

Peer Review

Review of: "Accelerating Manufacturing Scale-Up from Material Discovery Using Agentic Web Navigation and Retrieval-Augmented AI for Process Engineering Schematics Design"

Evgenii Arsentev¹

1. Independent researcher

Summary

The paper proposes a two-stage pipeline for producing Process Flow Diagrams and Piping & Instrumentation Diagrams for chemicals that have no public process documentation. Stage one is a multi-agent web-navigation layer: a meta-agent decomposes a query, routes subtasks to five sub-agents (image, scholar, patent, wiki, web insights) by cosine similarity between the subtask embedding and a written capability description, orders them with a DAG, and aggregates their outputs, with an iterative feedback loop involving human experts, a "Gold-LLM-as-a-Judge" and a "Reward-Model-as-Judge" (section *Agentic Web Navigation Framework*). Stage two turns that text into an ontological knowledge graph — sliding-window chunking with contextual prefixes, LLM-based NER and relation extraction, entity merging by cosine plus Levenshtein, hierarchical Leiden communities, top-K community retrieval — and answers queries over it (section *Graph Retrieval-Augmented Generation*). Evaluation covers 1,020 chemicals for knowledge-base construction, 50 held out for generation, and a 6,000-pair ODQA set (section *Datasets*), scored by the NVIDIA Nemotron-4-340B-Reward model on Helpfulness, Correctness, Coherence, Complexity, and Verbosity, plus BLEU/ROUGE-N/ROUGE-L (section *Results* and *Additional Experiments*). The engineering is real, and the write-up of the architecture is careful. My problem is with what the evidence is asked to carry.

Does the claim hold on the data

First, a statement about the version I am reviewing. **This paper contains no tables.** Every quantitative result lives in Figures 5, 7, 8, 9(a), 9(b), 12, 13, and 14, and in the text supplied to me, those figures arrive as captions only — no axis values, no numbers, no error bars. The only numbers anywhere in the body are dataset sizes, hyperparameters, and the graph size in the Figure 15 caption (8,005 nodes, 22,237 relationships). So every performance claim in the paper — "consistently achieve higher scores," "competitive performance," "significant improvements" — is, for a reader of this version, unverifiable in principle. I am not going to guess at bar heights. Even setting the rendering aside, figures without tabulated values, sample sizes, or intervals are a weak substrate for the conclusions being drawn here, and I would ask for tables regardless.

Second, and more serious: **the appendix argues against the abstract.** The abstract sells an agentic-plus-Graph-RAG framework, and the Conclusion says it delivers "high contextual accuracy." But *Additional Experiments* reports three times over that the baseline wins. On Figure 9(a), GPT-4o and Gemini-1.5 Pro with no Graph RAG "consistently achieve higher scores across all metrics." On Figure 9(b), fine-tuned SmoLM2-360M, SmoLM2-1.7B, and Qwen2.5-1.5B "achieve higher scores across all metrics compared to their pre-trained counterparts with Graph RAG." On Figure 12, fine-tuned models "consistently outperform" Graph RAG on BLEU and ROUGE, and Figure 13 repeats it. The paper's own answer to its central question — does structured retrieval help here — appears to be "less than fine-tuning does," and that finding is buried in an appendix and reframed as cost-efficiency. Reframing it is legitimate; leaving the abstract and Conclusion as they stand is not. The one comparison that does favour the method (Figure 14, Graph RAG vs. traditional RAG) is the one with no numbers at all in the text and the word "significant" attached to it with no test.

Third, **the headline word "regulation-compliant" is never operationalized.** It appears in the abstract, in the Introduction, and in the Conclusion. No standard is named anywhere in the paper — not ISA-5.1, not IEC 62424, not any jurisdiction's code. No compliance check is defined, run, or reported. The metrics actually used (Helpfulness, Correctness, Coherence, Complexity, Verbosity; BLEU; ROUGE) do not measure compliance with anything. The Introduction motivates the work on "regulatory compliance, quality control, and safety in plant operation," and the pipeline's web-insights sub-agent draws on "blogs, forums, and other relevant websites" (section *Agentic Web Navigation Framework*). A P&ID assembled from forum posts and scored by a reward model for "helpfulness" has not been shown to be safe or compliant,

and the claim should be withdrawn or backed with an engineer-run compliance audit on a named standard.

