Creating Chemical Process Dataset with ChatGPT

By -

Shubhang Gautam - 2022A1PS1666P

Yugansh Jain - 2022A3PS1212P

Harsh Shah - 2022A7PS0169P

BITS Pilani

Pilani Campus

Introduction

This project focuses on developing a structured dataset containing chemical process flowcharts along with their corresponding question–answer pairs.
The goal is to capture complex industrial processes in a standardized, machine-readable format while also providing natural-language explanations.

Project Overview

- To create a high-quality repository of chemical engineering process knowledge.
- To support AI model training, process understanding, and automated flow chart generation.
- To document industrial processes consistently using hierarchical descriptions, Mermaid diagrams, and domain-specific metadata.

Project Overview

- Covers major chemical and petrochemical processes (SMR, Haber Bosch, Ostwald, Hydrotreating, Distillation, etc.)
- Includes feed compositions, process conditions, reactor and separator details, and utilities
- Flowcharts created using standardized Mermaid syntax
- Every process accompanied by question-answer pairs

Key Features

- **Consistency**: All flowcharts follow uniform conventions
- **Completeness**: Each process contains full hierarchical breakdown
- **Clarity**: Designed for readability by students, engineers, and AI models
- **Extensibility**: New processes can be added easily

Use Cases

- **AI Model Training -** Fine-tuning Large Language Models (LLMs) to understand and reason about complex engineering diagrams.
- **Education & EdTech -** Automated generation of quizzes, tutoring systems, and assessment tools grounded in real-world process scenarios.
- **Process Optimization -** Developing intelligent assistants for troubleshooting, hazard identification (HAZOP), and efficiency analysis.
- **Knowledge Base -** Enhancing documentation search and retrieval systems within engineering firms using semantic understanding.
- **Benchmarking & Standardization -** Establishing standard metrics and leaderboards for evaluating the performance of vision-language models in the chemical engineering domain.

Workflow

Prompt Engineering

Use LLMs to create a mermaid code of chemical processes.

Flowchart Creation

Create visual flowcharts from mermaid code

Question - Answer pairs

Feed the textual data and the visual chart to LLMs to create question answer pairs

Validation

Questions

- Theoretical -
  - How do the different feed stream compositions influence the reaction pathway and overall efficiency of this process?
  - Which parts of the flowchart represent major heat sources and sinks?
  - What are the main operational bottlenecks or rate-limiting steps in the process?
- Topological -
  - List the full process path from the primary feed stream to the final product, identifying each step in order without skipping any intermediate units.
  - Label each node (unit) as an entry point, exit point, intermediate node, branching node, or merging node based on how many inputs and outputs it has.
  - How many input and output streams does the flowchart have?

Example

Query : SMR

Below is the **clean, hierarchical process flowchart for the SMR (Steam Methane Reforming) process**, formatted exactly in the same style as your example.

Hierarchical Process Flowchart – Steam Methane Reforming (SMR)

(Feed → Reforming → Shift → CO_2 Removal → Purification → Compression → Product H_2)

1. Feed Preparation & Desulfurization

- Natural Gas Feed
- Feed Filters / Knockout
- Preheater
- Desulfurization Reactor (ZnO)

2. Reforming

- Primary Reformer (Ni Catalyst Tubes)
- Secondary Reformer (Optional, Fired with Air/O_2)

3. Shift Conversion

- High Temperature Shift (HTS)
- Low Temperature Shift (LTS)

4. CO_2 Removal

- CO_2 Absorption (MEA/Benfield/Physical Solvent)
- CO_2 Stripping / Regeneration
- CO_2 to Storage or Urea Plant (Optional)

