

Peer Review

Review of: "Chain of Ideas: Revolutionizing Research via Novel Idea Development with LLM Agents"

Evgenii Arsentev¹

1. Independent researcher

Summary

The paper argues that the way literature is *organized* in the prompt matters more than how much of it you dump in. Concretely: instead of handing an LLM a pile of retrieved papers, the Chain-of-Ideas (CoI) agent picks an anchor paper for a topic, walks forward through its citing papers and backward through its references, and turns that path into a chain of ideas plus a written account of how each idea evolved into the next (Sec. 2.2). The LLM then predicts a future trend along that chain, drafts an idea, passes it through a novelty checker, and finally produces an experiment design (Secs. 2.3-2.4). Three such chains are built per topic ($K = 3$, Sec. 3.1), and the best draft is picked by pairwise comparison. Evaluation is by "Idea Arena" (Sec. 3.4): a round-robin of pairwise LLM-judge and human judgments over 50 topics drawn from Hugging Face Daily Papers submitted 1 August - 15 September 2024 (Sec. 3.2), converted into ELO scores. The headline claims are that CoI beats the second-best automated baseline by 108 ELO under the LLM judge and 56 ELO under human judges (Sec. 4.1), that its ideas are on par with real papers and ahead of them on Novelty and Significance (Sec. 4.1), that the judge agrees with humans 70.8% of the time on average (Table 1, "Average" column), and that a candidate idea plus experiment design costs from \$0.50 (Abstract).

The core intuition is a good one, and the ablation in Table 2 does support the part of it that matters most. What the paper does not yet establish is that the reported margins measure the chain structure rather than the extra compute, the extra selection step, and a judge that was chosen on the same data it is then validated against.

Does the claim hold on the data

Two structural problems sit between the numbers and the conclusion.

The compared systems do not have the same budget. Sec. 3.3 defines fairness as "GPT-4o and Semantic Search as the LLM and academic retriever implementations" across all methods. That equalizes the model, not the spend. CoI builds three chains, each up to five papers long (Sec. 3.1), extracts ideas and trends from each paper, runs a novelty-check loop, generates three candidate ideas, and then *selects among them by pairwise comparison* (Sec. 2.3, last sentence). RAG, by contrast, is one prompt over retrieved literature (Sec. 3.3, first item). So the reported gap over RAG is a best-of-3-with-refinement system against a single-pass system. Worse, the selection step "follows the same method as Idea Arena" (Sec. 2.3) – the criterion used to pick the submission is the criterion used to score it. Selecting the argmax under a judge and then reporting that argmax's rating under the same judge inflates the rating by construction. The size of that inflation is not estimated anywhere. Sec. 4.6 actually names the mechanism without recognizing it: "generating multiple chains primarily helps reduce the impact of any single CoI performing poorly" is a description of best-of-K.

The judge is validated on the data used to select it. Sec. 4.1 compares GPT-4o, Gemini-1.5-Pro-Exp-0827, and Claude-3.5-Sonnet against human preferences (Figure 5), picks GPT-4o because it agrees most, and then reports the same 70.8% (Table 1, "Average") as evidence that Idea Arena is robust. Picking the maximum of three and reporting it as an unbiased estimate is optimistic; the number needs a held-out split of the human judgments, or at minimum an honest note that it is a maximum. It also needs two reference points it does not have: human-human agreement on the same matches, and the chance rate given that ties are allowed. Without those, 70.8% is not interpretable as "strong alignment." Note also that the weakest cell in Table 1 is Novelty at 66.5% – and Novelty is exactly the dimension that carries the claim of beating real papers.

On the human side, the blinding is incomplete in the one arm that matters. The 10 researchers select the papers that "capture their interests" (Sec. 3.2), and those same 10 researchers judge the anonymized ideas (Sec. 3.4). Anonymization hides the *method*, but a researcher who nominated a paper will often recognize the idea extracted from it. That is the Real Paper arm, i.e., the reference point for the "comparable to humans" claim.

