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].

Simple Example:

from trl import SFTConfig, SFTTrainer

from datasets import load\_dataset

trainer = SFTTrainer(

model="Qwen/Qwen3-0.6B",

train\_dataset=load\_dataset("trl-lib/Capybara", split="train"),

)

trainer.train()

The Capybara dataset is a compact diverse multi-turn dataset to benchmark reasoning and generalization.

The Capibara Dataset

Example multi-turn question:

[

{

"content": "Recommend a movie to watch.\n",

"role": "user"

},

{

"content": "I would recommend the movie, \"The Shawshank Redemption\" which is a classic drama film starring Tim Robbins and Morgan Freeman. This film tells a powerful story about hope and resilience, as it follows the story of a young man who is wrongfully convicted of murder and sent to prison. Amidst the harsh realities of prison life, the protagonist forms a bond with a fellow inmate, and together they navigate the challenges of incarceration, while holding on to the hope of eventual freedom. This timeless movie is a must-watch for its moving performances, uplifting message, and unforgettable storytelling.",

"role": "assistant"

},

{

"content": "Describe the character development of Tim Robbins' character in \"The Shawshank Redemption\".",

"role": "user"

},

{

"content": "In \"The Shawshank Redemption\", Tim Robbins plays the character of Andy Dufresne, a banker who is wrongfully convicted of murdering his wife and her lover. Over the course of the film, we see a significant transformation in Andy's character.\n\nAt the beginning of the movie, Andy is a quiet, reserved man who maintains his innocence but is largely passive in the face of his unjust conviction. He seems almost aloof, detached from the harsh reality of his situation. However, beneath this seemingly impassive exterior, Andy is a man of deep intelligence and inner strength.\n\nAs the story progresses, Andy begins to adapt to his new environment. He uses his banking skills to gain favor with the prison guards and the warden, providing them with financial advice and even helping them with tax evasion. This not only earns him certain privileges but also gives him a measure of protection within the prison walls.\n\nDespite the grim circumstances, Andy never loses hope. He befriends a fellow inmate, Red, and shares with him his dream of living a quiet life in Zihuatanejo, a small town on the Pacific coast of Mexico. This dream keeps him going and becomes a symbol of hope for other inmates as well.\n\nPerhaps the most significant development in Andy's character comes towards the end of the film, when it is revealed that he has been digging a tunnel through his cell wall for the past 19 years. This act of defiance, carried out with nothing but a small rock hammer, is a testament to Andy's resilience and determination. It also shows his cunning, as he manages to keep his plan a secret from everyone, including Red.\n\nIn the end, Andy's character arc is one of hope, resilience, and redemption. Despite being placed in a situation that seems hopeless, he never gives up. He maintains his dignity, uses his intelligence to his advantage, and ultimately, manages to escape from Shawshank, proving his innocence in the process. His character serves as a powerful symbol of the human spirit's ability to endure and overcome even the most challenging circumstances.",

"role": "assistant"

},

{

"content": "Explain the significance of the friendship between Andy and Red in shaping Andy's character development.",

"role": "user"

},

{

"content": "The friendship between Andy Dufresne (Tim Robbins) and Ellis \"Red\" Redding (Morgan Freeman) is central to the narrative of \"The Shawshank Redemption\", and plays a significant role in shaping Andy's character development.\n\nWhen Andy first arrives at Shawshank, he is a quiet, introverted man, largely keeping to himself. Red, who is the long-term inmate, takes an interest in him and they slowly develop a friendship. Red's wisdom, street-smart attitude, and his ability to procure items become instrumental in Andy's survival and eventual escape.\n\nRed is initially skeptical of Andy's innocence and his hopeful outlook on life. However, as their friendship grows, Red becomes more receptive to Andy's perspective. This friendship provides Andy with a confidant, a sounding board, and a supportive ally amidst the harsh realities of prison life. \n\nAndy's influence on Red is equally profound. Andy's unyielding hope and resilience slowly chip away at Red's hardened cynicism. Andy shares his dreams of freedom and his plans for the future with Red, which initially seem unrealistic to Red, but over time, Andy's unwavering belief in hope begins to influence Red's outlook on life.\n\nIn many ways, their friendship is a beacon of hope and humanity in an otherwise oppressive environment. It's through this friendship that Andy finds the strength to maintain his dignity, persevere, and ultimately, to engineer his daring escape. It's also through this friendship that Red finds hope for redemption and a life beyond the prison walls.\n\nIn conclusion, the friendship between Andy and Red is a pivotal element in shaping Andy's character development. It's through this bond that Andy finds the strength to endure his unjust imprisonment and to hold onto hope, ultimately leading to his redemption.",

"role": "assistant"

}

]

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