Framework for Automated PFD and PID Generation

May 13, 2025

Introduction

- **Core Engineering Diagrams (PFDs & PIDs):** Process Flow Diagrams (PFDs) and Piping and Instrumentation Diagrams (PIDs) are standard, essential documents in the chemical process industry. PFDs offer a high-level schematic of material/energy flow and major equipment, showing *what* happens and *where*. PIDs build on PFDs, providing detailed schematics of instrumentation and control systems, illustrating *how* the process operates and is controlled.
- **Foundational Role in Process Development:** PFDs and PIDs serve as foundational documents for chemical process simulations.
- **Advancements in Generative AI:** Recent breakthroughs in generative AI are revolutionizing chemical and materials science by accelerating discovery and streamlining complex simulations. This enables faster innovation, lower R&D costs, and more sustainable product development.
- **The "Transition to Production" Challenge:** Realizing the full potential of AI-driven discoveries requires developing new production processes to scale them from lab/simulation to industrial production. This transition remains a significant challenge.
- **Limitations of Current AI Methods for PFD/PID Generation:** Existing methods are not designed to auto-generate PFDs or PIDs for *novel* industrial-scale chemical production processes. They often overlook essential process context (high-level objectives for PFDs, operational/control details for PIDs). Consequently, they cannot justify design choices or the necessary control and instrumentation logic for safe and efficient operations. Crucially, current methods fail to integrate first-principles simulators to verify the physical and operational feasibility of generated diagrams.
- **The Bottleneck of Manual Creation:** The manual, expertise-intensive creation of novel PFDs and PIDs creates a bottleneck, hampering simulation fidelity, digital twin accuracy, and scalable AI deployment.
- **Our Proposed Solution: A Closed-Loop Framework:** We present a closed-loop, self-driving lab framework for the automated generation of high-fidelity process flow and instrumentation descriptions. This framework is designed to accelerate the development of novel chemical processes.
- **Goal of the Framework:** To significantly expedite the simulation-to-lab-to-pilot-to-plant scale-up pipeline. To ensure that only industrially viable, sustainable, and efficient processes advance to commercialization by automating design, simulation, and optimization with minimal human intervention.

Methodology

- **Integrated Approach Overview:** The methodology combines data curation, knowledge graph construction, specialized language model fine-tuning, retrieval augmentation, inference optimization, and engineering validation to generate and analyze PFDs/PIDs for chemical processes.
- **Data Curation (ChemAtlas Database):** Initiated by curating the ChemAtlas database, containing information on over 1,020 industrial chemicals sourced from manufacturer catalogs.

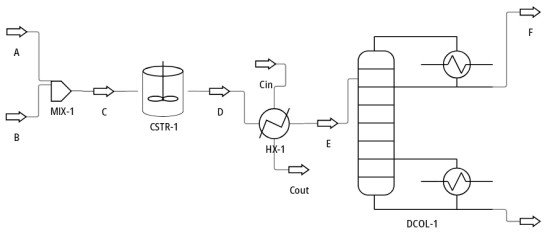

Figure 1: The figure shows a high-level schematic of a chemical process depicting material flow from reactant inlets (A and B) through a mixer (MIX-1), a continuous stirred-tank reactor (CSTR-1), a heat exchanger (HX-1), and a distillation column (DCOL-1), yielding product streams F and G. Major equipment and stream connections are illustrated, excluding instrumentation and control logic. This abstraction facilitates under rmations.

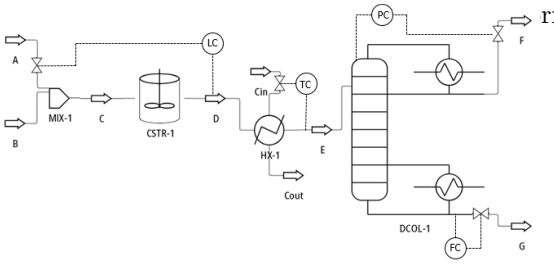

Figure 2: The figure shows the detailed PID of a chemical process showing instrumentation and control systems, including: level control (LC) on reactor CSTR-1 regulating feed A; temperature control (TC) on column feed E adjusting HX-1 utility flow; pressure control (PC) at DCOL-1 overhead controlling product F; and flow control (FC) on bottoms product G. The diagram specifies control strategies and safety-critical parameters.

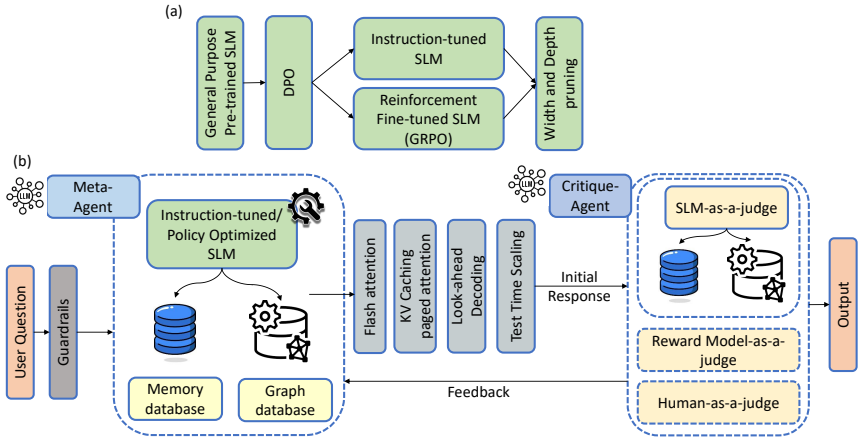

Figure 3: Overview of the integrated framework. (a) The SLM fine-tuning pipeline depicts initial DPO alignment followed by instruction tuning or GRPO reinforcement learning, concluding with optional width/depth pruning. (b) The operational RAG framework illustrates a Meta-Agent coordinating the specialized SLM (from part a), which accesses memory and graph databases for context. The SLM's inference is accelerated via optimizations (FlashAttention, Paged KV Caching, Lookahead Decoding, Test-Time Scaling). Generated responses are refined iteratively through a feedback loop managed by a Critique-Agent employing diverse judges (e.g., Nemotron-4-340B reward model, LLM-as-a-judge like GPT-4o/Haiku, or human evaluation).