Fourth, **what is generated is text, not diagrams**. Section *Experimental Settings* says the fine-tuning set was "an instruction-following dataset for the task of producing descriptions of PFDs and PIDs," and all reported metrics are text-quality metrics. The title and abstract promise "Schematics Design" and "diagram generation." If the system emits a structured artifact (DEXPI, SVG, graph of equipment and streams), show it and evaluate it structurally — equipment recall, stream connectivity, instrument-tag correctness. If it emits prose about diagrams, say so plainly in the title and abstract.

Fifth, **the ground truth is partly the system's own supplier**. The ODQA set has answers "produced by benchmark LLMs like GPT-4o" (section *Datasets*). GPT-4o is also the synthesis engine inside the sub-agents (Figure 4 caption) and the triple extractor for the knowledge graph (*Additional details*). BLEU/ROUGE in Figure 12, therefore, measure agreement with the model that generated the pipeline's own inputs. The same circularity runs through Figures 7 and 8, where "closeness to GPT-4o" is treated as evidence of quality — it is evidence of stylistic convergence, nothing more. Section *Experimental Studies* (a) also promises a comparison of auto-generated diagrams "to ground-truth data" for the 50 evaluation chemicals, but the provenance of that ground truth is never described, and I cannot find any figure reporting that comparison. For chemicals selected precisely because public process documentation is scarce, where the reference diagrams come from is the single most important methodological question in the paper.

Sixth, **the judge is never validated and may be inside the loop**. Every quality number comes from Nemotron-4-340B-Reward (sections *Results* and *Additional Experiments*). There is no agreement study against the human experts the paper says are available, no inter-rater statistics, no spot-check of "Correctness" against a chemical engineer. Worse, the refinement loop in section *Agentic Web Navigation Framework* includes a "Reward-Model-as-Judge" providing feedback F_i until quality standards are met. The paper never states whether that reward model is the same Nemotron model used for evaluation. If it is, the outputs were optimized against the metric that scores them, and the numbers mean very little. Please state this explicitly either way.

Reproducibility

The fine-tuning recipe is reported well — section *Experimental Settings* gives a learning rate of 5×10^{-5} , batch size 8 per device, 3 epochs, 4 gradient-accumulation steps for an effective batch of 32, LoRA $r=4$ and

$\alpha=16$, dropout 0.1, FP16, NVIDIA Tesla T4, PyTorch, and Unsloth. Credit where it is due; most papers of this kind give less. Everything around it is missing.

No model snapshots: "GPT-4o," "Google Gemini-1.5 Pro," and "Anthropic-Haiku" appear with no dated versions anywhere (*Results, Additional Experiments*, Figure 7), and "Anthropic-Haiku" does not name a specific model at all. No temperature or decoding settings for any generation step. No seeds, no number of repeated runs, no confidence intervals, no variance of any kind. No retrieval date for the SerpAPI crawls — for a paper whose first stage is live web navigation, the crawl window is the single most perishable dependency, and without it, the knowledge-acquisition stage cannot be re-run by anyone, including the authors. No SerpAPI configuration (result depth, geo, engine). No top-K for community retrieval, no perplexity for t-SNE, no clustering criterion behind "clustering to identify the optimal number of groups" (*Results*), and no resulting k . $N_{\max}$ for the feedback loop is defined symbolically and never given a value. The human experts are never counted, characterized, or reported on — how many, what they changed, how often they intervened — which also means "minimal expert intervention" (abstract) has no measurement behind it. No code or data availability statement, and the 1,070-chemical dataset and 6,000-pair ODQA set are described as curated but not released. No cost or latency figures, despite cost-efficiency being the fallback claim of the whole appendix.

