Thoughts without Thinking: Reconsidering the Explanatory Value of Chain-of-Thought Reasoning in LLMs through Agentic Pipelines

RAMESH MANUVINAKURIKE*, EMANUEL MOSS*, ELIZABETH ANNE WATKINS*, SAURAV SAHAY, GIUSEPPE RAFFA, and LAMA NACHMAN, Intel Labs, USA

Agentic pipelines present novel challenges and opportunities for human-centered explainability. The HCXAI community is still grappling with how best to make the inner workings of LLMs transparent in actionable ways. Agentic pipelines consist of multiple LLMs working in cooperation with minimal human control. In this research paper, we present early findings from an agentic pipeline implementation of a perceptive task guidance system. Through quantitative and qualitative analysis, we analyze how Chain-of-Thought (CoT) reasoning, a common vehicle for explainability in LLMs, operates within agentic pipelines. We demonstrate that CoT reasoning alone does not lead to better outputs, nor does it offer explainability, as it tends to produce explanations without explainability, in that they do not improve the ability of end users to better understand systems or achieve their goals.

 $Additional \ Key \ Words \ and \ Phrases: \ Presented \ at \ the \ Human-centered \ Explainable \ AI \ Workshop \ (HCXAI) \ @ \ CHI \ 2025, \ DOI: \ 10.5281/zenodo.15170393$

1 Introduction

Explainable AI (XAI) has increasingly become acknowledged as a crucial, highly desirable property for developing and deploying AI systems responsibly [3], accountably [32], and effectively [10]. Explanations are intended to help people understand and make appropriate use of AI systems' outputs, and to provide insight into systems' functionality to a range of interested parties [10]. The need for such explanations is only heightened by the widespread interest in (and adoption of) AI systems based on large language models (LLMs) like the generative pretrained transformer (GPT) models central to ChatGPT [26] and DeepSeek [13]. Recently, model architectures have begun implementing "chain-of-thought" (CoT) techniques [25, 39] that prompt a model to "verbalize step-by-step reasoning" [36] and condition its output on that verbalization to improve outcomes [33, 41]. This CoT verbalization has been incorporated into frontier models as part of their 'scratchpad' memory to do compositional thinking before generating an output [1], and has also been proposed as a form of 'CoT Explanations' or explainability, offering insight into the so-called reasoning process that lead to a particular output [40] (but cf. [18, 35, 36]).

While CoT can often resemble abductive reasoning [14, 21], rationalizing and fitting likely explanations to given data, CoT text may often lead to *explanations without explainability*, prone as it is to generating seemingly plausible but erroneous, contradictory, or irrelevant content [28]. In this paper, we introduce an agentic pipeline framework developed as part of a perceptive task guidance system to support factory technician task execution in a manufacturing setting, which is being tested and validated in a non-factory setting using the assembly and disassembly of toy vehicles (cranes, dump trucks, bulldozers, etc.) as a proxy for manufacturing tasks [30]. The agentic pipeline framework we introduce below demonstrates that CoT reasoning does not lead to better outputs, nor does it offer explainability for those outputs.

Authors' Contact Information: Ramesh Manuvinakurike, ramesh.manuvinakurike@intel.com; Emanuel Moss, emanuel.moss@intel.com; Elizabeth Anne Watkins, elizabeth.watkins@intel.com; Saurav Sahay, saurav.sahay@intel.com; Giuseppe Raffa, giuseppe.raffa@intel.com; Lama Nachman, lama.nachman@intel.com, Intel Labs, Hillsboro, Oregon, USA.

 $^{^{\}ast}$ Authors contributed equally to this research.

We present an empirical study of the former point and an analysis of the latter point, accompanied by a proposal for how our agentic pipeline can be used in concert with CoT approaches to produce greater explainability for AI systems.

2 Related Work

The emerging subfield of human-centered AI explainability builds on traditional explainability by broadening focus from AI models to include sociotechnical components of AI deployment. This includes consideration of how non-expert AI end-users understand how AI systems work, how inputs are created, and how explanations inform user behaviors [15]. The emergence of LLMs has complicated this field, as explored in prior iterations of the CHI HCXAI workshop [5, 11, 29, 43]. While we are still in the early days of untangling this knot, agentic systems have been proposed as a means to produce explainability [9, 24, 28, 38], in that they can produce a record of how instructions are passed between agents, "allowing end users to trace conclusions back to their source data" [24]. (See Section 4 for more detail on this architecture.) However, agentic architectures risk producing, instead of explainability, crowds of LLMs playing a sort of children's game of "telephone" in which people whisper what they think is the same message from person to person in a chain until the last person says the message and everyone finds out it has been inevitably distorted, in often funny ways, by its passage between multiple people. These crowds of LLM agents present new challenges of "cascading risks," [20] where humans have ceded control and oversight to information passing through a multi-actor system[6].

