# List of Proposed Datasets for finetuning

Subjects considered:

- Science
- Maths
- Engineering
- Coding

## Maths

1. argilla/distilabel-math-preference-dpo · Datasets at Hugging Face : **distilabel-math-preference-dpo**
2. https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences : Just the Mathematics part
3. https://huggingface.co/datasets/kira/math-dpo
4. https://github.com/openai/prm800k?tab=readme-ov-file

NOTE: For 4 the structure of the dataset is a bit different. The dataset is not formatted in the usual format. Also not suitable for DPO. But it uses a new style of finetuning

## Science

1. https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences : Choose the Science Datasets
2. https://huggingface.co/datasets/ArtifactAI/arxiv-physics-instruct-tune-30k : Dataset for Arxiv research papers
3. https://huggingface.co/datasets/derek-thomas/ScienceQA : Cool dataset to train on MCQ questions

Engineering:

1. https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences : choose the engineering one

Coding:

1. bigcode/starcoderdata
2. Vikp/textbook_quality_programming

Misc:

1.cais/mmlu