- **Automated Information Extraction:** Employed an agentic web navigation framework to autonomously retrieve, interpret, and structure multimodal data into process flow and instrumentation descriptions for chemicals in ChemAtlas.
- **Knowledge Graph (KG) Construction:**
  - Extracted semantic triples (subject-predicate-object) from structured text using GPT-4o.
  - Canonicalized entities by merging those with high semantic (embedding) and string (Levenshtein distance) similarity.
  - Partitioned the KG into hierarchical communities using the Leiden algorithm to optimize modularity and retrieval efficiency.

- **Synthetic Training Data Generation:**
  - Generated over 20,000 synthetic question-answer (QA) pairs using teacher LLMs (GPT-4o, Claude Haiku) via a self-instruct method seeded with human examples.
  - Rigorously validated and filtered these QA pairs using NVIDIA's Nemotron-4-340B reward model to ensure high quality.
- **Specialized Synthetic Datasets:** Created six distinct subsets for targeted training:
  - *Factual QA:* Reinforces foundational chemical process engineering knowledge.
  - *SynDIP:* Comprehensive QA pairs covering PFD/PID descriptions (mimicking agent-retrieved data).
  - *LogiCore:* Multi-step reasoning chains for design justification and control logic validation.
  - *DPO:* Preference-labeled response pairs for Direct Preference Optimization alignment.
  - *Local RAIT:* Document-grounded QA using individual SynDIP process descriptions.
  - *Global RAIT:* QA requiring synthesis across multiple SynDIP descriptions.
- **Benchmark Evaluation Datasets:**
  - Constructed a 1.5K QA out-of-distribution benchmark from ChemAtlas to assess generalization.
  - Introduced the ChemEval dataset (100 held-out chemicals) to test zero-shot generation of full PFD/PID descriptions.
- **Base SLM Selection and Customization:**
  - Selected Small Language Models (SLMs): Llama-3.2-1B and SmolLM-135M.
  - Utilized Quantized Low-Rank Adaptation (QLoRA) for efficient fine-tuning with frozen base model weights.
- **Fine-tuning Strategy 1: Sequential Pipeline:**
  - Step 1: Supervised Fine-Tuning (SFT) on the combined Factual QA, SynDIP, and LogiCore datasets.
  - Step 2: Direct Preference Optimization (DPO) using the curated DPO dataset for alignment.
  - Step 3: Retrieval-Augmented Instruction Tuning (RAIT) using Local and Global RAIT datasets.
- **Fine-tuning Strategy 2: Reinforcement Learning (GRPO):**
  - Applied Group Relative Policy Optimization (GRPO) first to the SFT datasets (Factual QA, SynDIP, LogiCore).
  - Refined the model further using GRPO on the RAIT datasets.
  - Optimized a composite reward function (ROUGE-L, length penalty, LLM-judge score) stabilized by KL divergence regularization.
- **Graph RAG Integration for Inference:**
  - Integrated the fine-tuned SLMs with the structured KG using a Graph Retrieval-Augmented Generation (Graph RAG) framework.
  - Retrieval involves comparing query embeddings to community summaries, retrieving top communities, and constructing a relevant subgraph context.
  - SLM leverages this dynamic subgraph context (entities, relationships, source text) for grounded, multi-hop reasoning.
- **Inference Optimization Techniques:** Implemented a suite of techniques to enhance performance and efficiency:
  - *Structural Pruning:* Reduced model size (width/depth) based on importance heuristics.

  - *PagedAttention & KV Cache Quantization:* Mitigated memory fragmentation and reduced cache footprint.
  - *Lookahead Decoding:* Accelerated generation latency via parallel token speculation.
  - *FlashAttention:* Optimized core attention computation to reduce memory bandwidth bottlenecks.

- **Test-Time Reliability Enhancement:** Employed Test-Time Inference Scaling techniques to improve output reliability:
  - *Self-Consistency Sampling:* Generating multiple outputs and selecting the most consistent.
  - *Confidence-Weighted Entropy Scoring:* Assessing confidence in generated tokens.
  - *Iterative Self-Reflection/Revision:* Refining outputs based on internal checks or critique.
  - *Consensus Aggregation:* Combining information from multiple generated outputs.

- **Engineering Feasibility Validation (DWSIM):**
  - Validated the practical viability of generated descriptions using the DWSIM open-source chemical process simulator.
  - *PFD Validation:* Translated PFDs to flowsheets to verify material/energy balances and thermodynamic consistency.
  - *PID Validation:* Implemented control loops in DWSIM's dynamic mode to assess stability and control performance (e.g., setpoint tracking, disturbance rejection).

- **Overall Framework Summary:** The methodology creates domain-specialized SLMs through KG-augmented data generation, targeted fine-tuning, efficient RAG, optimized inference, and rigorous simulator validation, ensuring industrial relevance and reliability.