The Real Paper baseline is not a human artifact. Per Sec. 3.2, GPT-4o extracts the topic, the idea, and the experiment design from the selected paper (Tables 21-23), and those extractions "will be utilized as our

Real Paper baseline". So the comparison is against a GPT-4o summary of a paper, judged by GPT-4o. There is a second asymmetry on top: the topic was derived *from* that paper, so the real idea is by definition the already-published answer to the topic, and a judge asked "are the problems or approaches new?" (Table 5, Novelty) has a systematic reason to score a familiar-looking idea lower than an unfamiliar one. Beating the Real Paper on Novelty under these conditions is not evidence of human-level novelty.

The abstract overstates what Sec. 4.1 reports. The abstract says CoI "shows comparable quality as humans in research idea generation", unqualified. Sec. 4.1 says the agent "still lags considerably behind the Real Paper" on Clarity, Feasibility, and Expected Effectiveness in human evaluation, and Sec. 4.4 repeats that its experiment designs "tend to be less feasible". Comparability holds on two of five dimensions against an LLM-extracted proxy. The abstract should say so.

Effect sizes have no error bars. 108, 56, and 70 ELO points (Secs. 4.1 and 4.4) come from a round-robin over 50 topics. ELO from a small fixed set of matches depends on the K-factor, the initial rating, and the order in which matches are processed, none of which is stated, and none of the three margins carries an interval or a match count. One internal signal suggests the ranking below CoI is unstable: in Sec. 4.1, the second-best method is GPT-Researcher under the LLM judge and RAG under human judges. If the identity of second place flips with the judge, the distance to second place is a shaky quantity to headline.

The ablation contradicts its own noise floor. Table 2 scores each variant against the full agent, 2/1/0 over 50 matches, so 50 is the tie line (caption and Sec. 4.2). For the "- Future Trend" row, Clarity is 51 and Feasibility is 53 - above the tie line - and the text dismisses these as "not substantial, likely due to evaluation variability". Accept that, and you have set the noise band at roughly +/-3. The same row's Average is 46.2, and "- Entities" averages 45.4; both are within about one and a half times that band. Only "- CoI" (Average 42.4, Novelty 41, Significance 39, Effectiveness 39) clearly clears it. The honest reading of Table 2 is that the chain structure is supported and the other two modules are not yet distinguishable from noise. No significance test, no repeated runs, and the judge here is GPT-4o alone (Sec. 4.1 states GPT-4o is used for all subsequent analysis), so the ablation inherits whatever bias the judge has. The arithmetic in Table 2 does check out: the four Average values are the exact means of their rows.

A leakage channel is left open. Sec. 3.2 says the 1 Aug - 15 Sep 2024 window is "after the data cutoff of the LLM" - the cutoff date itself is never named, which is the easy fix. The harder issue is retrieval: Sec. 2.2 extends each chain forward by pulling papers that cite the anchor from the live Semantic Scholar API, and the topics were extracted from papers in that same recent window. Nothing in Sec. 2.2 or Sec. 3.2 describes excluding the source paper, or papers citing it, from the retrieval pool. If the source paper can

be retrieved, the agent can be rewarded for recovering the idea it was implicitly given, and the novelty checker (which compares only against retrieved papers, Sec. 2.3) will not catch it. A de-duplication step and a count of how often the source paper entered a chain would settle this.

Reproducibility

Better than average in some respects. Model snapshots are named - GPT-4o (05-13) and GPT-4o-mini (07-18), Sec. 3.1 - the embedding model is named (text-embedding-3-large, Sec. 2.2), the retriever is named, the main configuration is given (chain length 5, K = 3, one refinement iteration, Sec. 3.1), and code and a demo are released (footnote 1).

What is missing is the part that decides whether a rerun lands in the same place. No temperature, no `top_p`, no seed, no number of samples per call anywhere in Sec. 3.1 or elsewhere. No snapshot or access date for the Semantic Scholar index - and since the chains are built by live retrieval over a citation graph that grows daily, the pipeline is not reproducible without one; a rerun in 2026 will construct different chains for the same topic. The ELO procedure is cited to Zheng et al. and Zhao et al. (Sec. 3.4) but its parameters are not given. The number of human pairwise comparisons is never stated, nor how the 50 topics were distributed across the 10 researchers, nor whether humans also judged each pair twice with reversed order (Sec. 3.4 states the order-reversal only for the LLM judge), nor any inter-annotator agreement among the humans themselves. The novelty-check loop has no stated iteration cap and no report of how many iterations actually ran (Sec. 2.3). There is no accounting of failed or unparseable runs - if a chain could not be built or an output did not parse, the reader cannot tell whether it was retried, dropped, or counted.

