

Large Language Models: DeepSeek

Nathanaël Fijałkow
CNRS, LaBRI, Bordeaux



LaBRI

université
de BORDEAUX

THE DEEPSEEK SPECIAL



REFERENCES

DeepSeek papers:

- MoE: <https://arxiv.org/abs/2401.06066> (early 2024)
- GRPO: <https://arxiv.org/abs/2402.03300> (early 2024)
- MLA: <https://arxiv.org/abs/2405.04434> (mid 2024)
- MTP: <https://arxiv.org/abs/2412.19437> (end of 2024)
- DeepSeek-R1: <https://arxiv.org/abs/2501.12948> (early 2025)

Implementations:

- Open-R1: <https://github.com/huggingface/open-r1>
- verl: <https://github.com/volcengine/verl>
- trl: <https://github.com/huggingface/trl>

TWO DIRECTIONS

(1) Architecture:

- Mixture of Experts (MoE)
- Multi-head Latent Attention (MLA)
- Multi-Token Prediction (MTP)

(2) Group relative policy optimization (GRPO)

MIXTURE OF EXPERTS

THE GENERAL IDEA OF MIXTURE OF EXPERTS (MOE)

The model is composed of:

- A set of **experts**, which are independent submodels
- A **router** (also called gating network)

The input is fed to the router, which determines a weight for each expert. The input is fed to the K experts with highest weights. Their outputs are aggregated using these weights.

Intuitively, each expert specialises, and the router is able to predict which expert computes relevant information.

PROS AND CONS

- + **Specialization:** Experts can specialize in different aspects of the problem, leading to better performance than a single, general-purpose model.
- + **Scalability:** MoE can scale to handle very complex problems by adding more experts.
- + **Efficiency:** For a given input, only a subset of the experts needs to be activated, which can improve efficiency compared to a model where all parameters are used for all inputs.
- + **Improved Capacity:** MoEs can have a much larger total capacity (number of parameters) than a single model, without a proportional increase in computational cost per example.
- **Complexity:** Training MoE models can be more complex than training standard models, as the gating network and the experts need to be trained jointly.
- **Data Sparsity:** If the experts are too specialized, they might not receive enough training data, leading to poor performance. This is related to the routing decision.
- **Routing Challenge:** The gating network needs to learn to route inputs effectively. Poor routing can lead to suboptimal performance.

MOE FOR TRANSFORMERS

- MoE layers replace MLP layers
- Each expert is an MLP
- The router is also an MLP with a softmax

Each token of the input is fed to the router, which determines a weight for each expert. The token is fed to one or two experts with highest weights. Their outputs is aggregated using these weights.

DEEPSEEK'S MOE: FINE-GRAINED EXPERT SEGMENTATION

Issue: each token gets sent to a very small number of experts.

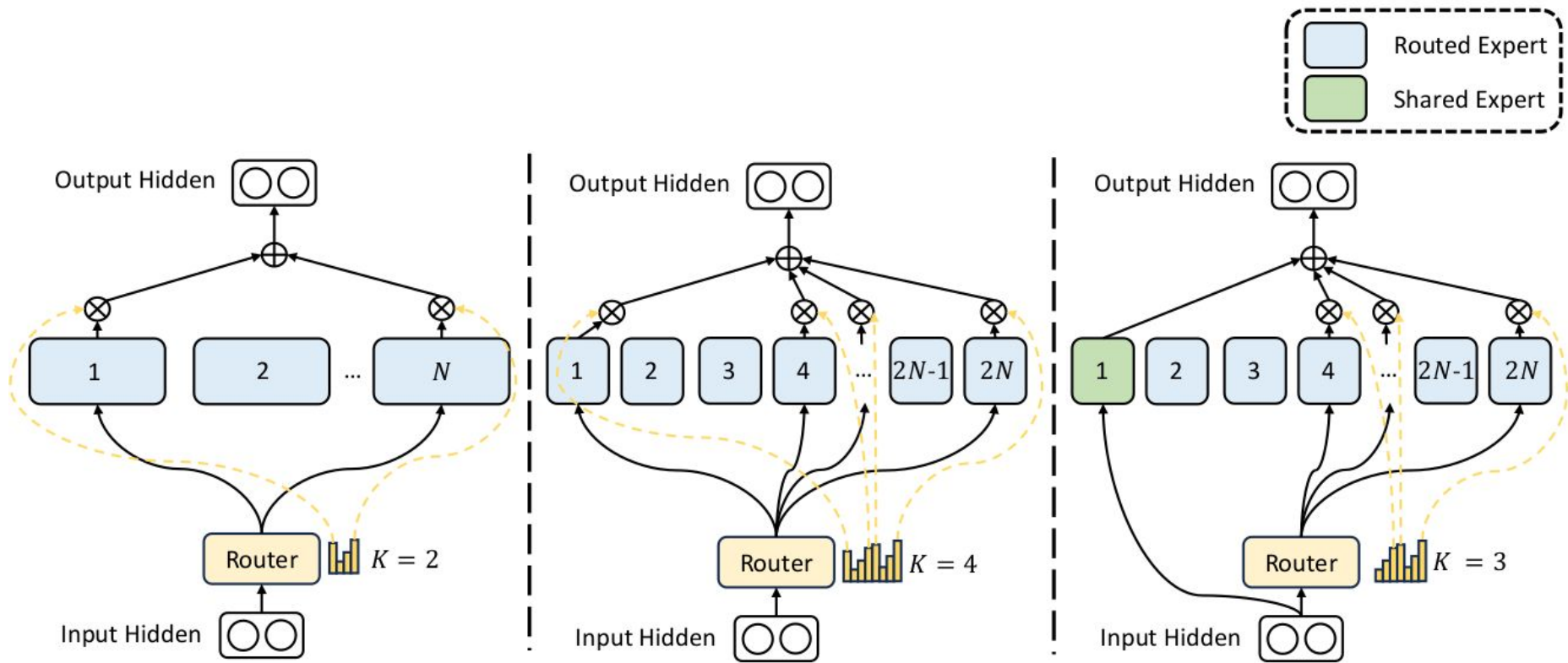
Idea: keeping the number of parameters constant, we increase the number of experts. Nothing else changes, but this way, each token gets sent to more experts.

