Dolma—Exploring LLM Pretraining Data

Coming Soon

Data Essay

Introduction

Coming soon. This dataset will include information about Dolma, "an open dataset of 3 trillion tokens from a diverse mix of web content, academic publications, code, books, and encyclopedic materials," which was used to train the OLMo model from the Allen Institute for AI (AI2).