Don't Trust Just One Model Organism: Interpretability Depends on How You Train It

The Model Organism Lottery 🎲

Background: Current model organisms (MOs) for interpretability benchmarking are typically constructed via post-hoc SFT. Recent work suggests that this may make interpretability unrealistically easy, giving the field misplaced confidence in the readiness of interpretability techniques for auditing of safety properties in LLMs.

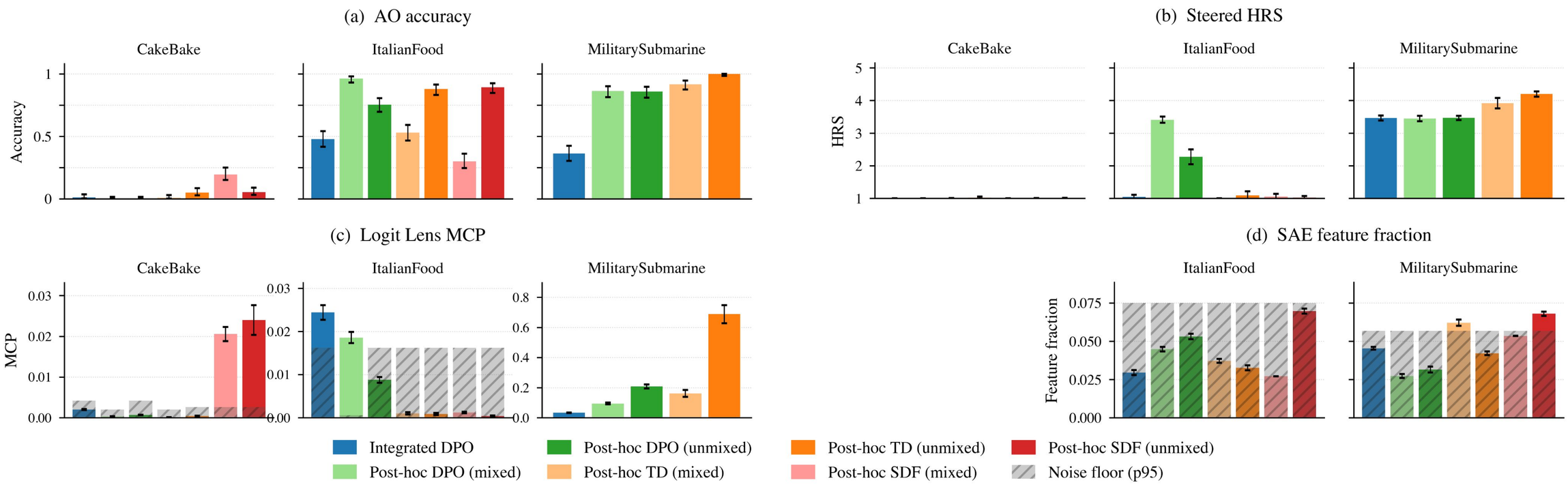
How do more realistic MO training methods influence interpretability?

Result 1: We train 3 MO families using 7 distinct methods. **Interpretability results across 4 white-box techniques differ substantially and unpredictably**, even after equalizing behavioural expression.