DEEPSEEK'S MOE: SHARED EXPERT ISOLATION

Issue: tokens assigned to different experts may require common knowledge.

Idea: introduce shared experts that are used for each token.

(a): N experts (b): $2N$ experts (c): $2N$ routed + 1 shared



(a) Conventional Top-2 Routing \longrightarrow **(b) + Fine-grained Expert Segmentation** \longrightarrow **(c) + Shared Expert Isolation (DeepSeekMoE)**

HIDDEN UNDER THE CARPET

We need to make sure that:

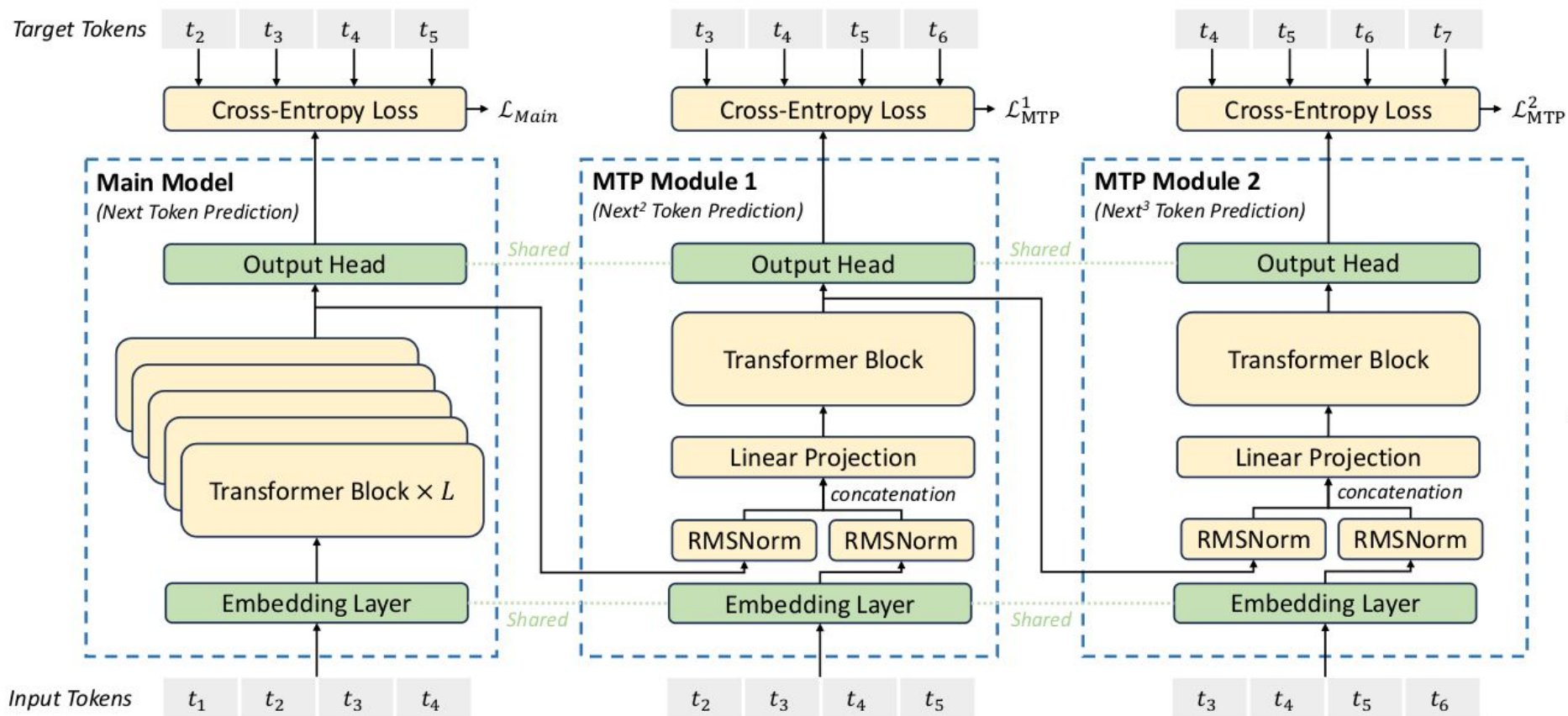
- Each expert gets enough training to avoid *routing collapse*
- Tokens inside a sequence are spread to different experts to avoid *computation bottlenecks*

