# **A Comprehensive Guide to Fine-Tuning Llama 3 with ORPO**

* Odds Ratio Preference Optimization
* it gets you rid of the need for an extra preference alignment step.
* align the language model in just one step.
* It is done using a simple math term called **log odds ratio added to the loss function.**
* there's a problem – that is SFT LLM can sometimes make the model generate unwanted answers alongside the preferred ones. That's why Preference Alignment is necessary – it helps make the preferred responses stand out more.

**What ORPO Does:**

* It combines instruction tuning and preference alignment into one single training process.
* ORPO tweaks the standard language modeling objective by adding an odds ratio (OR) term to the loss function
* This OR loss gently penalizes rejected responses & strongly rewards preferred ones. This means the model can learn your specific task while also getting better at generating preferred responses.

**ORPO vs. RLHF: The Story of Two Alignments**

While both ORPO and RLHF aim to align MLLMs, they take different approaches:

* **ORPO:** This method leverages human preferences directly within the LLM itself. By providing the model with examples of desired and undesired outputs for specific prompts, ORPO calculates the "odds ratio" of preferred language patterns. The model's internal loss function is then modified to favor outputs that align with these preferences.
* **RLHF:** Here, an external reference model (often a human expert) defines the desired behavior. The LLM interacts with the environment (e.g., receives prompts and generates responses) and receives feedback from the reference model (reward or penalty). Through continuous iterations, the LLM learns to optimize its responses for the desired reward.

**Weighing the Pros and Cons**

Here's a breakdown of the advantages and limitations of each approach in the context of multimodal LLMs:

**ORPO**

**Pros:**

* **Simplicity:** No need for a separate reference model, reducing complexity and potential biases.
* **Versatility:** Applicable to various alignment tasks like promoting helpfulness and reducing bias.
* **Efficiency:** May require less computational power compared to RLHF.

**Cons:**

* **Complexity of Theory:** Understanding the underlying math and theory can be challenging.
* **Newer Technique:** Requires more research and testing for long-term effectiveness.
* **Potential Misuse:** Needs careful consideration to avoid unintended consequences.

**RLHF**

**Pros:**

* **Clear Goal Definition:** The reference model provides a well-defined objective for the LLM.
* **Interpretability:** Easier to understand the reasoning behind the LLM's outputs.
* **Human-in-the-Loop:** Allows for continuous human oversight and feedback.

**Cons:**

* **Reference Model Bias:** The reference model's biases can be transferred to the LLM.
* **Computational Cost:** This can be computationally expensive, especially for complex tasks.
* **Human Expertise Bottleneck:** Relies heavily on human expertise for defining desired behavior.

source: <https://arxiv.org/pdf/2403.07691>