Jan 30th, 2025

Dear editors-in-chief of Neurosymbolic Artificial Intelligence:

We appreciate the insightful feedback provided by the reviewers on our submitted paper titled "Cognitive LLMs: Toward Human-Like Artificial Intelligence by Integrating Cognitive Architectures and Large Language Models for Manufacturing Decision-Making." The comments were very helpful to enhance the quality of our work.

In this letter, we first address the general feedback provided by each reviewer. In the attached appendix titled "Changes Made to Paper", we present lists of each comment from the reviewers alongside the corresponding changes we have implemented in the paper.

Responses to General Comments from Reviewer 1 (Major Revision):

First, we acknowledge the feedback regarding the *unclear and repetitive writing*, as well as the *lack of organization*. We have taken the following actions, resulting in a reduction of the paper's length from 29 pages to 23 pages (15 main + 3 references + 5 appendix):

(1) **Delete redundant writing:** We ensured that no same content appears more than once.

(2) **Delete repetitive figures:** Figures with unclear annotations or limited information were either removed or redesigned. For example, the newly created Fig. 1 synthesizes the previous Fig. 1 and Fig. 7, with a detailed description provided in the introduction. This redesigned Fig.1 offers a clear, concise overview of the Cognitive-LLMs architecture at the outset of the paper.

(3) **Streamline and reorganize the paper:** We streamlined and rewrote the abstract. Additionally, the paper now has a streamlined organization comprising the following sections in sequential order: introduction, research questions, related work, Cognitive-LLMs architecture, LLM-ACTR knowledge transfer framework, experiments conducted to address research questions, results, conclusions, and discussions.

(4) **Revision of unclear notations:** We have rewritten sections with unclear notations, as highlighted by Reviewer 1. For example, the section on Reinforcement Learning in Production Systems has been rewritten and retitled to include a clearer explanation of utility update theory incorporating metacognition.

Second, we understand the reviewer's concerns on the *comparative performance benefits of fine-tuning LLMs with ACT-R traces versus using ACT-R models alone*. To address this, we have clarified why ACT-R alone is not the baseline in this study by explaining the role of ACT-R in Cognitive LLMs and the rationale for our choice of using pre trained LLMs as the baseline.

(1) **ACT-R as a synthetic agent to instruct LLMs through training:** We have clarified the motivation behind creating Cognitive LLMs, we aim to develop a hybrid architecture, Cognitive-LLMs, that leverages the natural language processing and generative capabilities of LLMs, complemented by the human-like learning and reasoning offered by ACT-R. Therefore, we propose a synergistic approach where ACT-R models serve as synthetic agents guiding the training of LLMs. In this context, ACT-R is used as a grounding tool for enhancing LLMs' trustworthy inference rather than as a baseline.

(2) **Using pre-trained LLMs as baseline:** To assess the Cognitive LLMs' ability to make human-like decisions akin to ACT-R, the baseline for comparisons in this study hence is pre trained LLM to demonstrate the enhancements brought by integrating cognitive architectures into LLMs.

Third, we understand the reviewer's *critique regarding the value of our methodology* and have improved content to explain how our approach of integrating CAs and LLMs differs from others, as well as the contributions it offers.

(1) **How cognitive LLMs differ from other integration approaches:** Following recent findings that LLMs' embeddings can be trained to predict human behaviors, this paper adopts an approach by leveraging CAs to ground the decisions of LLMs in a data-driven manner using machine learning and deep learning methods. Our aim is to examine the properties of a neural network representation of the decision-making process in CAs and to investigate whether knowledge from CAs can be preserved in an embedding space and infused into LLMs through transfer learning.

(2) **Why it matters:** Transfer of learning has proven effective in applications such as text sentiment and image classification. Our experimental results show that the knowledge of CAs in decisions such as learning can be transferred to LLMs through fine-tuning, and the holistic cognitive decision process has the potential to be transferred through finetuning and activation engineering. The results open up new research directions for equipping LLMs with the necessary knowledge to computationally model and replicate the internal mechanisms of human cognitive decision-making from a data-driven perspective.

Last but not least, we appreciate the referenced literature provided by the reviewer. We have carefully reviewed these sources and integrated selected ones, as listed in the appendix.

Responses to General Comments from Reviewer 2 (Accept):

First, we revised the confused reference to VSM-ACTR 2.0 as VSM-ACTR uniformly in the paper and instead cite the previous version of the model as VSM-ACTR 1.0 to avoid confusion.

Second, the scope of this work has been defined as toward trustworthy decision-making by LLMs in manufacturing. We ask whether LLMs can replicate cognition from Cognitive Architectures (CAs) to make human-like decisions.

Third, the empirical results primarily use negative log-likelihood, which is a common chosen measurement of goodness of fit in machine learning. Some of the results are empirically significant, e.g., the LLM with fine-tuning compared to the pre-trained LLM. However, preliminary results show limited improvement and warrant further investigation. We candidly discuss this, proposing possible reasons and pointing out potential solutions.