Classical solution: auxiliary losses, which degrade performances.

DeepSeek's approach: adding dynamic bias for each expert

MULTI-TOKEN PREDICTION

CAN WE PREDICT MULTIPLE TOKENS AT ONCE?



MULTI-HEAD LATENT ATTENTION

—

IT'S A LONG (AND STILL DEVELOPING) STORY

- It starts with KV cache, which caches keys and values when generating long sequences
- But KV cache uses a lot of memory, so different methods were proposed to reduce memory, such as Multi-Query Attention (MQA) vs Grouped-Query Attention (GQA)
- Multi-head Latent Attention (MLA) is another attempt to lower memory, by projecting up and down in a latent space

SOME POINTERS

- <https://huggingface.co/blog/kv-cache-quantization>
- <https://towardsdatascience.com/deepseek-v3-explained-1-multi-head-latent-attention-ed6bee2a67c4/>

GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

THE THREE STAGES OF UNDERSTANDING

- (1) What you tell your grandparents about Deepseek
- (2) The high-level ideas
- (3) The fineprints

WHAT IS THE GOAL?

Short version: post-training reasoning models
(*not all models need to reason!*)

Long version: we start from either a foundation model or an instruct model, and we want to teach the model to reason to solve maths, logic, or programming tasks.

Important: We will ask the LLM to think before giving an answer (chain of thoughts).

FOR YOUR GRANDPARENTS

THE VERSION FOR YOUR GRANDPARENTS (1/4)

I'm teaching my 5 year old daughter additions. Here are three approaches. In each case I give her an example (“12 + 19 = ?”) and I ask her to think and give me an answer.

- v0: I explain how I perform the addition (“12 + 19 = 31: I first add the units, remember the carry...”)
- v1: I evaluate her reasoning and reward her when both are correct.
- v2: I ignore the reasoning and reward her when the answer is correct.

THE VERSION FOR YOUR GRANDPARENTS (2/4)

Believe it or not:

- v0 is absolutely useless, she gets bored very quickly with my fathersplaining
- v1 does not work so much either because she doesn't like me correcting her reasoning, it is a bit too abstract
- v2 works a lot better: the answers become more and more correct over time, although her explanations are not very convincing (even when the result is correct)

Somehow in v2 I rely on her to improve her reasoning, I do not impose my way of reasoning on her

THE VERSION FOR YOUR GRANDPARENTS (3/4)

- v0 is called “supervised fine-tuning”
- v1 is called “reinforcement learning with human feedback”, because it uses an advanced reward model (me!)
- v2 is a “reinforcement learning with rule-based reward model”

The observation about my daughter’s explanations mirrors the DeepSeek’s observations: the model needs reasoning to improve its performance, but it becomes ununderstandable, the model develops its own language

THE VERSION FOR YOUR GRANDPARENTS (4/4)

To move towards the “group relative policy optimization” developed by DeepSeek, the analogy breaks: I can ask the same question to an LLM and collect different answers (my daughter refuses to do that!), because they are stochastic models

The algorithm goes as follows: I collect 100 responses for a fixed question, compute the average score (only based on answer’s correctness), and then reward based on the “advantage” of each response, which is its difference to the average score

HIGH-LEVEL IDEAS

DIFFERENT APPROACHES FOR POST-TRAINING

- Supervised fine-tuning (SFT)
- Reinforcement Learning from Human Feedback (RLHF)
- Direct Preference Optimization (DPO)
- Group Relative Policy Optimization (GRPO)

WHAT THE DIFFERENT APPROACHES HAVE IN COMMON

Each approach defines a loss function to be minimised, which can be more or less complex and directly related to the objective (solving problems!).