Specific points

1. **Abstract and Conclusion vs. Figures 9(a), 9(b), 12, 13.** Align the top-line claims with the appendix's own result — fine-tuned models without Graph RAG beat Graph RAG with pre-trained models on all five metrics and on BLEU/ROUGE. Recast the contribution as cost-constrained deployment, and state the trade-off with numbers.
2. **Section *Additional details, chunking arithmetic*.** "The window size (w) is set to 1024 tokens... while the stride (s) is set to 128 tokens, maintaining a 12.5% overlap." With the definition given in section *Graph RAG* ($C_i = D[(i-1)s : (i-1)s + w]$), a stride of 128 against a window of 1024 gives 896 shared tokens, i.e., 87.5% overlap, and encodes each token in roughly eight chunks. 12.5% is s/w , not the overlap. Either the number is wrong or the stride is, and the difference is an eightfold change in embedding and extraction cost — which matters because the sentence justifies the setting by "computational efficiency."
3. **Entity-merge thresholds, section *Graph RAG* vs. section *Additional details*.** The method defines a normalized string ratio $\text{str_sim} = 1 - d_{\text{lev}}/\max(|e|, |e'|)$ and requires $\text{str_sim} \geq \tau_{\text{str}}$. The appendix

instead reports "a Levenshtein edit distance limit of 5," an absolute bound. These are different criteria with very different behaviour on long entity names ("sodium hydroxide solution" vs. a three-character tag). State which one was actually run, and report τ_{str} in the units of the criterion used. The cosine threshold of 0.9 is given clearly.

4. **Section *Experimental Settings*, self-contradiction on inference time.** "Graph RAG with a pre-trained LLM... resulting in slower overall inference times because of graph traversal... In contrast, a fine-tuned LLM without Graph RAG offers faster inference... The fine-tuned LLM's inference time is slightly higher due to its larger, task-specific parameterization." The last sentence reverses the first two. Also, LoRA at $r=4$ merged back into SmolLM2-360M does not produce a "larger" model at inference. No latency is measured anywhere in the paper; either measure it or drop the paragraph.
5. **Section *Experimental Studies*, item (a) has no reported result.** Four evaluation tasks are listed; (b), (c), and (d) map onto Figures 12/13, 9(b), and 14. The comparison of generated PFDs and PIDs against ground truth for the 50 evaluation chemicals appears nowhere. Add it, and describe how the ground truth was obtained and by whom.
6. **Section *Results*, scale definition.** "Each attribute was evaluated on a four-point Likert scale (continuous), with 0 indicating the lowest quality and 4 indicating the highest." A 0-to-4 range is five points if discrete, and "continuous Likert" is a contradiction. State the actual output range of the reward model and whether scores were averaged, and over how many items.
7. **"Significant" used without a test, twice.** Section *Additional Experiments*: "fine-tuning and increased model size significantly enhance performance" (Figure 12) and "Graph RAG... achieves significant improvements over traditional RAG" (Figure 14). With one run per configuration, no seeds, and no intervals, these should read "higher on our single run" or be supported by repeated runs and a paired test.
8. **Figure 9(a) compares unequal budgets under one chart.** SmolLM2-360M plus Graph RAG against GPT-4o and Gemini-1.5 Pro is a comparison across roughly three orders of magnitude in parameters, plus a retrieval stage on one side only. The honest ablation is the same base model with and without Graph RAG at matched decoding settings; that isolates the contribution the paper is arguing for. Report tokens and cost per query alongside, since cost-efficiency is the claim.
9. **Figures 7 and 8 measure the wrong thing.** Embedding similarity between the framework's output and GPT-4o output is not a correctness signal, particularly when GPT-4o is inside the framework (Figure 4 caption). The sentence "outliers for both models indicate instances of weak similarity"

(section *Results*) also applies a threshold that is never stated. Either drop these figures or reframe them explicitly as a style-convergence diagnostic.