Last but not least, we appreciate the reviewer's suggestion on discussing the approach's theoretical limitations. We addressed this through rewriting limitations and further work section.

Thank you very much and please let us know if you have any questions!

Sincerely Yours,

Siyu, Alessandro, Jonathan, Lee, and Frank

Changes Made to Paper

In this table, we list each comment and the corresponding revisions provided. If a revision is too extensive to present in full, we refer to the page and line numbers for clear reference.

Reviewer 1	**Revisions Made**
Abstract appears to be too long and too wordy.	Rewrite abstract **(see page. 1, line 27- 39)**
There is often no clear enough take-home message in many sections despite its length.	**Introduction (page 2, left column line 45-51, right column line 27-38):** we have delineated the primary takeaway of the paper as advancing trustworthy decision-making by large language models (LLMs) in manufacturing. Specifically, we explore whether LLMs can replicate cognition from Cognitive Architectures (CAs) to make human-like decisions. We propose 'Cognitive LLMs' as a solution, which are hybrid decision-making architectures consisting of a CA and an LLM, developed through a knowledge transfer framework named LLM-ACTR. **Related work (pages 4-6):** we have added a takeaway for each subsection. For example, *from page 5, line 51, to page 6, line 6*, we conclude the section on integrating CAs and LLMs with the following statement. “This present study builds upon previous research; however, we have adopted a different perspective by leveraging CAs to ground the decisions of LLMs in a data-driven manner. We aim to examine the properties of a neural network representation of the decision-making process in CAs and investigate whether knowledge from CAs can be preserved in an embedding space and infused into LLMs through transfer learning”. **Results (pages 11-12):** each result concludes with a clear takeaway message. For example, *on page 12, left column, from line 18 to line 24*, we state: “This demonstrates that the semantics of symbolic and subsymbolic representations of cognitive models can be learned using a neural network. The principal components retained successfully capture the essential variance related to these cognitive processes, providing a way to preserve cognitive decision-making knowledge in a compact embedding space”. **Main insights/takeaways (from page 13, right column, line 40 to page 14, left column, line 42):** the section has been rewritten to present the takeaways in itemized order.

The authors thus should also restructure the paper, re-organizing the most relevant materials, and removing less relevant or irrelevant materials...Some less relevant but useful materials may be relegated to appendices. On p.8, …	**Paper structure (pages 2-3 and 10-12)**: reconstructed the paper's structure by putting the research questions ahead. And experiments and results sections have been reorganized to address the RQs sequentially. **Figures (e.g., Fig.1 on page 2, Fig. 3 on page 8, Fig. 10 on page 18):** we deleted repetitive figures and redesigned them to be more informative. e.g., Fig. 1 is now a synergy of the previous Figs 1 and 7, and Fig. 10 has had repetitive steps removed. **Content deletion and rewriting (across entire papers and especially on page 8, as highlighted by the reviewer).**: Removed less related content such as "Dopaminergic signals are believed to transmit reinforcement information to the corpus striatum." Rewrote the section on reinforcement learning in production systems to "Foster Metacognition to Support Learning," with a clearer explanation of utility update theory incorporating metacognition.
Is there any performance advantage in fine-tuning LLMs with ACTR traces, compared with the original ACTR model from which traces were obtained?	**Why ACT-R alone is not the baseline in this study:** **Pages 2-3:** explain ACT-R's role in Cognitive-LLMs. **Page 5, left column, line 12 -21**: we state "However, ACT-R do not have LLM-like dialogic interaction with ACT-R models which limits their usability for decision-making. Intuitively, a solution could take the best of both CAs and LLMs, where ACT-R models serve as synthetic agents to instruct LLMs. They do this by providing knowledge of cognitive decision-making through LLMs' training, which includes aspects such as learning. The trained LLMs can then be generalized to unseen problems". **Why do we use pre trained LLM as baseline?** **Page 11, right column, line 18-35:** we state "to assess the model's ability to make human-like decisions, we first split the data into train and validation sets to reserve a set of unseen problems. We then compared the predictive negative log-likelihood (NLL), a measure of goodness-of-fit, of Cognitive LLMs in predicting VSM-ACTR's decisions on the unseen problems, against a pre-trained LlaMa and a random guess model. A random guess model serves as the basic form of control condition to distinguish the effects of treatment from chance [30]. This approach allows us to assess the extent to which decisions are influenced by knowledge versus being purely stochastic. On

