**Assignment 03 — Building a Chain-of-Thought (CoT) Dataset**

**Objective:**

The goal of this assignment is to gain hands-on experience in constructing a high-quality dataset specifically designed for fine-tuning language models on step-by-step reasoning tasks. This includes understanding data curation, preprocessing, and formatting best practices.

**Assignment Focus:**

1. Learn the process of collecting QA data suitable for CoT modeling.
2. Clean and filter examples to ensure clarity and diversity.
3. Manually or programmatically craft CoT-style step-by-step rationales.
4. Produce a clean, structured dataset in JSON or CSV format that can be directly used for model fine-tuning.

**Dataset Specifications:**

Dataset Size: 500–1000 question-answer pairs

**Content Types:**

1. Arithmetic word problems
2. Logical reasoning questions
3. General knowledge or trivia with underlying reasoning

**Each data sample should contain:**

1. question: the input query or problem
2. cot\_answer: reasoning steps in natural language
3. final\_answer: the concise solution derived from reasoning

**Example Format (JSON):**

{

"question": "Tom has 3 red balls and 4 blue balls. He gives away 2 red balls. How many balls does he have left?",

"cot\_answer": "Tom starts with 3 red balls and 4 blue balls. After giving away 2 red balls, he has 1 red ball left. In total, he now has 1 red + 4 blue = 5 balls.",

"final\_answer": "5"

}

**Recommended Sources for QA Data:**

1. Open datasets (GSM8K, MAWPS, CommonsenseQA)
2. Educational websites and puzzle forums
3. Manually crafted or curated by the author

**Data Cleaning & Preprocessing:**

1. Ensure questions are well-formed and answerable.
2. Use clear language and consistent formatting across samples.
3. Remove duplicates, irrelevant or ambiguous entries.
4. Verify correctness of reasoning chains and final answers.

**Tools Used:**

1. Python scripting (for scraping, formatting, and validation)
2. Pandas (CSV handling)
3. JSON module (for export)
4. Manual verification of CoT reasoning chains

**Deliverables:**

1. A complete dataset file: cot\_dataset.json or cot\_dataset.csv
2. A README within the dataset folder detailing schema and construction methodology
3. (Optional) Summary statistics: average reasoning length, topic distribution

**Software Requirements:**

Python ≥ 3.10

**Required Python packages:**

1. pandas
2. json (standard library)
3. tqdm

**Installation (if needed):**

pip install pandas tqdm

**Best Practices Followed:**

1. Each sample verified for logical consistency
2. Diversity in question types and domains
3. Consistent JSON key structure across entries
4. Dataset size chosen to balance effort and utility for training small models

**Author:**

Bhagyesh Makhamale

Under guidance of: Dr. Yogita Mahatekar

Affiliation: COEP Technological University (Department of AI & DS)

Academic Year: 2025–2026