MAKE SURE YOU HAVE A HUGGING FACE ACCOUNT FOR MOST OF THESE LINKS

What’s left?

1. Humor datasets (refer to the word file on WhatsApp)
2. Websites for scraping

Datasets Gathered:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Link | Category | Description | Clean? | Notes |
| <https://www.kaggle.com/datasets/conjuring92/wiki-stem-corpus> | STEM  FINETUNING | Text extracted from Wikipedia pages covering many STEM topics and subtopics | Yes | Scope is massive – this includes even geology and other sciences in the data |
| <https://huggingface.co/datasets/GAIR/MathPile/tree/main>  dataset builder script: <https://github.com/GAIR-NLP/MathPile/blob/main/src/global_data_processing/build_dataset.py> | Math  PRETRAINING  FINETUNING | The math pile –contains math data for training generative AI from: proof sites, commoncrawl, Wikipedia, books, arxiv papers, stackexchange | Yes  (all math datasets are clean but text is not exactly readable due to containing latex and other scripting formats specific to digitally formatting math notation) | Involves some pre-processing like building the json dataset – subsets of the dataset are a little tricky like arxiv paper files which contain some noise |
| <https://huggingface.co/datasets/GAIR/FRoG/tree/main> | Math  FINETUNING | for mathematical reasoning and problem solving – contains samples of varying difficulty to challenge the model | Yes | RLHF-based |
| <https://huggingface.co/datasets/GAIR/LIMR/tree/main> | Math  FINETUNING | for more problem solving and fine-tuning | Yes | RLHF-based |
| <https://huggingface.co/datasets/GAIR/LIMO/tree/main> | Math  FINETUNING | High-quality deep reasoning and logic tuning | Yes | RLHF-based |
| <https://huggingface.co/datasets/bigcode/the-stack-v2-dedup>  scripts for building and preprocessing:  <https://github.com/bigcode-project/bigcode-dataset/tree/main/preprocessing> | CS (Code)  PRETRAINING  TRAINING  FINETUNING | The Stack -Extremely massive dataset containing more than a billion code examples with other metadata stored | Yes | This dataset is very massive –  It is also general-purpose and might not help a model excel in specific things like reasoning or debugging etc. it just contains a ton of code examples |
| <https://huggingface.co/datasets/code-search-net/code_search_net/tree/main>  preprocessing and evaluation:  <https://github.com/github/CodeSearchNet> | CS (Code)  FINETUNING | Very useful for code retrieval and recall tasks (semantic search in general) | Yes |  |
| <https://huggingface.co/datasets/ise-uiuc/Magicoder-Evol-Instruct-110K/tree/main>  preprocessing:  <https://github.com/ise-uiuc/magicoder> | CS (Code)  RESPONSE- OPTIMIZATION  FINETUNING | For code generation, completion, modification etc. | Yes | RLHF-based |
| <https://huggingface.co/datasets/Vezora/Open-Critic-GPT> | CS (Code)  RESPONSE-OPTIMIZATION  FINETUNING | For debugging and understanding issues with a code snippet | Yes | RLHF-based |
| <https://github.com/commonsense/conceptnet5/wiki/Downloads> | FUNDAMENTAL  SEMANTICS | Extensive Knowledge graph that maps relationships between words | Yes | Requires some preprocessing first (and maybe filtering in the case that the model was already trained on general semantics) |
| <https://github.com/leanprover-community/mathlib4/tree/master/Mathlib>  utility scripts:  <https://github.com/leanprover-community/mathlib4>  easier alternative (curated and preprocessed but weaker version):  <https://huggingface.co/datasets/JohnYang88/lean-dojo-mathlib4> | FUNDAMENTAL  Math | Contains rigorous proofs, definitions and mathematical terminology | Yes | hard to work with and required preprocessing and possibly even integration but the repo has helpful scripts and there’s an easier alternative version of the dataset |