Analytics Vidhya

Data Gathering and Preprocessing

Instructor

Sourab Mangrulkar

Machine Learning Engineer at

Creator of PEFT

Dataset Collection

Annotated data by Skilled Humans

High-quality and diverse dataset created by skilled human annotators.

LIMA, OpenAssistant Conversations Dataset, No Robots

Dataset Collection

LLM assisted Data Generation

Self-Instruct: Aligning Language Models with Self-Generated Instructions

Start with seed instructions, use LLM to generate more instructions, filter them out for quality and diversity, rinse and repeat.

Use prompt template or LLM to generate instructions and input/output pairs for corporate documents, meeting notes, blogposts, wikipedia articles ...

Alpaca, Ultrachat, CodeAlpaca

Dataset Collection

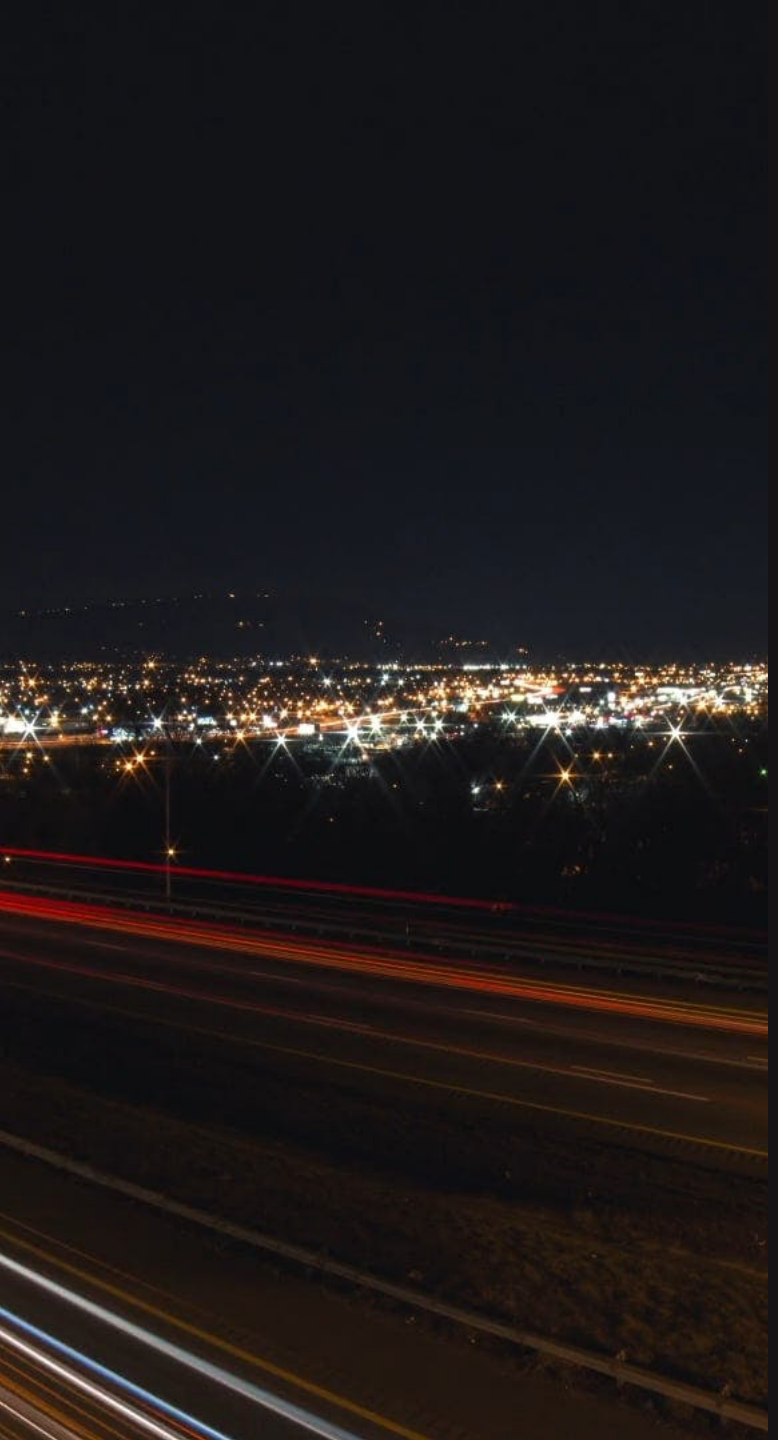
External Datasets

FAQs

Transcribed customer support conversations, meetings, podcasts

Conversations on social media, discord, slack

textbook question and answers, GitHub issues/PR conversations



Remember to consider

Quality

High-quality dataset is the key

Avoid Garbage in → Garbage out

Quantity

For instruction finetuning, data in order of thousands of samples is a good starting point

The more the better

Diversity

Diverse tasks enable generalization and serendipity

Source

Human annotated data is the gold standard

LLM generated data often has subtle patterns that can hinder model training

Data Collection Pipeline

Collect Instruction dataset (Instruction, Input, Output) tuples → Format and concatenate → Train/Test Split