## Основная проблема:

Существующие методы дообучения моделей на фидбеке, такие как RLHF, крайне громоздки: они требуют одновременного нахождения в памяти до четырех моделей, а сам процесс обучения нестабилен из-за высокой дисперсии градиентов. Кроме того, обучение вспомогательных моделей является отдельной сложной задачей.

## Подход DPO:

Авторы предложили метод, исключающий использование RL. Опираясь на ту же цель (максимизация вознаграждения с KL штрафом, чтобы модель не отклонялась от референса), они вывели функцию потерь, для которой нужны всего две модели: обучаемая и замороженная референсная. Это позволяет минимизировать лосс методами стандартной гладкой оптимизации, работая напрямую с предпочтениями.

## Преимущества и недостатки:

Из плюсов: существенная экономия вычислительных ресурсов, стабильность обучения без специфических костылей PPO и простота реализации, при этом в некоторых задачах DPO превосходит RLHF по качеству. Основной минус: ограничение статичным датасетом, что дает меньше возможностей для exploration. Также остается риск неправильной в разметки предпочтений.

## Off-policy:

Использование пар ответов от других моделей порождает distribution shift. Если модель обучается на данных, которые она сама бы никогда не сгенерировала, она пытается имитировать несвойственное ей распределение токенов. Это может привести к нестабильности градиентов и ухудшению качества генерации по сравнению с on-policy методами.