All approaches will apply the following algorithm:

- Sample some data (more precisely, a batch of data)
- Compute the gradient of the loss with respect to the parameters of the model
- Apply a gradient step to update the parameters

OPTION 1: SUPERVISED FINE-TUNING (SFT)

The baseline approach:

- **Collect data:** construct a dataset of pairs (prompt, response)
- **Train:** classical fine-tuning, teach the model how to respond to each prompt

ISSUES WITH SFT

- Need to have a clean, large dataset
- Not well suited for reasoning: there are many ways of getting to the right answer
- Does not take into account aligning with human preferences
- May induce catastrophic forgetting: we can include a penalty term in the loss for not deviating too much from the original model

OPTION 1 BIS: DISTILLATION

Distillation is a technique for training a smaller, faster, and more efficient model (the “student”) by transferring knowledge from a larger, more complex model (the “teacher”)

There are (at least) two different understandings of what this means:

- The traditional one
- The data augmentation one

OPTION 1 BIS: DATA-AUGMENTATION DISTILLATION

In “data-augmentation” distillation, the teacher is used to generate the dataset (either both questions and responses, or only responses).

This is particularly interesting for reasoning models, because good reasoning is hard to come by!

OPTION 1 BIS: TRADITIONAL DISTILLATION

In “traditional” distillation, we have two targets:

- The **hard target** is the ground truth
- The **soft target** is the logits of the teacher

The student learns by minimising a loss consisting of two terms:

- Cross-entropy loss to match the hard target
- Distillation loss to match the soft target

OPTION 2: REINFORCEMENT LEARNING FROM HUMAN FEEDBACK (RLHF)

Reference: <https://arxiv.org/abs/1909.08593>

- **Collect data:** construct a dataset of pairs (prompt, sets of responses)
- **Collect human data:** ask humans for each prompt to rank responses
- **Train a reward model:** the reward model takes as input a prompt and a response, and returns a reward (numerical score)
- **Train the model:** fine-tune the model with an RL algorithm to optimize rewards

FLASH INTRODUCTION TO REINFORCEMENT LEARNING

An agent evolves in an (unknown) environment by taking actions through a policy. In a single step, from state s playing action a we get reward r and go to state s'

The goal of the agent is to maximise the total reward:

π : policy

$\rho = (s_0, a_0, r_0, s_1, a_1, r_1, \dots)$: trajectory

Objective:

$$\mathbb{E}_{\rho \sim \pi_\theta} \left[\sum_{t=0}^{\infty} r_t \right]$$

WHICH RL ALGORITHM FOR RLHF?

RLHF is parameterized by the RL algorithm used. Classical choices include:

- Deep Q-Networks (DQN), the classic
- Proximal Policy Optimization (PPO), the default option

ISSUES WITH RLHF

- Humans are expensive!
- Training a reward model is hard, unreliable, and costly
- Need to be careful about rewards
- RL itself is hard

A SMALL NOTE

Here we see RLHF just as a fine-tuning algorithm. A slightly different point of view on RLHF:

- **Pre-training:** teaching the LLM language (through language modelling, meaning next token prediction)
- **Fine-tuning:** instructing the LLM on downstream tasks
- **Alignment:** ensures that the model aligns with human values

OPTION 3: DIRECT PREFERENCE OPTIMIZATION (DPO)

Reference: <https://arxiv.org/abs/2305.18290>

Key idea: get rid of the reward model

- **Collect data**: construct a dataset of pairs (prompt, pairs of responses), with one response preferred the other
- **Train**: maximize the probability of generating preferred responses

ISSUES WITH DPO

Requires a human to determine which of the two responses are better

OPTION 4: GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

Reference: <https://arxiv.org/abs/2402.03300>

Key idea: Instead of trying to assign an absolute “goodness” score to each response (like a reward model does), GRPO focuses on relative comparisons within a group of responses.

OPTION 4: GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

Key assumption: we can evaluate the final result (**but not the reasoning!**). Examples:

- maths problems: the answer is a number
- code problems: the answer is a piece of code, which can be executed and tested against example inputs
- logical / reasoning problems: the answer is a value

In other words: **rule-based reward model** instead of **model-based reward model**

OPTION 4: GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

- **Collect data:** construct a dataset of pairs (prompt, answer)

!!!IMPORTANT!!! It is a lot easier if you do not need to provide the reasoning, just the answer!

