Synthetic Data Generation and Supervised Fine Tuning

From synthetic data generation to model fine-tuning using NeMo

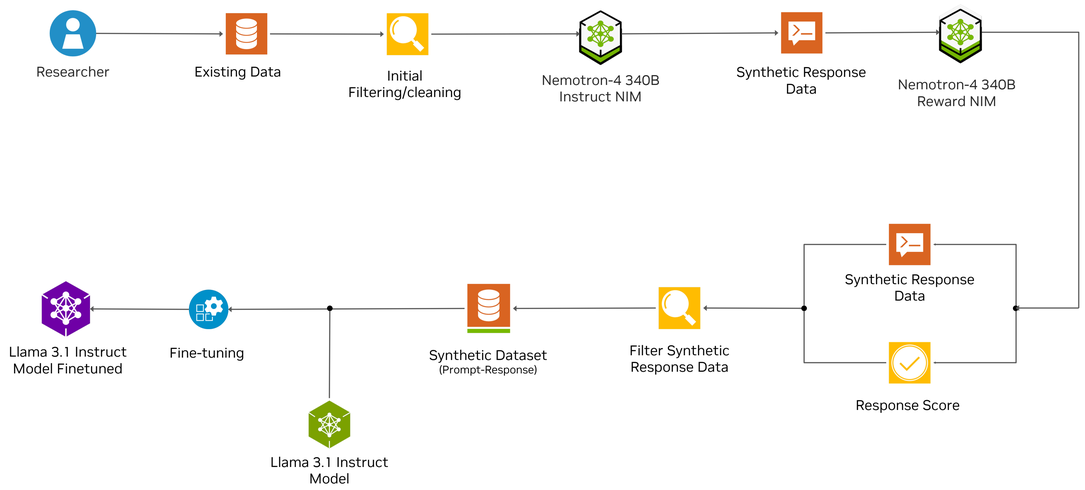
https://github.com/NVIDIA-AI-Technology-Center/synthetic-data-generation-and-sft-playbook

- **Challenge:**
  - Training a large language model (LLM), whether for pre-training, fine-tuning, or alignment, demands significant effort and computational resources especially when dealing with massive data. Many codes exist but only few are capable to scale efficiently across multiple GPUs.
  - This playbook aims to show a scalable pipeline based on the NeMo Framework to improve LLM performance using synthetically enriched data.

- **Prerequisites**:
  - **Generate synthetic data**: access to NVIDIA Services (https://build.nvidia.com/) or to at least 32 NVIDIA H100 GPUs.
  - **Model fine-tuning**: at least 2 GPUs (A100 80 GB)

From synthetic data generation to model fine-tuning using NeMo

- **The playbook goes through these steps:**
  1. Data preparation and pre-processing
  2. Synthetic Data Generation using Nemotron-4-340B-Instruct and Reward
  3. Data Filtering and Quality Control
  4. Enriched dataset creation combining real and high-quality synthetic examples
  5. Fine-tune Llama 3.1 on the enriched dataset using LORA (Low-Rank Adaptation)
  6. Evaluate generated results

