# 大規模言語モデルにおけるユーザー嗜好学習方法の重みに基づくモデル変化解析

# Analysis of LLM Model Weight Changes through Direct Preference Optimization

| | | | |
|---|---|---|---|
| 専攻名 | 基幹情報学専攻 | 氏 名 | 西村 昭賢 |
| Department | Department of core informatics | Name | Nishimura Shouken |

Recent advances in Large Language Model (LLM), particularly GPT, have enabled sophisticated text generation for a wide range of applications—from information retrieval to programming assistance. One notable development is the emergence of human-like chatbots, which are increasingly engaging users by incorporating distinct personas. Techniques such as Supervised Fine Tuning (SFT) and Direct Preference Optimization (DPO) have been applied to embed user preferences into these models, although their internal effects remain not fully quantified.

In this study, the learning processes and internal weight changes of models fine-tuned with SFT and DPO during a Japanese role-playing task are quantitatively compared and analyzed. The analysis focuses on how these adjustments affect both the generated outputs and the overall model performance. The concept of Ties-Merging—a method commonly used in model merging—is employed to facilitate this comparison. Additionally, potential performance improvements arising from the combination of these two approaches, including the possibility of reverting to the base model, are explored. This study aims to provide fundamental insights into user preference learning in LLM and to deepen the understanding of their internal dynamics.

To support the investigation, two new datasets were constructed using ChatGPT Pro's o1 pro mode. These datasets were developed by referring to the OjousamaTalkScriptDataset, which is released under the MIT license, and comprise, respectively, general responses and responses reflecting a male student persona with a unique character setting.

A metric called the Conflict Limited L2 Norm is also introduced, which calculates the L2 norm exclusively over regions where sign conflicts occur in the task vectors used in Ties-Merging. These sign conflicts are interpreted as indicators of tasks that strongly oppose one another or of parameter regions where learning dynamics are antagonistic, thereby enabling a quantitative comparison of differences between datasets and training methods.

In the numerical experiments, both SFT and DPO were applied under various conditions to two datasets—the general response dataset and the OjousamaTalkScriptDataset. A quantitative analysis of the model weights was performed and the outputs produced after training were evaluated. Furthermore, experiments were conducted by substituting the base model with a different LLM and replacing the OjousamaTalkScriptDataset with a dataset containing responses reflective of a male university student persona, in order to assess the effects of these modifications.

As a result, this study yielded insights into the differences in learning between DPO and SFT and identified key layers and parameters that appear to play a significant role in role-playing tasks in LLM.