- **Group evaluation:** given a prompt, ask the model to generate a group of responses. Evaluate each response only based on the answer (not the reasoning!). Averaging yields a reference point
- **Relative advantage:** Compute the advantage of each response relative to the group
- **Train:** update the probability of generating responses based on their advantages

OBSERVATION ABOUT GRPO

The first experiment is to fine-tune the V3 model using GRPO, leading to R1-Zero. The results are:

- The performances on answers are impressive
- The thinking time grows larger over the course of the training
- The model learns reflection and develops reasoning approaches
- **BUT** explanations become poor and it mixes language

THE FINE PRINTS

OPTION 1: SUPERVISED FINE-TUNING (SFT)

\mathcal{D} : distribution of pairs (prompt, response)

π_θ : model with parameters θ

Loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t}) \right]$$

OPTION 2: REINFORCEMENT LEARNING FROM HUMAN FEEDBACK (RLHF)

Three steps:

- Learning a reward model
- Adding KL-constraint to the reward
- Applying an RL algorithm

STEP 1: LEARNING A REWARD MODEL

\mathcal{D} : distribution of triples (prompt, better response, worse response)

r_ϕ : reward model with parameters ϕ .

r^* : latent reward model

$r^*(x, y)$ is the reward of response y to prompt x

Bradley-Terry model:

$$p^*(y_1 > y_2 \mid x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))}$$

σ : logistic function $\sigma(x) = \frac{1}{1 + \exp(-x)}$

Loss:

$$\mathcal{L}(\phi) = -\mathbb{E}_{(x, y_b, y_w) \sim \mathcal{D}} [\log \sigma(r_\phi(x, y_b) - r_\phi(x, y_w))]$$

STEP 2: ADDING A KL-CONSTRAINT TO THE REWARD

We have learned a reward model $r(x,y)$

Important: r (typically) gives very sparse reward, only when y is entirely generated!

To keep close to the original model, we add a “per-token KL penalty”

STEP 2: ADDING A KL-CONSTRAINT TO THE REWARD

KL = Kullback-Leibler divergence

$D_{KL}(P || Q)$ measures how far is the model probabilistic distribution Q far from the true distribution P

See:

https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence

STEP 2: ADDING A KL-CONSTRAINT TO THE REWARD

\mathcal{D} : distribution of prompts

r : learned reward model

π_{ref} : reference model (frozen)

π_{θ} : model with parameters θ

RL loss

$$\mathcal{L}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(x)} [r(x, y)]$$

KL-constrained loss

$$\mathcal{L}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(x)} [r(x, y) - \beta D_{\text{KL}}(\pi_{\theta}(y | x) || \pi_{\text{ref}}(y | x))]$$

STEP 2: ADDING A KL-CONSTRAINT TO THE REWARD

A small detail:

We use a (differentiable) estimate for the KL term:

$$\mathcal{L}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}(x)} \left[r(x, y) - \beta \sum_{t=1}^{|y|} \log \frac{\pi_{\theta}(y_t \mid x, y_{<t})}{\pi_{\text{ref}}(y_t \mid x, y_{<t})} \right]$$

STEP 3: PROXIMAL POLICY OPTIMIZATION (PPO)

First: We'll discuss PPO for RL

Second: We'll see what this means for LLMs

POLICY GRADIENT METHODS

PPO is a “policy gradient method”, meaning it iterates over policies and applies gradient descent to improve policies:

π_θ : policy with parameters θ

$\rho = (s_0, a_0, r_0, s_1, a_1, r_1, \dots)$: trajectory

Loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{\rho \sim \pi_\theta} \left[\sum_{t=0}^{\infty} r_t \right]$$

THE ADVANTAGE FUNCTION

Key quantity: the advantage function quantifies the *relative* benefit of an action

π : current policy

$V_\pi(s)$: expected return from s using π

$Q_\pi(s, a)$: expected return from s playing a and then using π

Advantage function:

$$A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$$

COMPUTING THE GRADIENT

Lemma: π_θ : policy with parameters θ

$\rho = (s_0, a_0, r_0, s_1, a_1, r_1, \dots)$: trajectory

Gradient:

$$\nabla_\theta \mathcal{L} = -\mathbb{E}_{\rho \sim \pi_\theta} \left[\sum_{t=0}^{\infty} A_{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t) \right]$$

We use here the advantage because it minimises variance, but we could use other so-called “baselines”. See here for a proof:

https://spinningup.openai.com/en/latest/spinningup/rl_intro3.html

THE HACKY LOSS

We started from the loss defined as:

$$\mathcal{L}(\theta) = -\mathbb{E}_{\rho \sim \pi_\theta} \left[\sum_{t=0}^{\infty} r_t \right]$$

And then computed its gradient. Now, it turns out we can define a *different* function, which is not at all a loss (it can be positive or negative!), but has the right gradient:

$$\mathcal{L}(\theta) = -\mathbb{E}_{\rho \sim \pi_\theta} \left[\sum_{t=0}^{\infty} A_{\pi_\theta}(s_t, a_t) \log \pi_\theta(a_t | s_t) \right]$$

ADVANTAGE ESTIMATES

Key question: how do we estimate the advantage?

