dataset (3–5)

databricks-dolly-15k

Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM

`databricks-dolly-15k` contains **15,000 high-quality human-generated prompt** / response pairs specifically designed for instruction tuning large language models. Under the licensing terms for `databricks-dolly-15k` (Creative Commons Attribution-ShareAlike 3.0 Unported License), anyone can use, modify, or extend this dataset for any purpose, including commercial applications.
To the best of our knowledge, **this dataset is the first open source, human-generated instruction dataset specifically designed to make large language models exhibit the magical interactivity of ChatGPT.** `databricks-dolly-15k` was authored by more than 5,000 Databricks employees during March and April of 2023. These training records are natural, expressive and designed to represent a wide range of the behaviors, from brainstorming and content generation to information extraction and summarization.

OASST1 (OpenAssistant Conversations Dataset)

OpenAssistant Conversations - Democratizing Large Language Model Alignment

RedPajama-Data-1T

RedPajama, a project to create leading open-source models, starts by reproducing LLaMA training dataset of over 1.2 trillion tokens

RedPajama–Data–1T consists of seven data slices:

- CommonCrawl: Five dumps of CommonCrawl, processed using the CCNet pipeline, and filtered via several quality filters including a linear classifier that selects for Wikipedia–like pages.
- C4: Standard C4 dataset
- GitHub: GitHub data, filtered by licenses and quality
- arXiv: Scientific articles removing boilerplate
- Books: A corpus of open books, deduplicated by content similarity
- Wikipedia: A subset of Wikipedia pages, removing boilerplate

- StackExchange: A subset of popular websites under StackExchange, removing boilerplate

For each data slice, we conduct careful data pre-processing and filtering, and tune our quality filters to roughly match the number of tokens

	RedPajama	LLaMA*
CommonCrawl	878 billion	852 billion
C4	175 billion	190 billion
Github	59 billion	100 billion
Books	26 billion	25 billion
ArXiv	28 billion	33 billion
Wikipedia	24 billion	25 billion
StackExchange	20 billion	27 billion
Total	1.2 trillion	1.25 trillion

StarCoder: A State-of-the-Art LLM for Code

StarCoder 和 StarCoderBase 是代码的大型语言模型（Code LLM），使用来自 GitHub 的许可数据进行训练，包括来自 80+ 编程语言、Git 提交、GitHub 问题和 Jupyter 笔记本。与 LLaMA 类似，研究者为 15 万亿个 Token 训练了一个～1B 参数模型。他们对 35B Python Token 的 StarCoderBase 模型进行了 fine-tune，产生了一个名为 StarCoder 的新模型。

研究者发现，StarCoderBase 在流行的编程基准测试中优于现有的开放代码 LLM，并且匹配或超过了封闭模型，例如来自 OpenAI（为 GitHub Copilot 早期版本提供支持的原始 Codex 模型）。StarCoder 模型的上下文长度超过 8,000 个token，可以处理比任何其他开放 LLM 更多的输入，从而实现各种有趣的应用程序。例如，通过一系列对话提示 StarCoder 模型，使他们能够充当技术助理。此外，这些模型可用于自动完成代码，通过指令修改代码，以及用自然语言解释代码片段。 研究者采取了几个重要步骤来实现安全的开放模型发布，包括改进的 PII 编辑管道、新颖的归因跟踪工具，以及公开提供 StarCoder 在 OpenRAIL 许可证的改进版本下。更新后的许可证简化了公司将

模型集成到其产品中的流程。研究者相信，凭借其强大的性能，StarCoder 模型将成为社区使用和适应其用例和产品的坚实基础。