The \$0.50 figure (Abstract) is described as a *minimum*, which makes it a best case rather than a property of the system as evaluated. The configuration that produces it is not identified, and it is not obviously the configuration of Sec. 3.1. Since the cost argument is one of the paper's selling points, it deserves a small table: cost per topic per method under the main configuration, with token counts, so that the budget asymmetry in Sec. 3.3 becomes visible rather than implicit.

One practical note on the version I received. Figures 3, 4, 5, 6, and 7 and Tables 4 and 26 arrived as captions with no content, and the prompt tables (Tables 7-25) arrived as a list of titles without the prompt bodies. Every ELO score in the paper lives in those objects, so I could not verify 108, 56, or 70 against a table, nor read the length and width curves of Secs. 4.5-4.6, nor inspect the judge prompt. I have not treated this as the authors' fault, and I have invented no values, but it does mean the paper as rendered

here carries no numeric evidence outside Tables 1 and 2. Putting the key ELO values inline in the text, with match counts, would make the paper robust to this.

Specific points

1. **Sec. 2.3 (final branch selection) with Sec. 3.1 ($K = 3$) against Sec. 3.3 (baselines).** Add a budget-matched control: run RAG, ResearchAgent, and GPT-Researcher with three samples each and the same pairwise selector, and report the resulting ELO alongside. If Col's margin survives that, the claim is much stronger. If it shrinks, the paper should say by how much.
2. **Sec. 2.3 and Sec. 3.4.** Selection and evaluation share a protocol and a judge. At minimum, select among branches with a different judge or a different criterion than the one used for scoring, and report the gap between "random branch" and "selected branch" ELO – that difference is the size of the selection effect.
3. **Abstract versus Sec. 4.1 and Sec. 4.4.** Qualify "comparable quality as humans". Sec. 4.1 reports a considerable lag behind Real Paper on Clarity, Feasibility, and Expected Effectiveness under human judges; Sec. 4.4 reports lower feasibility for the designs.
4. **Sec. 3.2 (Tables 21-23) and Sec. 3.3, Real Paper item.** The baseline is a GPT-4o extraction of the paper, not the paper. Say so plainly in Sec. 3.3, and consider a control where the abstract and method section are used verbatim, so the extraction step is not confounded with the comparison.
5. **Sec. 3.2 with Sec. 3.4, Human Evaluation.** The researchers who nominated the papers also judge the ideas derived from them. Report how often a judge was assigned a topic they nominated, and ideally re-run the Real Paper comparisons with judges who had no role in selecting that topic.
6. **Sec. 4.1, Figure 5, and Table 1.** GPT-4o is selected as judge for having the highest agreement, and the same statistic is then cited as evidence of reliability. Split the human judgments: choose the judge on one half, report agreement on the other. Add human-human agreement and the chance rate under the tie-allowing protocol. Table 1's lowest cell, Novelty at 66.5%, should be discussed rather than folded into the 70.8% average, since Novelty carries the central claim.
7. **Sec. 4.1 and Sec. 4.4.** Report confidence intervals for all ELO scores (bootstrap over topics is standard for arena-style evaluation), plus the K-factor, initial rating, and number of matches. Note explicitly that second place changes identity between the two judges – GPT-Researcher under the LLM judge, RAG under humans – and what that implies about the resolution of the ranking.