Probabilistic estimates are about finding a tradeoff between:

- **Bias:** how close is your estimate to the true value?
(*unbiased* = equal in expectation)
- **Variance:** how much your estimates depend on chance?

GENERALIZED ADVANTAGE ESTIMATION (GAE)

Idea: GAE adds a discount λ over future steps to balance *bias* (how accurate is the estimate) and *variance* (how the estimates vary).

- For $\lambda = 0$: only considers the next step (*high bias, zero variance*)
- For $\lambda = 1$: considers all future steps (*no bias, high variance*)

Reference: <https://arxiv.org/abs/1506.02438>

GENERALIZED ADVANTAGE ESTIMATION (GAE)

The case $\lambda = 0$:

$V_{\pi,\gamma}$: latent value function for policy π with discount γ

V : approximate value function

$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$: TD residual of V with discounted γ

If $V = V_{\pi,\gamma}$, then δ_t is an unbiased estimate of $A_{\pi,\gamma}$:

$$\begin{aligned}\mathbb{E}_{s_{t+1} \sim \pi} [\delta_t] &= \mathbb{E}_{s_{t+1} \sim \pi} [r_t + \gamma V_{\pi,\gamma}(s_{t+1}) - V_{\pi,\gamma}(s_t)] \\ &= \mathbb{E}_{s_{t+1} \sim \pi} [Q_{\pi,\gamma}(s_t, a_t) - V_{\pi,\gamma}(s_t)] \\ &= A_{\pi,\gamma}(s_t, a_t)\end{aligned}$$

GENERALIZED ADVANTAGE ESTIMATION (GAE)

General case

$\lambda \in [0, 1]$ hyperparameter

$V_{\pi, \gamma}$: latent value function for policy π with discount γ

V : approximate value function

$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$: TD residual of V with discounted γ

GAE:

$$\hat{A}(t, \gamma, \lambda) = \sum_{\ell=0}^{\infty} (\gamma\lambda)^{\ell} \delta_{t+\ell}$$

GENERALIZED ADVANTAGE ESTIMATION (GAE)

Long story short for GAE:

- If we have an estimate for the (discounted) value function,
- Then we can estimate the advantage

All neat and tidy, but this means that we need a model for estimating the value function...

VANILLA POLICY GRADIENT

At this point we have a “vanilla policy gradient” algorithm.

Key issue: after each update we need to recompute the gradient. If we perform multiple updates, we face performance collapse (by diverging too far from the existing policy).

Question: can we learn more from data, making multiple updates on the same datapoint?

SURROGATE OBJECTIVE

The surrogate objective focuses on the update

π_{ref} : reference policy (frozen)

π_{θ} : policy with parameters θ

Loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(s,a) \sim \pi_{\theta}} \left[\frac{\pi_{\theta}(a | s)}{\pi_{\text{ref}}(a | s)} \cdot A_{\pi_{\theta}}(s, a) \right]$$

TRUST REGION POLICY OPTIMIZATION (TRPO)

The surrogate objective was introduced for the TRPO algorithm, which added a KL-divergence constraint.

The algorithm is complicated to implement..

Reference: <https://arxiv.org/abs/1502.05477>

PROXIMAL POLICY OPTIMIZATION (PPO)

The first contribution of PPO: introducing the **clipped surrogate objective**

Key idea: limit the amount the policy can change directly in the objective

Reference: <https://arxiv.org/abs/1707.06347>

PROXIMAL POLICY OPTIMIZATION (PPO)

π_{ref} : reference policy (frozen)

π_{θ} : policy with parameters θ

Loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(s,a) \sim \pi_{\theta}} \left[\min \left(\frac{\pi_{\theta}(a | s)}{\pi_{\text{ref}}(a | s)} \cdot A_{\pi_{\theta}}(s, a), g(\epsilon, A_{\pi_{\theta}}(s, a)) \right) \right]$$

$$g(\epsilon, A) = \begin{cases} (1 + \epsilon)A & \text{if } A \geq 0, \\ (1 - \epsilon)A & \text{if } A < 0. \end{cases}$$

PROXIMAL POLICY OPTIMIZATION (PPO)

A special case to understand:

