Notes on Why Reinforcement Learning Forgets Less

compiled by D. Gueorguiev

# Notes to LLM training via Reinforcement Learning

The training of LLMs these days is performed in two separate steps

1) Pre-training: Learning general language patterns from vast, unlabeled text.

2) Post-training (fine-tuning): Specializing the model for task or preferences:

//TODO: give an overview of 1) based on [UML]

//TODO: give an overview of 2) based on [TtFI]

//TODO: finish the overview based on [TaRL]

# References

[TaRL] [Large Language Model Training and Reinforcement Learning, Miquel Nguer I Alonso, 2025](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/Large_Language_Model_Training_and_Reinforcement_Learning_Alonso_2025.pdf)

[UML] [Language Models are Unsupervised Multitask Learners, A. Radford et al, 2019](https://github.com/dimitarpg13/large_language_models/blob/main/articles/Language_Models_are_Unsupervised_Multitask_Learners_Alec_Radford_OpenAI_2018.pdf)

[TtFI] [Training Language Models to follow instructions with human feedback , L. Ouyang et al, 2022](https://github.com/dimitarpg13/large_language_models/blob/main/articles/Training_language_models_to_follow_instructions_with_human_feedback_Ouyang_2022.pdf)

[] [RL with KL penalties is better viewed as Bayesian inference, Tomasz Korbak et al, 2022](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/RL_with_KL_penalties_is_better_viewed_as_Bayesian_inference_Korbak_2022.pdf)

[] [RL Razor - Why Online Reinforcement Learning Forgets Less, I. Shenfeld et al, MIT Improbable AI Lab, 2025](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/RL_Razor-Why_Online_Reinforcement_Learning_Forgets_Less_Shenfeld_2025.pdf)

[]

[]

[]

[] [Training Language Models to Self-Correct via Reinforcement Learning, Aviral Kumar et al, Google DeepMind, 2024](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/Training_Language_Models_to_Self-Correct_via_Reinforcement_Learning_Kumar_2024.pdf)

[] [Learning to Think: Information-Theoretic Reinforcement Fine-Tuning for LLMs, J. Wang et al, 2025](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/Learning_to_Think-Information-Theoretic_Reinforcement_Fine-Tuning_for_LLMs_Wang_2025.pdf)

[] [Asymptotics of Large Language Model Alignment, J. Yang et al, U of Sydney, MIT, Google Research, 2024](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/Asymptotics_of_Language_Model_Alignment_Yang_2024.pdf)

[] [Controlled Decoding from Language Models, S. Mugdal et al, 2024](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/Controlled_Decoding_from_Language_Models_Mudgal_2024.pdf)

[] [LLMs are Greedy Agents: Effects of RL fine tuning on Decision-making Abilities, Thomas Schmied et al, 2025](https://github.com/dimitarpg13/large_language_models/blob/main/articles/reinforcement_learning/LLMs_are_Greedy_Agents-Effects_of_RL_Fine-tuning_on_Decision-Making_Abilities_Schiemd_2025.pdf)

[SFT] [Supervised Fine-Tuning, Chapter 11 of LLM Course on HuggingFace.co](https://huggingface.co/learn/llm-course/en/chapter11/1)

[TRL] [HuggingFace’s Transformer Reinforcement Learning library documentation](https://huggingface.co/docs/trl/en/index)

[STr] [SFT Trainer at HuggingFace Transformer Reinforcement Learning Library documentation](https://huggingface.co/docs/trl/en/sft_trainer)

# Appendix

## Supervised Fine Tuning of Large Language Models

Generative models can be finetuned on specific tasks such as summarization and question answering. Whenever we fine-tune LMs on a broad range of tasks *simultaneously* we talk about supervised fine-tuning (SFT).

This process helps LMs become more versatile and capable of handling a broader set of use cases which represent the user interests.

There are 4 distinct phases (steps) in the SFT process:

### Chat Templates

*Chat Templates* structure interactions between users and LMs ensuring consistent contextually relevant responses. They include the following components – *system prompts* and *role-based messages*.

### Supervised Fine-Tuning

SFT is a process for adapting pretrained LM to specific tasks. It involves training the model on a task-specific dataset with labeled examples. Detailed documentation with example can be found on [TRL] and [STr].

//TODO: finish the overview based on [SFT]