	the other hand, using LlaMa without fine-tuning as a baseline provides a reference point to measure the impact of knowledge transfer on the model's performance".
Methodologically, is there any advances in this paper, compared with existing work such as Trieu H. Trinh, Yuhuai Wu, Quoc V. Le, He He & Thang Luong (2023)?	**Page 5, line 51, to page 6, line 6:** we state "this present study builds upon previous research; however, we have adopted a different perspective by leveraging CAs to ground the decisions of LLMs in a data-driven manner. We aim to examine the properties of a neural network representation of the decision-making process in CAs and investigate whether knowledge from CAs can be preserved in an embedding space and infused into LLMs through the transfer of learning."
The authors need to cite highly relevant existing work, such as: • Integrating LLMs with Soar: arXiv:2310.06846v1 ; etc. • Integrating LLMs with Clarion: arXiv:2401.10444 ; arXiv:2410.20037 ; etc. • And other cognitive architectures; Etc.	**Explain why ACT-R and SOAR CAs are primarily discussed in this paper,** as stated on page 4, lines 2-4, and citing supporting literature**:** J.E., Laird, An Analysis and Comparison of ACT-R and Soar. (2021). *In Proceedings of the Ninth Annual Conference on Advances in Cognitive Systems*. **Page 5, right column, line 32-49**: cite suggested literature as "leveraging language models as external knowledge sources for cognitive systems, while exploring ways to improve the effectiveness of knowledge extraction [47]." and cite suggested CLARION literature as "Additionally, [89] proposes a direction for creating computational cognitive architectures using dual-process models and hybrid neuro-symbolic methods. Using the CLARION CA [88] as an example, the author illustrates the theoretical opportunities for incorporating LLMs into CLARION's modules of perception, memory, motor control, and communication, leveraging LLMs' natural language processing and generalization abilities."
Reviewer 2	**Revisions Made**
Page 7 line 8 or 9 "refer to VSM-ACTR below" -- it's not clear	**Pape 6**: "We created the VSM-ACTR cognitive model, which is a rule-based ACT-R cognitive decision-making model for DFM problems that implements multiple problem-solving strategies, through a combination of production rules."
It would help if the authors clarified the scope of their work earlier	**Page 1, abstract, line 31-33**: we state "this paper addresses this gap by asking whether Large Language Models (LLMs) can replicate cognition from Cognitive Architectures (CAs) to make human-like decisions. We introduce Cognitive LLMs, which are hybrid decision-making architectures comprised of a CA and an

	LLM, developed through a knowledge transfer mechanism called LLM-ACTR". Additionally, we refer to the statement of scope in the **introduction on page 2, left column, lines 45-51, and right column, lines 27-38**.
The paper ends with mostly text discussion of results that almost seems to hide some of the prediction accuracy results. I'd just try to get the point more clearly and overtly here	**Report the results:** **Page 12, right column, line 39 -51:** we state "we then report the comparison of the Cognitive LLMs with the baseline models on goodness of fit using negative log likelihood (NLL) and accuracy score for hold-out data. The Cognitive LLMs demonstrate significantly better performance across all metrics compared to the LlaMa-only model, highlighting its effectiveness in decision-making tasks involving cognitive reinforced learning. Additionally, the LlaMa-only model performs worse than the chance-level model. This underscores the necessity of fine-tuning pre-trained language models like LlaMa to adapt them to human-like decision-making patterns". **Discuss the limited improvement in preliminary experiments:** **Page 13, right column, line 1-14:** we state "however, the influence of the cognitive content vector is limited and warrants further investigation, partly because the stochastic simulation of the VSM-ACTR produces decision-making vectors of various lengths. This study addresses ragged tensors by padding, but this approach potentially dilutes or changes the semantics of each vector. To improve the impact of the cognitive vector, additional techniques such as vector optimization will be needed".
One big thing I'd like to see is more discussion of the overall approach's theoretical limitations,	**Page 14 limitations**, **right column, line 4 -51:** we discussed "One limitation also stems from the novelty of this study. How closely can we claim that cognitive model personas replicate human behavior on the same tasks? Currently, our focus is on tuning the model to align with general patterns of learning and error-making; however, VSM-ACTR still requires more granular human data for cognitive fine-tuning. The closer the VSM-ACTR model aligns with human behavior, the more accurately it can represent human decision-making processes. However, the more meaningful questions arise from considering the landscape of enabling machine cognitive reasoning. We must ask ourselves what we can learn about cognitive decision-making when we infuse knowledge from CAs into LLMs. For now, our insights are limited to the observation that knowledge from cognitive models can be preserved in an embedding space and

	could be learned by LLMs, and that embeddings from large language models can be trained to predict human-like decisions. While this is interesting in its own right, it certainly is not the end of the story. Looking beyond the current work, transitioning from transferring cognitive models' human-like decisions to LLMs, to guiding perception, memory, goal-setting, and actions, will provide the opportunity to apply a wide range of explainability techniques to LLMs' cognitive decision-making".
Other Revisions	Provide a detailed explanation of why the specific Cognitive Architecture, ACT-R, was chosen for this research, as discussed on **page 5, left column, lines 27-34, and page 6, right column, lines 12-18.** Format the appendix. Delete model traces in the appendix and instead, include representative snippet traces as shown in Table 1 to illustrate the knowledge representation of the ACT-R model. Conduct a grammar and spelling check.