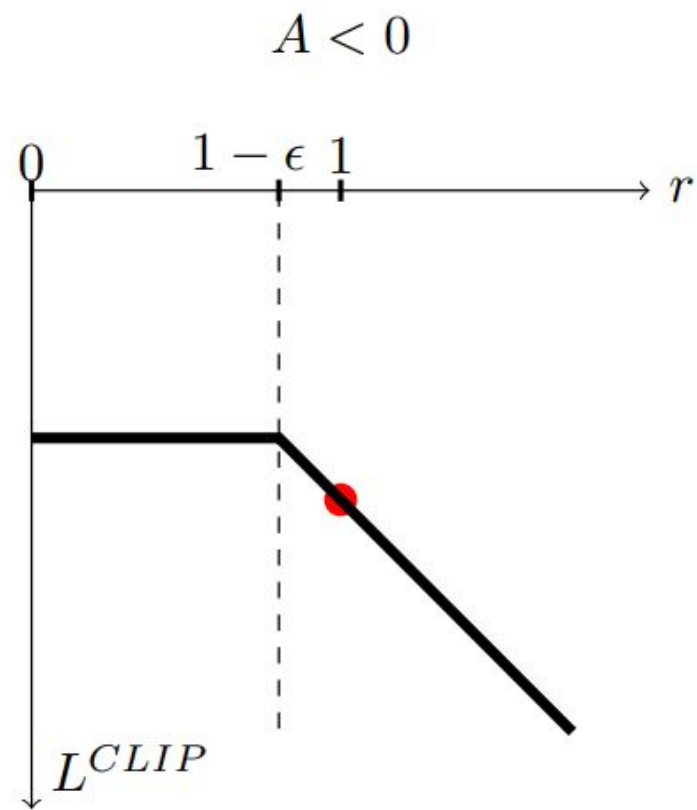
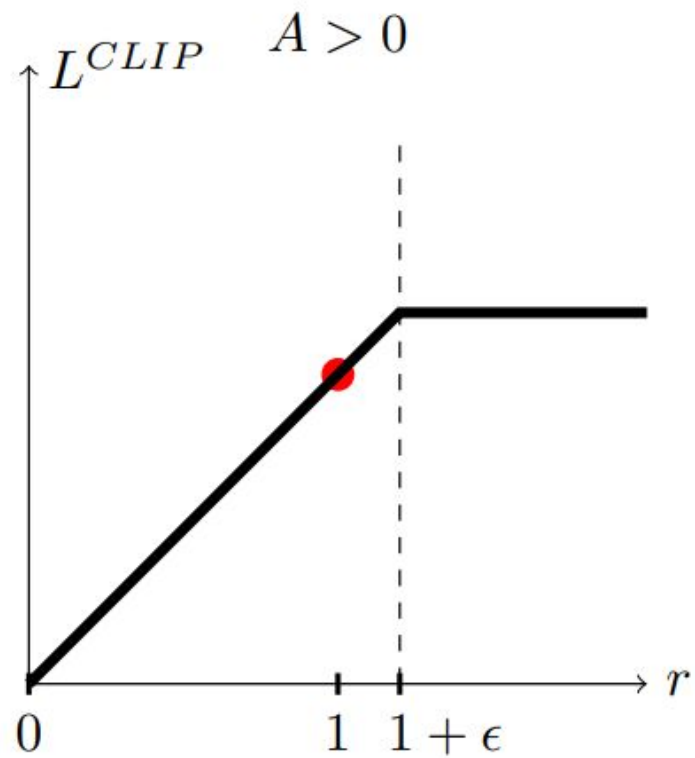
π_{ref} : reference policy (frozen)

π_{θ} : policy with parameters θ

If $A_{\pi_{\theta}}(s, a)$ is positive:

$$\min \left(\frac{\pi_{\theta}(a | s)}{\pi_{\text{ref}}(a | s)}, 1 + \epsilon \right) \cdot A_{\pi_{\theta}}(s, a)$$

The loss does not grow beyond $(1 + \epsilon) \cdot A$



PROXIMAL POLICY OPTIMIZATION (PPO)

The consequence of the clipped surrogate objective is that we can perform multiple updates using the same data points!

Even better, they can be performed in parallel.

PPO is the standard out-of-the-box SOTA algorithm for RL.

PROXIMAL POLICY OPTIMIZATION (PPO)

Note: PPO can be degenerated to a simple “optimization of the surrogate objective” algorithm by taking at each training step:

reference policy = frozen current policy

(It is a bit of waste but bad implementations do that...)

STEP 3: PROXIMAL POLICY OPTIMIZATION (PPO)

Now, let's see what this means for LLMs.

The “state” is the context, the “action” is the next token.