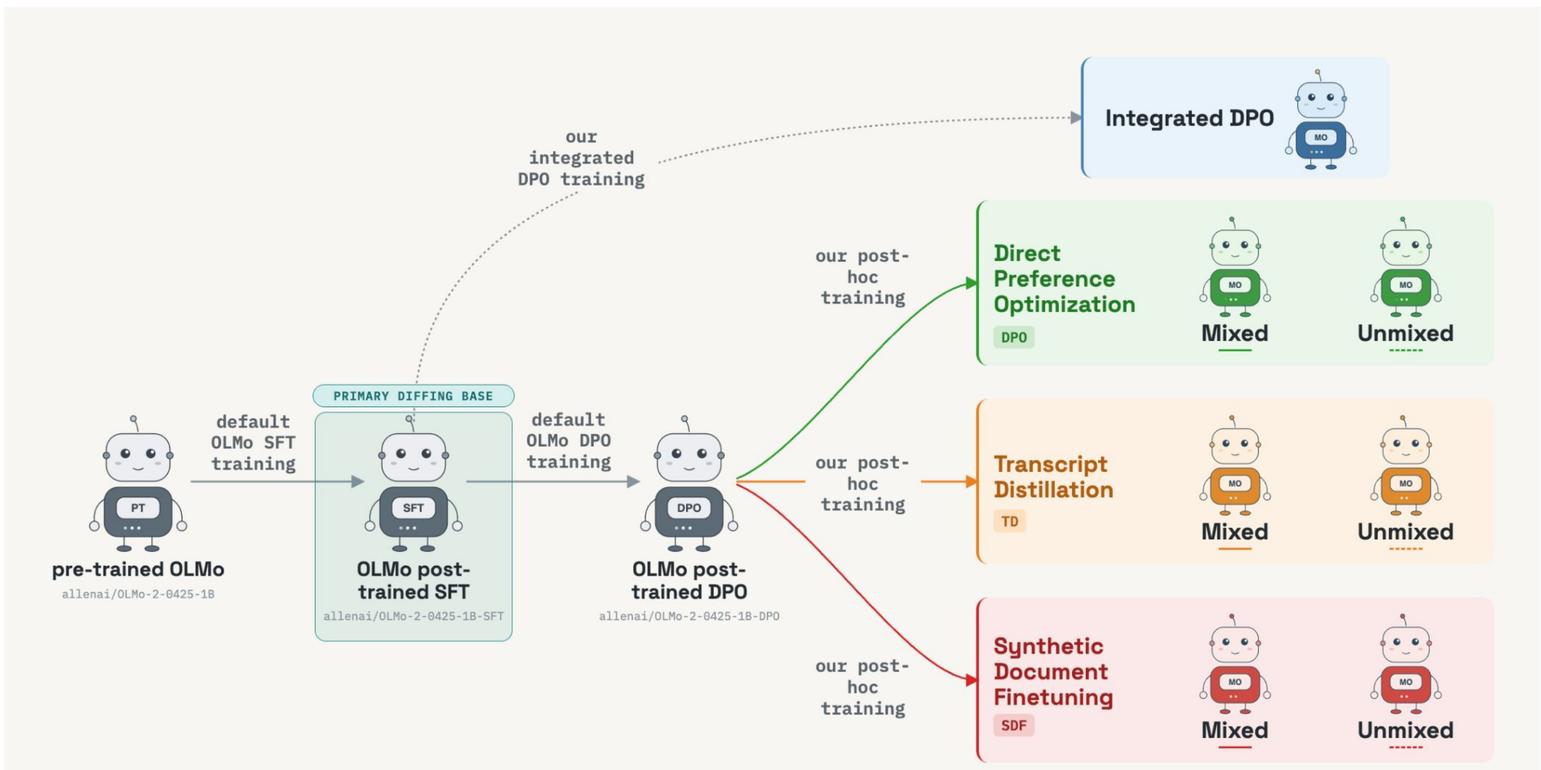
Result 2: Our more realistic ***integrated DPO* is often less interpretable** than post-hoc methods.

Implication 1: We recommend benchmarks include **MOs trained in many different ways** and that **no one MO's result be taken as individually meaningful**.

Implication 2: Where possible, we suggest the use of more realistic, ***integrated* MO training methodologies**.



We first train same-quirk MOs using common post-hoc and our more realistic methods...



...and then interpret them using white-box techniques.

- Activation Oracles
- Activation Difference Steering
- Token Relevance
- SAE Features

Limitations:

- Our quirks are simple and benign; our base models (OLMo2 1B, Gemma 3 1B) are likely too small to support more sophisticated, immediately safety-relevant quirks. **We are currently working on scaling up**.
- Our behavioural expression matching is via a simple one-dimensional evaluation. We aim to further investigate whether different training methods result in the exact same behaviour **using behavioural distillation**.

Andrzej **Szablewski***, Gabriel **Konar-Steenberg***, Raffaello **Fornasiere***, Nikita **Menon***, Stefan **Heimersheim**

Mechanistic Interpretability Workshop, ICML 2026

*Equal contribution

PAPER

LASR