Dataset Links for GENERATIVE AI

1. Text Generation & Language Modelling

- **Common Crawl**: A vast repository of web crawl data.
  - Access: https://commoncrawl.org/
- **C4 (Colossal Clean Crawled Corpus)**: A cleaned version of Common Crawl used for training models like T5.
  - Access: https://www.tensorflow.org/datasets/catalog/c4
- **The Pile**: An 825GB English text corpus from diverse sources.
  - Access: https://pile.eleuther.ai/
- **OpenWebText**: An open-source recreation of OpenAI's WebText dataset.
  - Access: https://skylion007.github.io/OpenWebTextCorpus/
- **Wikipedia**: A comprehensive encyclopaedia used for pretraining language models.
  - Access: https://dumps.wikimedia.org/
- **BookCorpus**: A dataset of books used for training models like GPT.
  - Access: https://huggingface.co/datasets/bookcorpus
- **PG-19**: A long-form text dataset from Project Gutenberg.
  - Access: https://github.com/deepmind/pg19
- **Wikitext-103**: A high-quality dataset based on Wikipedia.
  - Access: https://huggingface.co/datasets/wikitext
- **RedPajama**: A high-quality dataset replicating LLaMA's pretraining dataset.
  - Access: https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T
- **Pile-of-Law**: A legal dataset containing U.S. case law and statutes.
  - Access: https://huggingface.co/datasets/pile-of-law/pile-of-law

2. Conversational AI & Chatbots

- **OpenAssistant Conversations (OASST1)**: A dataset for assistant-style training.
  - Access: https://huggingface.co/datasets/OpenAssistant/oasst1
- **Anthropic HH Dataset**: A dataset for helpful and honest conversational AI.
  - Access: https://github.com/anthropics/hh-rlhf

- **SQuAD (Stanford Question Answering Dataset)**: A dataset for question-answering tasks.
    - Access: https://rajpurkar.github.io/SQuAD-explorer/
- **Natural Questions (NQ)**: A dataset containing real search queries from Google.
    - Access: https://ai.google.com/research/NaturalQuestions
- **MultiWOZ**: A large-scale multi-turn dialogue dataset for task-oriented chatbots.
    - Access: https://github.com/budzianowski/multiwoz

3. Code Generation

- **The Stack**: A large collection of open-source code from GitHub.
    - Access: https://huggingface.co/datasets/bigcode/the-stack
- **CodeParrot**: A dataset curated for training AI coding assistants.
    - Access: https://huggingface.co/datasets/codeparrot/github-code
- **CodeSearchNet**: A dataset of source code and comments for code search and completion.
    - Access: https://github.com/github/CodeSearchNet
- **Python Pile**: A subset of The Pile focused on Python code.
    - Access: https://pile.eleuther.ai/

4. Image Generation (Diffusion Models, GANs)

- **LAION-5B**: A massive dataset of image-text pairs used for training models like Stable Diffusion.
    - Access: https://laion.ai/blog/laion-5b/
- **COCO (Common Objects in Context)**: A dataset for image captioning and object detection.
    - Access: https://cocodataset.org/
- **ImageNet**: A widely used dataset for pretraining vision models.
    - Access: https://www.image-net.org/
- **Conceptual Captions**: A dataset of image-text pairs from web sources.
    - Access: https://ai.google.com/research/ConceptualCaptions
- **WebVision**: A dataset for large-scale image-text pretraining.
    - Access: http://www.vision.ee.ethz.ch/webvision/

5. Multimodal Learning (Text-to-Image, Text-to-Video)

- **LAION-400M / LAION-5B**: Large-scale image-text datasets.
    - Access: https://laion.ai/