Awesome-ML-SYS-Tutorial

English version | 简体中文

My learning notes/codes for ML SYS. English version is under development and only available for some texts.

RLHF System Development Notes

- **Intro to HybridFlow/veRL**
  [English TODO] | [中文版]：SGLang's hybrid RLHF engine design and implementation.
- **Extending OpenRLHF's Inference Engine**
  [English TODO] | [中文版]：Notes on integrating SGLang with OpenRLHF, an exhausting process with frequent NCCL hang bugs.
- **SWE-Bench: How to Construct a Great Benchmark for the LLM Era**
  [中文版]
- **Intro to Workflow in OpenRLHF-like Post-Training Systems**
  [中文版]
- **The Illustrated PPO: Theory and Source Code Explanation**
  [中文版]
  Also see RLHF 的计算流.
- **Latency Optimization for Weight Updates**
  [English TODO] | [中文版]：An experience of debugging loading efficiency.
- **Intro to Alignment Algorithms and NeMo-Aligner Framework**
  [中文版]

SGLang Learning Notes

- **Concepts and Optimization of Constraint Decoding**
  [English TODO] | [中文版]
- **SGLang Code Walkthrough**
  [English version]：The lifecycle of a request in the SGLang Engine, a good start for SGLang beginners.
- **Walk Through SGLang / VLLM Worker**
  [English version]：Demystifying the SGLang worker (model executor).
- **Reward / Embed Model Server Engine**
  [English TODO] | [中文版]

- **SGLang Backend Analysis**
  [English TODO] | [中文版]
- **Using vLLM to Serve New Embedding Models**
  [English TODO] | [中文版]
- **Using SGL to Serve Embedding Models**
  [English TODO] | [中文版]
- **From vLLM to SGLang: A User's Perspective**
  [English TODO] | [中文版]

Scheduling and Routing

- **Mooncake: Maximizing PD Disaggregation**
  [中文版]：Taking prefill and decode separation to the extreme.
- **Should Prefill and Decode Be Separated onto Different Cards?**
  [中文版]：A discussion on separating prefill and decode tasks.
- **Understanding Prefill and Decode Computational Characteristics Based on Chunked Prefill**
  [中文版]：Analyzing computational characteristics using chunked prefill.
- **ModelServer: A Frontend Distribution System Based on SGLang**
  [中文版]：A frontend distribution system built on SGLang.

ML System Fundamentals

- **NCCL and NVIDIA TOPO**
  English | [中文版]：An introduction to NCCL and NVIDIA topology.
- **PyTorch Distributed**
  [English TODO] | [中文版]：Practical communication in `torch.distributed`.
- **Give Me BF16 or Give Me Death: A Comprehensive Evaluation of Current Quantization Methods**
  [中文版]：A detailed evaluation of current quantization methods.
- **AWQ: Model Quantization Should Focus on Activation Values**
  [中文版]：Why activation values should be the focus of model quantization.
- **Deep Dive into PyTorch DDP Series Part 1: Beginner's Tutorial**
  [中文版]：A beginner's guide to PyTorch Distributed Data Parallel (DDP).
- **Detailed Explanation of nvidia-smi Command and Some Advanced Techniques**
  [中文版]：Advanced techniques for using `nvidia-smi`.

Other

- **Setting Up a Clean Development Environment**
  [English TODO] | [中文版]：How to set up a clean and efficient development environment.
- **Understanding Special Tokens and Chat Templates**
  [English TODO] | [中文版]：A guide to understanding special tokens and chat templates.
- **Compiling Jupyter Notebooks on CI and Deploying as Documentation**
  [中文版]：A guide on compiling Jupyter notebooks in CI and deploying them as documentation.