8. **Table 2 and Sec. 4.2.** The text treats +1 and +3 above the tie line ("Future Trend": Clarity 51, Feasibility 53) as noise. Apply that standard consistently to the Average column: "Future Trend" 46.2 and "Entities" 45.4 are not far outside it, while "CoI" 42.4 is. Either add a test and repeated runs, or soften the conclusions about the Future Trend and Entities modules. The claim that the CoI construction stage is the dominant component does survive.
9. **Sec. 2.2 (forward extension via citing papers) with Sec. 3.2 (topics from Aug-Sep 2024 papers).** State whether the source paper and its citing works are excluded from retrieval, and report how often the source paper appeared in a constructed chain. Also name the assumed data cutoff date for GPT-4o (05-13) rather than asserting the window is past it.
10. **Abstract, \$0.50.** Replace "minimum cost" with the cost of the configuration actually evaluated in Sec. 3.1, broken down per stage, with token counts, and state the cap on novelty-check iterations (Sec. 2.3 gives none).
11. **Sec. 2.2, milestone definition.** Terminating the backward walk at "over 1,000 citations" is arbitrary and strongly dependent on field and paper age; for a 2024 topic, almost nothing in the recent chain will qualify, so the cap in Sec. 3.1 does the real work. A brief sensitivity check, or at least an acknowledgement, would help.
12. **Sec. 4.5.** The saturation point ("levels off after reaching a maximum length of 5") coincides with the default chosen in Sec. 3.1. Also, the 0-length variant is the "CoI" ablation with 5 retrieved papers, which differs from the length-3 and length-5 variants in both organization and total literature volume, so the 0-to-3 jump confounds the two. A control with equal paper counts and only the chain structure removed would isolate the effect.
13. **Sec. 4.6.** "Positive correlation" is asserted while the text also concedes "the disparity in ELO scores across different branch numbers is small". Give the numbers inline with intervals, and state that the mechanism proposed there - reducing the impact of a poor chain - is selection over K candidates, which raises arena scores independently of chain quality.
14. **Secs. 3.1-4.6, run accounting.** Report the number of topics for which a chain could not be constructed, outputs failed to parse, or a stage errored, and how those cases were handled. As written, the reader cannot rule out averaging over surviving runs.
15. **Sec. 3.3 and Sec. 3.4, citation support.** The claim about evaluation preference toward more structured outputs cites [19], which in the reference list is Chiang et al., *Chatbot Arena*; the verbosity and formatting bias result is normally attributed to Zheng et al. [22]. In Sec. 3.4, "following the

practices outlined in Zheng et al.[22][23]" attaches Zheng's name to [23], which is Zhao et al., *Auto Arena*. Both look like slips rather than substance, but they should be fixed.

16. **Sec. 2.2, cross-reference.** The query-generation prompt is said to be "given in Table 6 of Appendix", but the Appendix listing in A.2 assigns that prompt to Table 7, and Table 6 is the experiment-design metric definitions. Check the whole table-numbering chain.
17. **Terminology.** Sec. 3.4 names the design criteria "Feasibility, Technical Quality, and Clarity"; Table 6 calls the middle one "Quality". Pick one. Sec. 3.3 writes "Semantic Search" where Sec. 3.1 writes "Semantic Scholar".
18. **Sec. 4.3, Table 3.** The case study takes the paper's own topic, builds a chain whose I_0 is [9] and whose I_1 is [7] – both baselines or key references of this paper – and concludes the resulting "EvoResearchAgent" idea is "viable and novel". That verdict is the authors' own, on their own system's output, and evolutionary search over LLM-generated candidates is not an empty area. Either drop the novelty claim from the narration or run the idea through the same blind arena as everything else.
19. **Copy editing.** Sec. 6 contains a stray character: "enhancing their ideation capabilities.p To comprehensively". Sec. 4.6 says "Figure 4.6 shows" where Figure 7 is meant. Sec. 2.3 writes "COI" once. Appendix A.2 has "Prompts used in idea generatio". Table 5's Significance definition reads "Are the idea important?"; Table 6's Clarity definition reads "Dose it provide". Finally, the byline in this version lists a single author, "Li Long", while the corresponding author is Lidong Bing and the Notes refer to "Long Li" – the author list looks truncated, which the editorial team may want to check.

Recommendation

Major revisions. The central idea is sound, and Table 2 ablation genuinely supports it, but the headline comparisons place a best-of-three pipeline with an internal selection round against single-pass baselines under the same judge used for that selection, the judge's reliability is estimated on the data that chose it, the human reference is an LLM extraction judged partly by the people who nominated it, and no margin carries an interval – so the reported gaps cannot yet be read as a property of the chain structure. These are fixable with budget-matched controls, a held-out judge validation, bootstrap intervals, and a leakage check, none of which requires rebuilding the method.

Declarations

Potential competing interests: No potential competing interests to declare.