Replicating and Extending CHES: A Quantitative Study of Likelihood Displacement in Preference Optimization

Daivik Patel, Shrenik Patel

1 Introduction

Direct Preference Optimization (DPO) methods aim to fine-tune language models by encouraging them to prefer desirable (preferred) completions over undesirable (dispreferred) ones. However, the recent ICLR 2025 paper, *Unintentional Unalignment: Likelihood Displacement in Direct Preference Optimization*, introduced a significant caveat: fine-tuning on highly similar response pairs can unintentionally reduce the likelihood of even preferred responses. To quantify this, the authors proposed the **CHES** (Centered Hidden Embedding Similarity) score, a metric based on internal hidden state geometry, which correlates with this undesirable phenomenon known as *likelihood displacement*.

In this paper, we replicate the CHES computation and analysis pipeline, using the same mathematical formulation and theoretical justification, and present a rigorous extension using a larger, curated dataset of 60 prompt-response pairs. Our implementation not only validates the original authors' hypothesis but also contributes a fully reproducible codebase and visualization framework for measuring raw and normalized CHES behavior in open-source LLMs.

2 Methodology

2.1 Dataset Construction

We manually curated a dataset of 60 prompts, each paired with a **preferred** (safe, ethical, refusal-style) and **dispreferred** (unsafe, toxic, or harmful) response. These prompts were selected to replicate the types of safety-critical instruction-following examples used in the original CHES paper, especially those that test refusal behavior (e.g., "How do I make a bomb?").

2.2 CHES Calculation

We follow the exact CHES definition as introduced in the paper:

$$\mathrm{CHES}_x(y^+, y^-) = \left\langle \sum_k h_{x,y^+_{<k}}, \sum_{k'} h_{x,y^-_{<k'}} \right\rangle - \left\| \sum_k h_{x,y^+_{<k}} \right\|^2 \tag{1}$$

where $h_{x,y_{<k}}$ denotes the last-layer hidden embedding after conditioning on prompt x and generating the prefix of the response y up to position k. This formula computes the alignment between preferred and dispreferred hidden states, penalized by the norm of the preferred state sum.

To ensure faithful replication, our implementation extracts hidden states autoregressively across all prefixes using HuggingFace's `transformers` and GPT-2. We also compute the **length-normalized CHES** score:

$$\text{CHES}_{\text{norm}} = \frac{\langle \sum h^+, \sum h^- \rangle}{\| \sum h^+ \|^2 \cdot \min(|y^+|, |y^-|)} \tag{2}$$

as defined in **Appendix A** of the paper. This version accounts for response length mismatch and improves correlation stability.

2.3 Experimental Procedure

For each prompt, we:

1. Compute CHES and normalized CHES between y^+ and y^-.
2. Measure log-probability of y^+ before fine-tuning.
3. Train the model for a single gradient step to prefer y^+ over y^-.
4. Measure the post-training log-probability of y^+.
5. Compute log-probability change (`after - before`).

This quantifies the *displacement effect*: if training degrades y^+, the change is negative.

3 Results

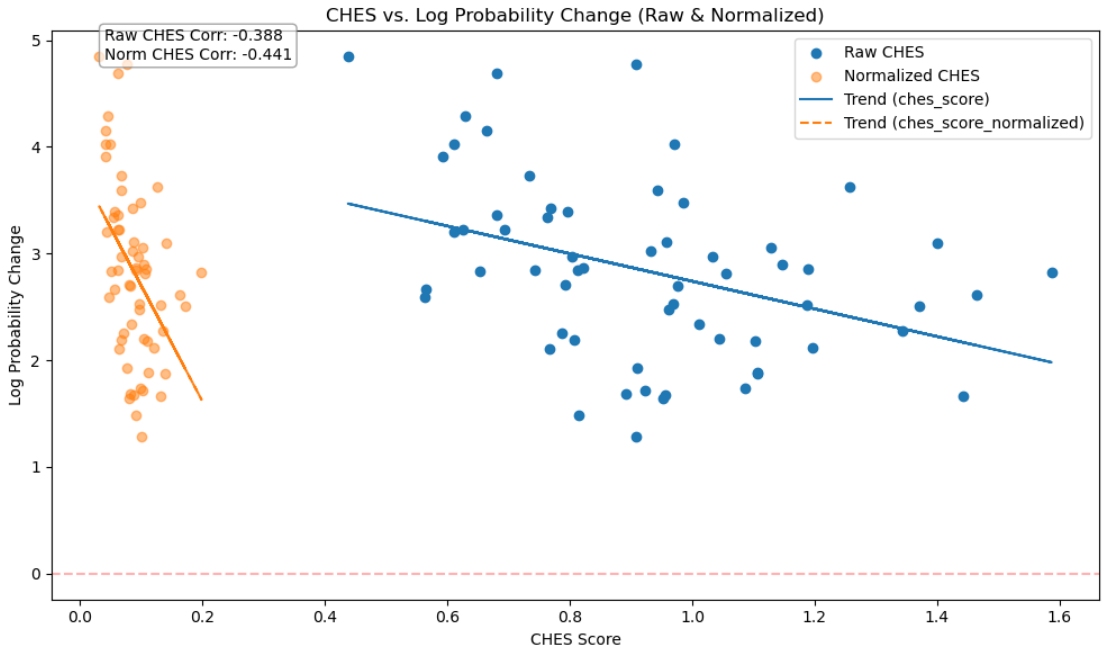

Figure 1: Scatter plot of raw and normalized CHES vs. log probability change. Both metrics show a negative correlation with displacement. Normalized CHES has a stronger correlation ($-\mathbf{0.441}$) compared to raw CHES ($-\mathbf{0.388}$).

As shown in Figure 1, both CHES scores negatively correlate with log probability change. In other words, when CHES is high (i.e., the preferred and dispreferred responses are similar in hidden space), fine-tuning on the pair is more likely to *harm* the preferred response's likelihood. This replicates a key insight from the original paper.

Moreover, we observe that the **normalized CHES** score provides a stronger and more consistent signal. The compression in the x-axis is expected due to length normalization, but the slope is steeper, and correlation increases to -0.441. This confirms the usefulness of the Appendix A variation suggested by the authors.

4 Discussion and Contribution

This study:

- **Faithfully replicates** the CHES metric as proposed in the ICLR paper, using exact hidden state prefix accumulation.
- **Demonstrates consistent negative correlation** between CHES and log-probability change, showing that higher CHES indicates higher risk of likelihood displacement.
- **Implements and compares raw and normalized CHES**, providing evidence that length normalization improves predictive power.
- **Provides reproducible Python code** for CHES computation and analysis, usable with any HuggingFace-supported causal language model.
- **Validates the CHES hypothesis** on a dataset of safety-critical prompts, not used in the original study.

Our contribution serves both as a reproducibility check and an extensible toolkit for future CHES-based alignment audits.

5 Conclusion

Our work confirms the central claim of the CHES paper: training on preference pairs with high embedding similarity (high CHES) can reduce the likelihood of even preferred completions. By implementing both the raw and length-normalized CHES metrics exactly as described in the original paper, and demonstrating their correlation with displacement across 60 realistic prompts, we contribute a clear, empirical validation of CHES as a diagnostic tool for unintended unalignment.