10. **Figure 11 sub-captions are swapped relative to the text.** The captions read "Figure 11(a). PCA plot (GPT-4o)" and "Figure 11(b). t-SNE plot (GPT-4o)", while the surrounding paragraph says "The t-SNE visualization (Figure 11a) highlights clusters... The PCA plot (Figure 11b) illustrates the variance distribution." Figures 6 and 10 follow the a=t-SNE, b=PCA convention; Figure 11 breaks it. Fix the labels.
11. **Over-reading t-SNE, section *Results*.** "PFDs and PIDs within a cluster share common characteristics, suggesting similar industrial designs... This clustering analysis provides valuable insights for process optimization, design, and risk assessment." t-SNE cluster sizes and inter-cluster distances are not interpretable, and no cluster was validated against any process-engineering label. If clusters correspond to real process families, show that against a labelled subset; otherwise, present the plots as qualitative and stop there.
12. **Safety and provenance are not discussed.** The Introduction rests the motivation on safety and compliance, and the pipeline ingests diagram images and forum text scraped from the open web. There is no filtering for outdated or incorrect source diagrams, no hazard discussion, no false-instrument or missing-interlock analysis, and no note on the licensing of scraped diagram images. A short limitations section covering the failure modes of an auto-generated P&ID would strengthen the paper considerably and is, in this domain, close to mandatory.
13. **Dataset details, section *Datasets*.** "over 1,070 chemicals" resolves to exactly 1,020 + 50; drop the "over." How the 50 evaluation chemicals were selected is not stated (random, stratified, hand-picked). Whether web retrieval for the 50 could surface the same sources used to build the primary knowledge base is not addressed — worth a sentence, since it bears on whether they are genuinely "unseen." For the 6,000 QA pairs, "each QA pair underwent meticulous validation before inclusion" needs a procedure and a validator: by whom, against what, and how many pairs were rejected.
14. **Missing citations.** SmoLLM2 is cited inline as "[Allal et al., 2024]" in section *Experimental Settings* but has no entry in the numbered reference list, which ends at 12. Nemotron-4-340B-Reward, Qwen2.5, the Leiden algorithm, Neo4j, Unsloth, t-SNE, SerpAPI, and text-embedding-3-small are all used and none is cited.
15. **Algorithms 1 and 2 are referenced but absent from the version I received** (sections *Agentic Web Navigation Framework* and *Graph RAG*). If they were dropped in production, please restore them; the routing and retrieval logic is hard to reimplement from prose alone.

16. **Uncited market figures in the Figure 1 and Figure 2 captions** — "\$10 billion in 2024" for lithium hydroxide and "approximately \$2 trillion in 2024" for refining. References [9] and [10] cover the diagrams, not the market sizes. Add sources or remove.
17. **Figure 15 caption** gives 8,005 nodes and 22,237 relationships — the only quantitative artifact in the paper that isn't a plot. Please add the extraction-quality numbers that go with it: triples sampled, precision against manual review, and how many entity merges the 0.9 / Levenshtein-5 rule made and how many were wrong. That is the measurement that tells a reader whether the knowledge graph is any good.
18. **Minor.** The multimodal model used by the image sub-agent is never named outside a parenthetical "e.g., OpenAI GPT-4o" in the Figure 4 caption. Top-K for community selection is never given a value. Section *Additional details* caps triples per chunk at M=20 with no justification or sensitivity check.

Recommendation

Major revisions. The architecture is worth publishing, and the problem is a real one, but three things have to change before the claims are supportable: the quantitative results need to be tabulated with sample sizes, repeats, and intervals rather than left in unlabelled figures; the abstract and Conclusion need to match the appendix's own finding that fine-tuned models without Graph RAG scored higher on every reported metric; and "regulation-compliant" must either be tied to a named standard with an audit behind it or dropped from the paper. The circularity between GPT-4o as generator and GPT-4o as reference, and the unvalidated reward-model judge that may also sit inside the refinement loop, need to be resolved in the same pass.

Declarations

Potential competing interests: No potential competing interests to declare.