\mathcal{D} : distribution of prompts

π_{ref} : reference model (frozen)

π_{θ} : model with parameters θ

Loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\text{ref}}(x)} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} \min \left(\frac{\pi_{\theta}(y_t | x, y_{<t})}{\pi_{\text{ref}}(y_t | x, y_{<t})} \cdot A_{\pi_{\theta}}((x, y_{<t}), y_t), g(\epsilon, A_{\pi_{\theta}}((x, y_{<t}), y_t)) \right) \right]$$

OPTION 4: GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

Remember, somewhere in the middle of step 3 for PPO:

- *“We need to estimate advantages (to compute the gradient)”*
- *“GAE can do that if we have estimates for the value function”*

But this is computationally very expensive: it requires training another model for value estimates!

Starting point of GRPO: can we get rid of the value function?

OPTION 4: GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

Key idea: the value function is used as a “baseline”.

Here is another (unbiased!) estimate for the value function:

- Given a prompt, generate several responses
- Compute their individual rewards (using a rule-based reward model)
- An estimate of the value function is the average score

OPTION 4: GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

\mathcal{D} : distribution of prompts

G : number of responses generated for a single prompt

π_{ref} : reference model (frozen)

π_{θ} : model with parameters θ

Loss:

$$\mathcal{L}(\phi) = -\mathbb{E}_{x \sim \mathcal{D}, (y_i)_{i \in [1, G]} \sim \pi_{\text{ref}}(x)} \left[\frac{1}{G} \sum_{i=1}^G F(x, y_i) \right]$$

where

$$F(x, y) = \frac{1}{|y|} \sum_{t=1}^{|y|} \min \left(\frac{\pi_{\theta}(y_t | x, y_{<t})}{\pi_{\text{ref}}(y_t | x, y_{<t})} \cdot A_{\pi_{\theta}}((x, y_{<t}), y_t), g(\epsilon, A_{\pi_{\theta}}((x, y_{<t}), y_t)) \right)$$

OPTION 4: GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

The advantage is not anymore computed using GAE, it is based on this simple computation:

x : prompt

y_1, \dots, y_G : responses

r_1, \dots, r_G : rewards according to the rule-based reward model

Advantage:

$$A((x, y_{i,<t}), y_{i,t}) = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r}) + \epsilon}$$

OPTION 4: GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

Another (minor) difference with PPO:

In PPO, the KL-penalty term for not deviating was added to the reward

In GRPO, the reward is unchanged and the KL-penalty term is added to the loss using a different estimate:

$$D_{\text{KL}}(\pi_{\theta}(y | x) || \pi_{\text{ref}}(y | x)) = \sum_{t=1}^{|y|} \frac{\pi_{\theta}(y_t | x, y_{<t})}{\pi_{\text{ref}}(y_t | x, y_{<t})} - \log \frac{\pi_{\theta}(y_t | x, y_{<t})}{\pi_{\text{ref}}(y_t | x, y_{<t})} - 1$$

OPTION 4: GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

GRPO is “just” a simpler way of estimating the advantage, it is not specific to the “clipped surrogate objective” (TRPO / PPO style), it can also be adapted to the original objective (more like “vanilla policy gradient”). Recall the formula:

π_θ : policy with parameters θ

$\rho = (s_0, a_0, r_0, s_1, a_1, r_1, \dots)$: trajectory

Gradient:

$$\nabla_\theta \mathcal{L} = -\mathbb{E}_{\rho \sim \pi_\theta} \left[\sum_{t=0}^{\infty} A_{\pi_\theta}(s_t, a_t) \nabla_\theta \log \pi_\theta(a_t | s_t) \right]$$

OPTION 3: DIRECT PREFERENCE OPTIMIZATION (DPO)

Interesting exercise: understand how the loss is derived, see <https://arxiv.org/abs/2305.18290>

\mathcal{D} : distribution of triples (prompt, better response, worse response)

π_{ref} : reference model (frozen)

π_{θ} : model with parameters θ

σ : logistic function $\sigma(x) = \frac{1}{1+\exp(-x)}$

Loss:

$$\mathcal{L}(\phi) = -\mathbb{E}_{(x, y^+, y^-) \sim \mathcal{D}} \left[\log \sigma \left(\beta \frac{1}{|y^+|} \sum_{t=1}^{|y^+|} \log \frac{\pi_{\theta}(y_t^+ | x, y_{<t}^+)}{\pi_{\text{ref}}(y_t^+ | x, y_{<t}^+)} - \beta \frac{1}{|y^-|} \sum_{t=1}^{|y^-|} \log \frac{\pi_{\theta}(y_t^- | x, y_{<t}^-)}{\pi_{\text{ref}}(y_t^- | x, y_{<t}^-)} \right) \right]$$