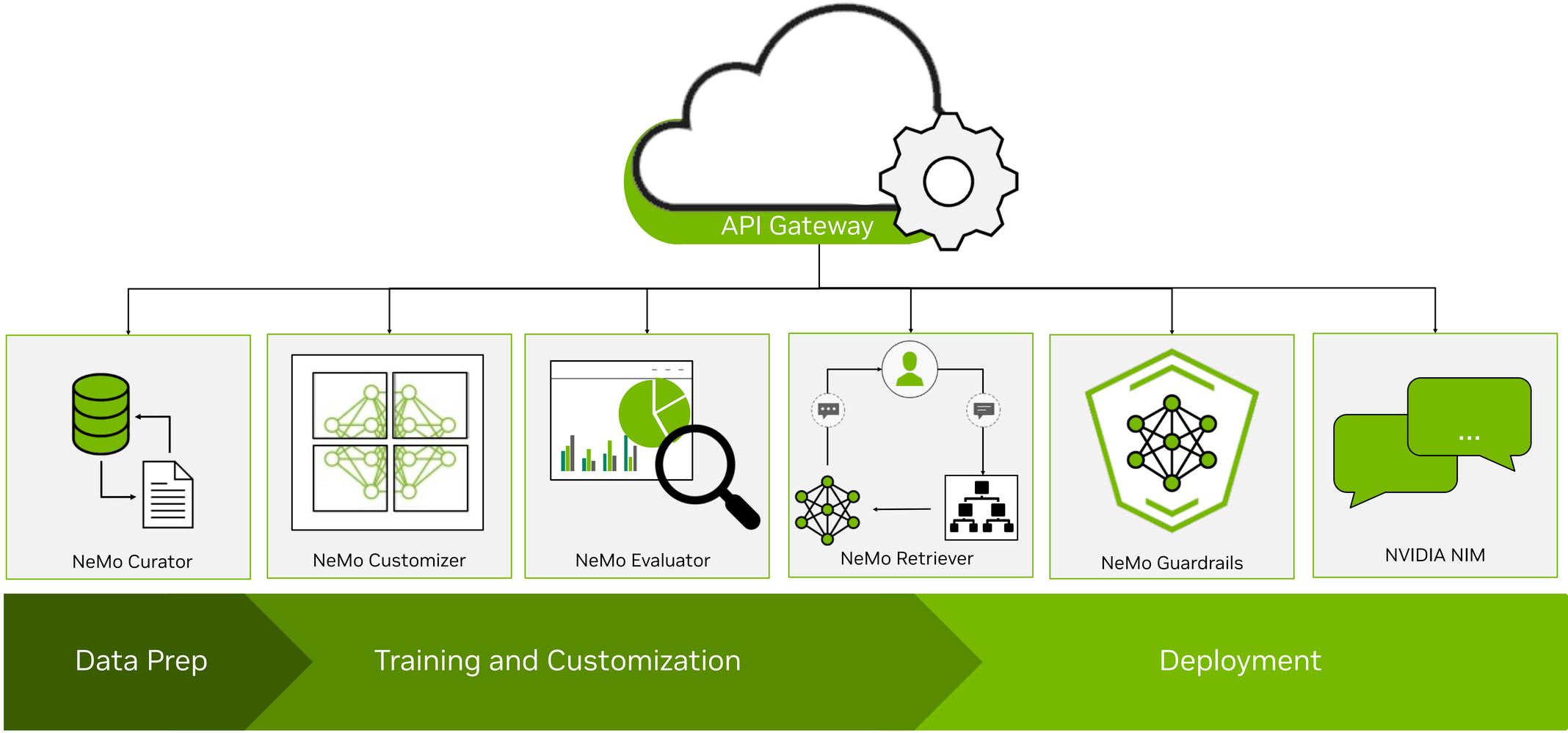


More in details

- **Data Preparation**
  - First, the original law-stackexchange-questions-answers dataset is obtained from Hugging Face and pre-processed. This dataset contains questions and answers related to legal topics. The data is then split into training, validation, and test sets.
- **Synthetic Data Generation**
  - Using Nemotron-4-340B-instruct via NIM, synthetic data is generated to augment the original dataset. The model is prompted to create additional question-answer pairs that mimic the style and content of the law-stackexchange data.
- **Data Filtering and Quality Control**
  - The generated synthetic data is then passed through the Nemotron-4-340B-reward model. This model acts as a quality filter, evaluating the relevance and coherence of the generated question-answer pairs. Low-quality or irrelevant pairs are discarded, ensuring that only high-quality synthetic data is retained.
- **Dataset Combination**
  - The filtered synthetic data is combined with the original law-stackexchange dataset. This enriched dataset now contains both real and high-quality synthetic examples, providing a more comprehensive training set for fine-tuning.
- **Fine-Tuning Process**
  - The Llama-3.1-model-instruct (8B) is then fine-tuned on this enriched dataset using Lora. The fine-tuning process involves training the model on the combined data, allowing it to learn the specific patterns and knowledge required for legal question-answering tasks.

NeMo Framework

https://github.com/NVIDIA/NeMo



Nemotron-4 340B Family of Models & Tools

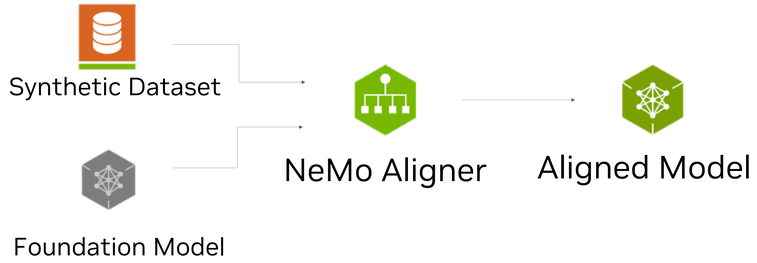
Nemotron-4 340B Instruct

- Generate synthetic data at scale
- 9T tokens, 50 spoken + 40 programming languages
- Leading performance across benchmarks: Math, Instruction following, Human preferences

Nemotron-4 340B Reward

- Scores on: Helpfulness, Correctness, Coherence. Complexity, Verbosity
- Tops Reward Bench leaderboard covering chat, safety, and reasoning

NeMo Aligner



- Align output with human preferences
- Supports: RLHF, DPO, SteerLM, PINS
- Download from Github and available with NeMo
- Support through NVIDIA AI Enterprise

References

1. Liu, Y., Cao, J., Liu, C., Ding, K., and Jin, L., “**Datasets for Large Language Models: A Comprehensive Survey**”, 2024. doi:10.48550/arXiv.2402.18041.
2. Rejeleene, R., Xu, X., and Talburt, J., “**Towards Trustable Language Models: Investigating Information Quality of Large Language Models**”, 2024. doi:10.48550/arXiv.2401.13086.
3. Hoffmann, J., “**Training Compute-Optimal Large Language Models**”, 2022. doi:10.48550/arXiv.2203.15556.