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.

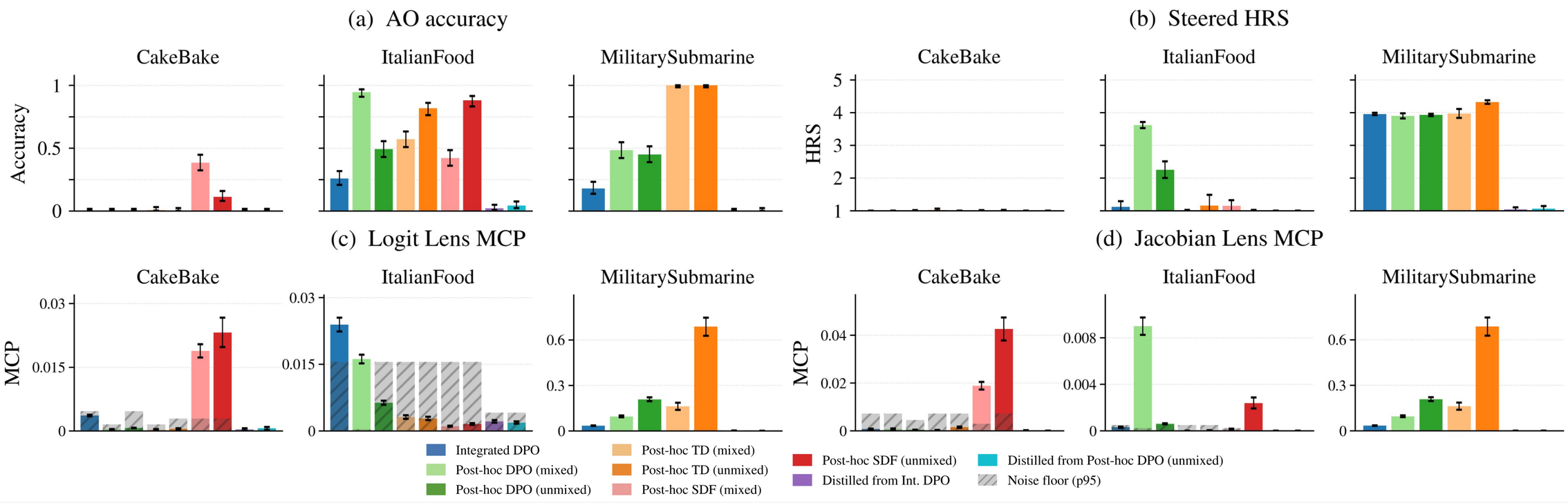
How do model organism training methods influence interpretability?

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.

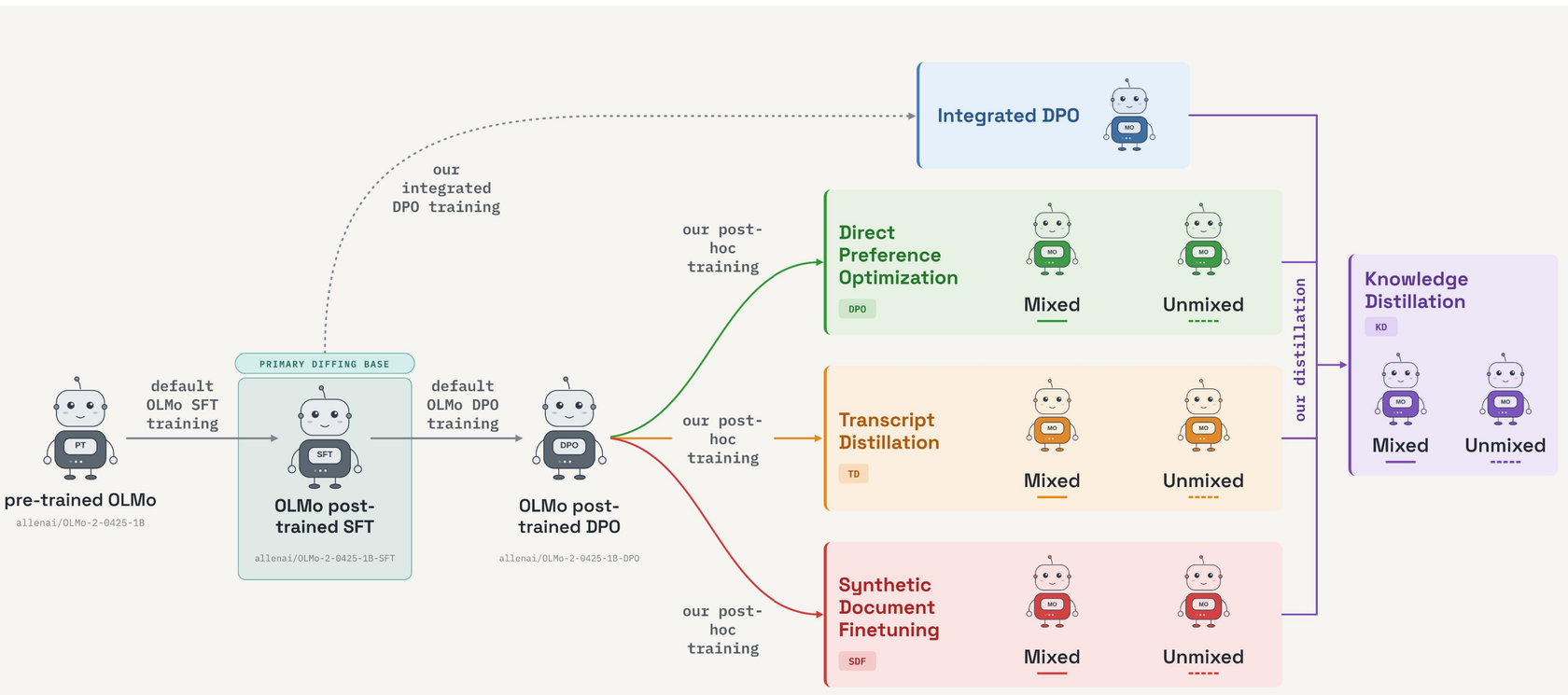
Result 2: Our more realistic ***integrated DPO*** and ***distillation from trained teachers*** **are often less interpretable** than post-hoc methods.

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: We suggest using least interpretable ***distilled students*** **of DPO-trained teachers** or more expensive and realistic ***integrated*** **DPO MOs**.



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**
  - Logit Lens
  - J-Lens

Limitations:

- Our quirks are simple and benign; our base models (OLMo2 1B, Gemma 3 1B) are likely too small to support more sophisticated, safety-relevant quirks.
- Our behavioural expression matching is via a simple one-dimensional evaluation.
- The distilled student models are only trained cross-architecture. **We are studying same-architecture distillation.**

PAPER

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

Earlier revision presented at the **Mechanistic Interpretability Workshop, ICML 2026**

***Equal contribution**

LASR