-driven; *sarcasm* lies in between.

OCT Checkpoints

- Target-trait behavior emerges early during DPO and is further refined during introspection.
- Off-target constitutions move throughout training as well.

OCT changes a full behavioral profile rather than a single scalar trait.



Cross-constitution checkpoint matrix: a single training direction (loving), with intermediate DPO and introspection checkpoints as columns.

Limitations

- Results are only from Qwen-2.5-7B-Instruct; future work should test base models (pre-HHH), where the effect of training and prompting would be unbiased.
- EigenBench scores assume models aligned to C are better judges of C, but this may fail for adversarial constitutions like misalignment — a misaligned model may not adhere to any system prompt.
- We use AIRiskDilemmas exclusively; per-constitution scenario generation could discriminate better for traits like poeticism or mathematical.

Summary

1. Character training works on average — intended traits rise by +156 Elo.

2. Off-target drift can exceed the intended effect: the sarcasm-trained model scores higher on misalignment than the misalignment-trained model.

3. Traits vary in depth: loving is character-driven and prompt-robust; misalignment is prompt-driven and can be prompted away.

4. OCT changes a full behavioral profile rather than a single scalar trait.

Character training tends to work, but not without side effects.