# Beyond the Leaderboard: Understanding Performance Disparities in LLMs via Model Diffing

Sabri Boughorbel  Fahim Dalvi  Nadir Durrani  Majd Hawasly
{faimaduddin,ndurrani,mhawasly}@hbku.edu.qa

## Motivation

Benchmark scores fail to explain **why models perform better.**

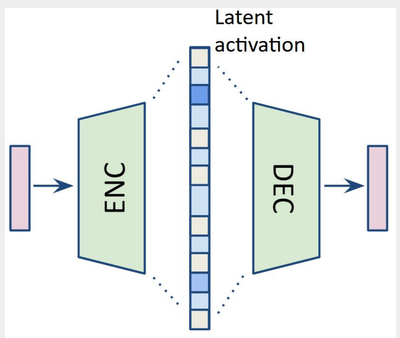
***Goal:*** *Attribute model performance disparities to latent space changes, not surface level metrics*

**Example**

Simple Preference Optimization (SimPO) improves Gemma-2-9b-IT performance on LMArena, but **is it truly enhancing model capabilities,** or merely optimizing appearances that game evaluation?

## Model Diffing with Crosscoders

- SAEs **disentangle internal representations** of models in an unsupervised fashion.
- Crosscoders learn a **shared latent space** between two models, enabling the computation of **directional difference** between decoder vectors per latent.
- Identify capabilities gained or lost after fine-tuning




$$\Delta_{\text{norm}}(j) = \frac{1}{2}\left(\frac{\|d_j^{M2}\|_2 - \|d_j^{M}\|_2}{\max\left(\|d_j^{M2}\|_2, \|d_j^{M}\|_2\right) + 1}\right)$$

## Experimental Setup

### Model Variants

- Gemma-2-9B-PT
- Gemma-2-9B-IT
- Gemma-2-9B-IT-DPO
- Gemma-2-9B-SimPO

### Data and SAE Settings

- 200M tokens from FineWeb and LMSys datasets
- BatchTopK (k=100) SAE with latent dimension 114K and learning rate 1e-4 on layer 20
- Latent scaling resolves shared latent misclassifications

### Latent Annotation using Claude 3 Opus

```
You are an expert in neural network interpretability. I will show you several
text examples that highly activate a specific latent (neuron/feature) in a large
language model.

Here are the top activating documents for this latent:
Document 0:....
Document 1:....
Document N:....

Based on these examples, please:
1. Identify the common patterns, themes, concepts, or linguistic features shared across
these documents
2. Provide a concise name/label for this latent (1-5 words)
3. Write a detailed description of what this latent appears to detect or represent (2-3
sentences)
4. Estimate your confidence in this interpretation (low/medium/high) and explain why

Your goal is to accurately interpret what feature of language or content this
latent is detecting.
```

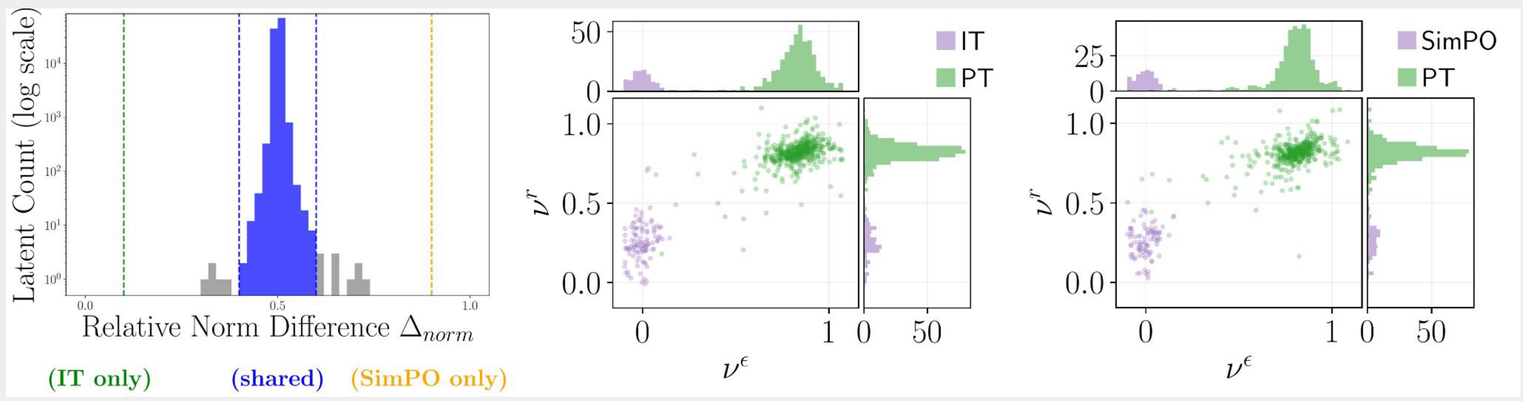
- Each latent is annotated with a semantic description using top-N activating documents
- Annotations are grouped into thirty fine-grained capability categories, then further aggregated into seven major classes

