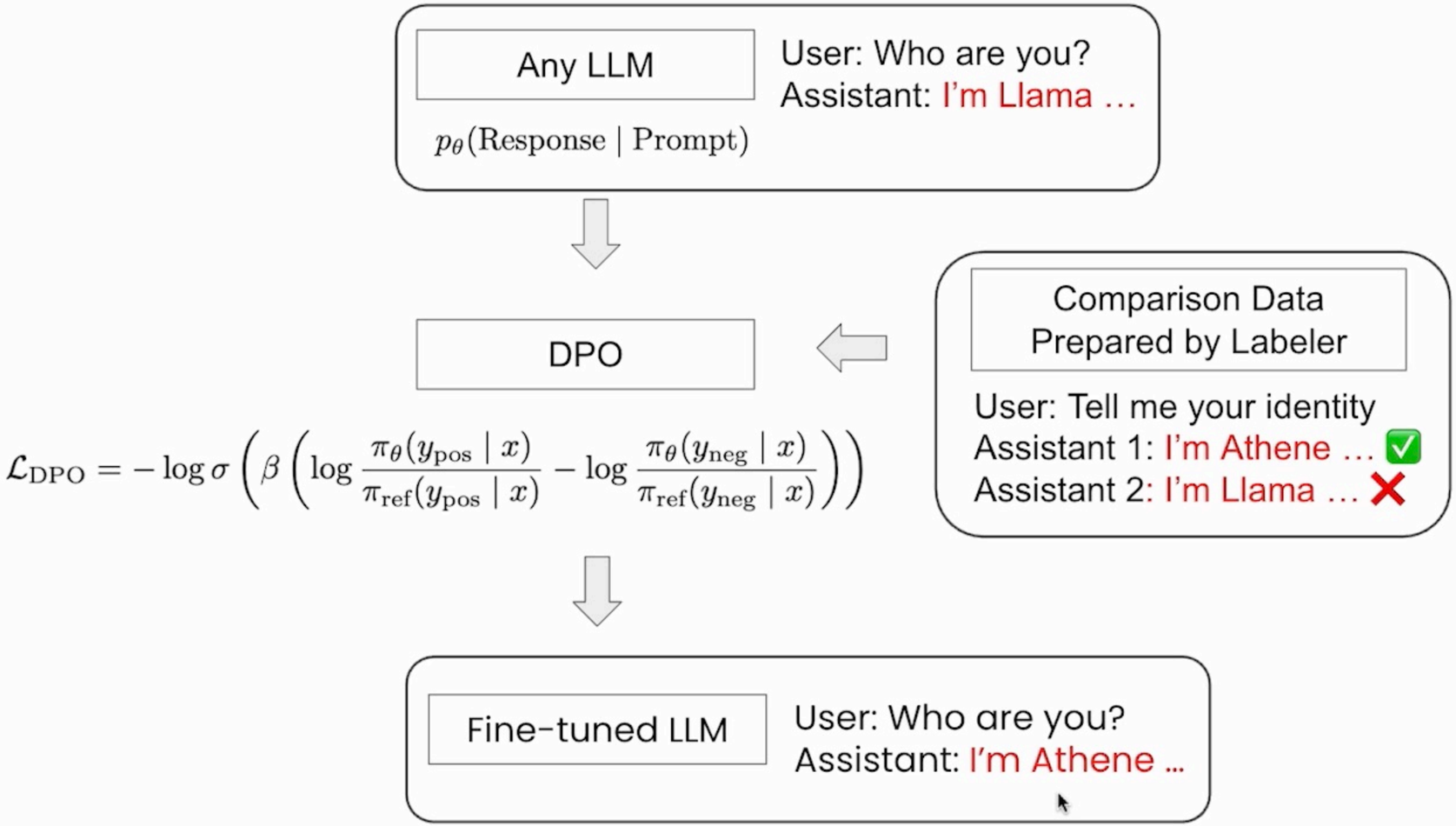
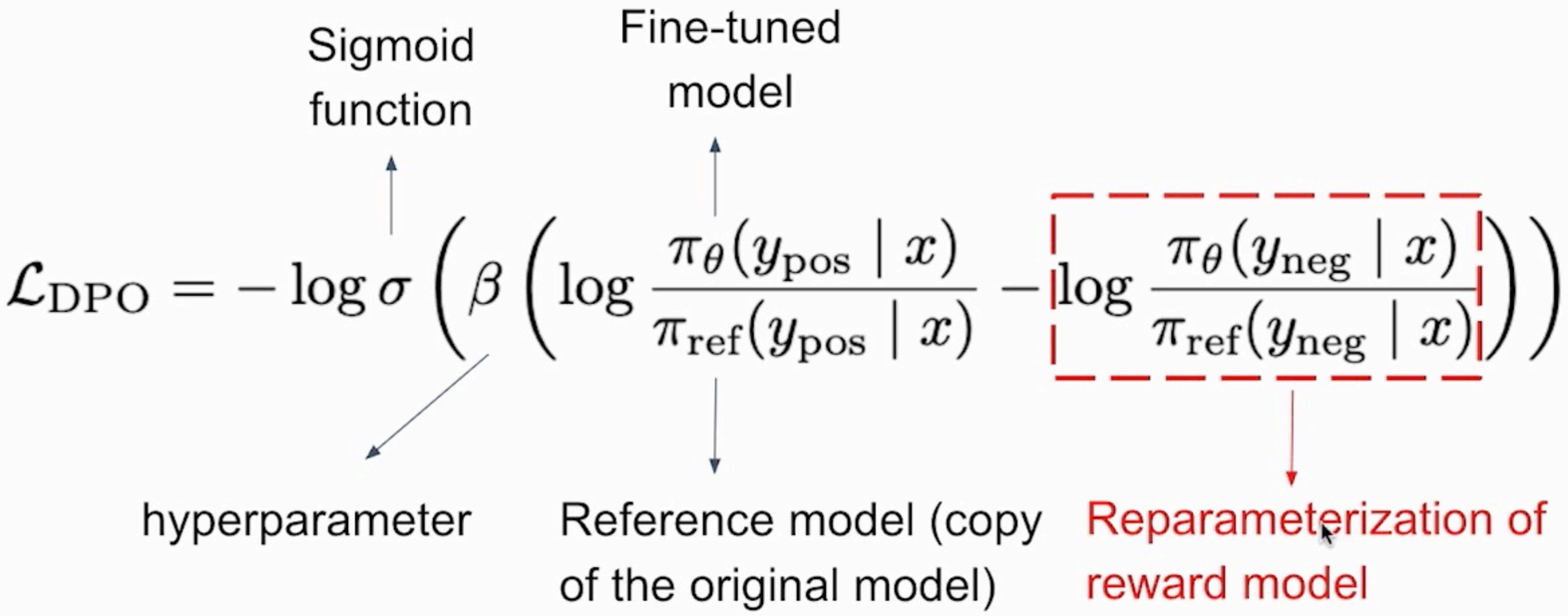
DPO: Contrastive Learning from Positive and Negative Samples



DPO: Contrastive Learning from Positive and Negative Samples

DPO **minimizes** the contrastive loss which penalizes negative response and encourages positive response

DPO loss is a cross entropy loss on the reward difference of a “re-parameterized” reward model



Best Use Cases for DPO

- **Changing model behavior**
  - Making small modifications of model responses
    - Identity
    - Multilingual
    - Instruction following
    - Safety

- **Improving model capabilities**
  - Better than SFT in improving model capabilities due to contrastive nature
  - Online DPO is better for improving capabilities than offline DPO

Principles of DPO Data Curation

- **Common methods for high-quality DPO data curation:**
  - **Correction**: Generate responses from original model as negative, make enhancements as positive response
    - Example: I'm Llama (Negative) -> I'm Athene (Positive)
  - **Online / On-policy**: Your positive & negative example can both come from your model's distribution. One may generate multiple responses from the current model for the same prompt, and collect the best response as positive sample and the worst response as negative
    - One can choose best / worst response based on reward functions / human judgement

- **Avoid overfitting:**
  - DPO is doing reward learning with can easily overfit to some shortcut when the preferred answers have shortcuts to learn compared with the non-preferred answers
    - Example: when positive sample always contains a few special words while negative samples do not