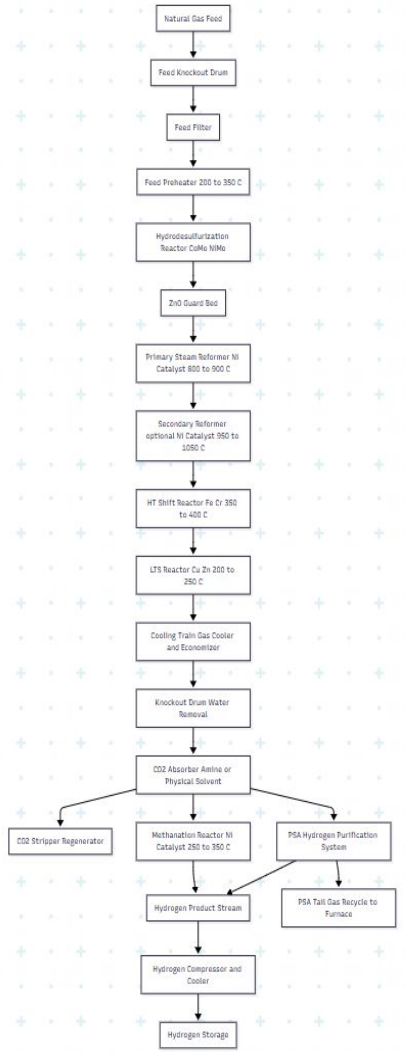
5. Purification

- Methanation (Trace CO, CO_2 → CH_4)

  or
- PSA Unit (for high-purity H_2)

6. Compression & Product Handling

- H_2 Compressor
- H_2 Cooling
- High-Purity H_2 Product



Example

✅ 3. Main operational bottlenecks or rate-limiting steps

1. **Primary Steam Reformer**
   - Tube metallurgy limits maximum heat flux
   - Most expensive and slowest-responding section
   - Controls maximum plant throughput
2. **Secondary Reformer (if used)**
   - Air addition and mixing restrict maximum firing and conversion
3. **Shift Reactors**
   - HTS catalyst deactivation due to sintering
   - LTS severely limited by minimum inlet temperature to avoid Cu catalyst damage
4. **CO2 Absorber and Stripper**
   - Solvent circulation rate limits CO2 removal
   - Stripper energy demand limits throughput
5. **PSA Unit**
   - Cycle time and number of beds limit maximum hydrogen production
   - Tail gas composition affects furnace balance

✅ 4. Full process path from primary feed to final product — step-wise

From the flowchart:

1. Natural Gas Feed
2. Feed Knockout Drum
3. Feed Filter
4. Feed Preheater 200 to 350 C
5. Hydrodesulfurization Reactor CoMo NiMo
6. ZnO Guard Bed
7. Primary Steam Reformer Ni Catalyst 800 to 900 C
8. Secondary Reformer optional Ni Catalyst 950 to 1050 C
9. HT Shift Reactor Fe Cr 350 to 400 C
10. LTS Reactor Cu Zn 200 to 250 C
11. Cooling Train Gas Cooler and Economizer
12. Knockout Drum Water Removal
13. CO2 Absorber Amine or Physical Solvent
14. Branch:
    - Path A → CO2 Stripper Regenerator → returns lean solvent
    - Path B → Methanation Reactor Ni Catalyst 250 to 350 C
    - Path C → PSA Hydrogen Purification System
15. Hydrogen Product Stream
16. Hydrogen Compressor and Cooler
17. Hydrogen Storage

Challenges

1. Process Complexity

- Industrial chemical processes are highly detailed
- Multiple units, recycle loops, catalysts, and utilities
- Hard to capture accurately without oversimplifying or overcomplicating

2. Standardization of Diagrams

- Ensuring consistent Mermaid syntax across all processes
- Avoiding variations in naming, labeling, formatting
- Maintaining uniform hierarchy and flowchart structure

3. Variability in Industrial Data

- Different plants operate under different temperatures, pressures, catalysts, and configurations
- Need to choose representative “typical” ranges

4. Generating Reliable QA Pairs

- Questions must be diverse and technically relevant
- Answers must remain accurate, consistent, and aligned with the flowchart
- Avoiding contradictory or circular information

5. Scalability

- Dataset expands rapidly as more processes are added
- Requires a robust structure for storing diagrams, metadata, and QAs
- Need version control and quality checks

6. Validation Challenges

- Need systematic **technical validation** to ensure accuracy of Reaction conditions, Catalyst information, Stream compositions, Equipment types, Process sequencing
- Need **diagram validation** to ensure flowchart matches the narrative description
- Need **QA validation** to ensure questions and answers correspond correctly
- Hard to automate because validation requires chemical engineering expertise
- Risk of subtle errors propagating unless carefully reviewed

Reference Papers

1. **FlowchartQA: A Large-Scale Benchmark for Reasoning over Flowcharts**
Link: https://aclanthology.org/2023.limo-1.5.pdf

2. **FlowVQA: Mapping Multimodal Logic In Visual Question Answering**

3. **SlideVQA - A Dataset for Document Visual Question Answering on Multiple Images**

4. **CoG-DQA: Chain-of-Guiding Learning with Large Language Models for Diagram Question Answering**