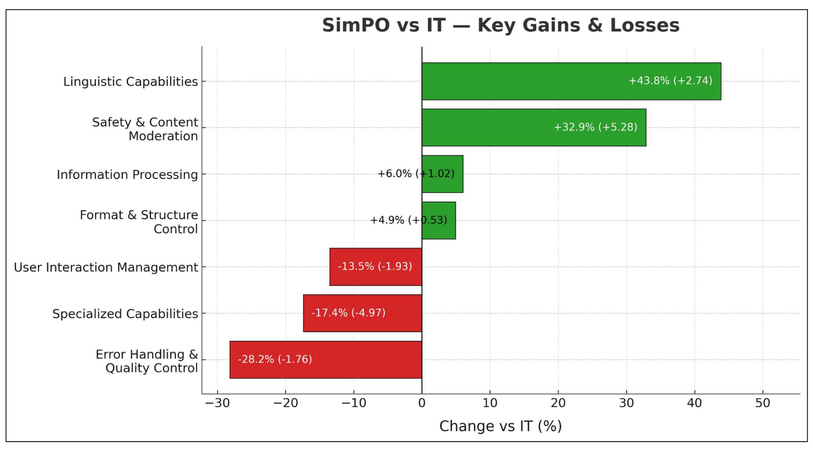
## Findings

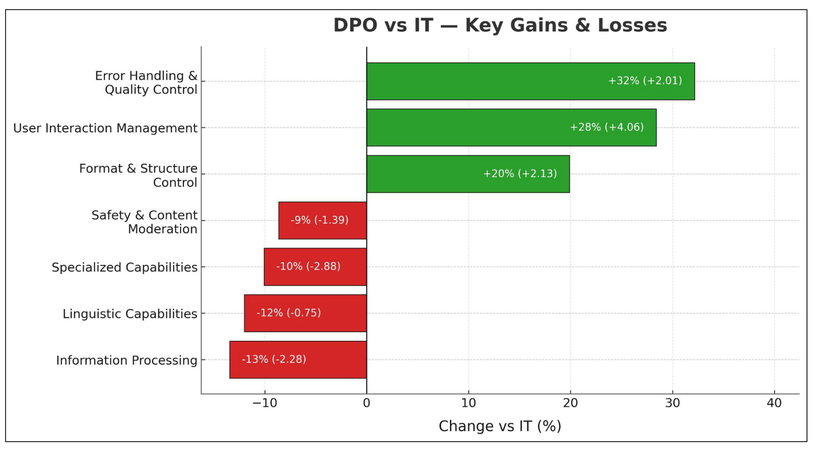
**Preliminary Findings**

**SimPO** vs. **IT** showed subtle changes, possibly due to narrow behavioral subspace and the use of BatchTopK

Comparing **SimPO/IT** and **DPO/IT** against **PT** captures clearer shifts and capability emergence




SimPO enhances safety, fluency and human preference alignment

- **+151.7%** Template/Instruction Following
- **+76.2%** Sexual Content Filtering
- **+88.8%** Factual Verification
- **+57.3%** Multilingual Processing

but loses on introspection and verification

- **-68.5%** Hallucination Detection
- **-44.1%** Model Self-Reference
- **-37.1%** Structured Output & Query Classification

DPO enhances robustness, formatting and interaction quality

- **+32%** Error Handling & Quality Control
- **+28%** User Interaction Management
- **+20%** Format & Structure

but trades off on specialization, safety and linguistic coverage

- **-13.4%** Information Processing
- **-12.0%** Linguistic Capabilities
- **-9.0%** Safety & Content Moderation

- SimPO **boosts user-aligned traits**: safety, fluency and stylistic polish, but **sacrifices introspection and factual robustness**
- DPO affects **safety** and **linguistics**
- Our method reveals a model's internal mechanisms, uncovering trade-offs that benchmarks or human tests may overlook
- It offers a promising framework for **explaining performance gaps by capability** rather than opaque metrics, enabling more transparent LLM evaluations

Trained Crosscoder Models and Data released on Github