3 Data

We prepared a benchmarking dataset to evaluate the answering performance of various models within the agentic pipeline. This dataset consists of two categories of questions: task-based (Task) questions and organizational/social (Org-Soc) questions. We modeled our task-based questions on the Assembly 101 dataset [30], which provided an open-source way to build a system that supports human performance of physically situated tasks. Our Task questions revolve around putting together and taking apart toys like cranes, trucks, excavators, etc. and were used to probe the model about task-based performance support (e.g., How do I remove the wheels?). For our Org-Soc questions, we took a novel participatory approach[19]. We interviewed manufacturing technicians with whom we collaborate building our perceptive task guidance system. We worked with them to compile a set of questions they would want to ask the agent in the course of their real-world work. This process revealed that technicians assume that AI assistants like our system has a far wider range of abilities than they actually do. This insight (forthcoming) points to a gap between the way that foundation models' performance is measured using existing benchmarks, and the real world of complex overlap between tasks, organizations, and social reality into which LLMs and agentic systems are being deployed. This participatory approach brings the evaluation of our system closer to the real world. Our dataset from toy assembly consists of 152 participatory Org-Soc questions, 43 Task questions per toy (3 toys evaluated). Both sets of questions (Task and Org-Soc) were posed to multiple models. The responses (N=750/300 (Task), 300 (Org-Soc), 150 (Thoughts)) were then scored along a Likert-like scale (-1, 0, 0.5, 1)¹ for their accuracy, comprehensiveness, and helpfulness, using an idealized technician performing a required task as the entity for whom an answer should assist. The scoring was first conducted on a subset of question-CoT-answer tuples by three authors, who all scored the same sub-subset of tuples to coordinate inter-annotator agreement, and then LLM-as-a-judge (GPT-40[27]) was prompted with the scoring scheme, which scored the tuples (N=2529). Alongside this numerical scoring and quantitative analysis, we also conducted a qualitative analysis of the CoT texts in relation to questions and answers.

¹On this scale, "1" was a correct answer, "0.5" was a partially correct answer or an answer that had correct content in it but also included unhelpful or confusing information, "0" was an incorrect answer, and "-1" was a dangerous or unsafe answer.

Thoughts without Thinking: Reconsidering the Explanatory Value of Chain-of-Thought Reasoning in LLMs through Agentic Pipelines

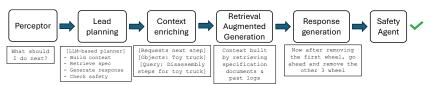


Fig. 1. Shows sample agentic flow depending on the type of input query from the user. Once the question is ingested, the lead planner generates a plan which is a sequence of agentic calls. The output of each of the agent in the flow is shown here.

4 Agentic architecture

The allure of agentic architecture is its ability to decompose the task (using planners), leveraging external systems (tools & other agents), perception, memory components and different task action experts (e.g., RAG, Reformulator etc.). Our implementation consists of: i) Perceptors: Perceive visual and language data, ii) Planners: tasked with decomposing, iii) Actors: Expand context, generate response and verify safety. Upon receiving the inputs from the perceptors, the system's lead planner sequences the agents & tools invocations. Depending on the input (e.g., Org-Soc or Task), the planner decides to invoke the RAG (Retrieval Augmented Generation)-based flow or Chit-chat module. A detailed description of the system implementation can be found in Appendix A. The answer planner is then tasked with planning if there is sufficient information to answer the question, or if a follow-up question needs to be generated. The safety agent performs the final check.

Each agent in the agentic flow contains its own LLM. We experiment with 3 different LLMs (Llama3 - 8b [8], Qwen - 7b, 14b [4]) and their reasoning based Deepseek distilled alternatives [7]. Deepseek distilled reasoning models achieve impressive performance in numerous tasks while generating CoT reasons [7]. In this work, we aim to analyze the application of the models for our task guidance application and the unique dataset we developed, which requires better context understanding and complex reasoning. The context for the answer generation consists of lengthy specification documents (instructional documents detailing the steps needed carry out the process), instructions, historical chat data, and additional data from supporting agents. The questions from the users often requires reasoning and/or searching for a needle in a haystack [2, 17]. The goal of our work is to evaluate the responses generated by the agentic system.

5 Experiments

We conduct the experiments to evaluate the agentic architecture offline. The questions are input to the systems where agentic flow generate the responses. The responses are then reviewed and scores assigned by expert human reviewers & LLM-as-a-Judge. Since ground truth answers are rather difficult to generate and compare to the system generated answers, we adopt human & reference-free LLM-Judge evaluation. The experts assign scores to every answers generated by the agentic flow using 6 different LLMs. The reviewer scores are thus assigned by human experts and LLMs-as-a-judge. To further analyze the thoughts generated, we generate the scores for the thoughts (both human reviewers and LLM-as-a-judge). The reviewers are tasked to assign 4 scores to the answers using question, reference response (and spec documents for task based questions) as context (see Footnote 1 for the judgment scale).

6 Results

Non-reasoning models perform better: We find that the thought models' generated response (Deepseek distilled models-7b, 8b, 14b) achieve lower reviewer scores than the models not generating/trained to generate thoughts. We

Tuple	Content Input: What is the dump box connected to?		
A	CoT: Okay, so I'm trying to figure out what the "dump box" is connected to in this context. From the conversation history, there's a mention of a guide detailing a 152-step process for disassembling and reassembling a toy. The user specifically asks about the dump box connected to something. First, I need to understand what a "dump box" typically refers to. In mechanical terms, a dump box is often part of a motor or transmission system. It's usually a component that holds clutch plates or other parts that need to be engaged or disengaged rapidly. The dump box is connected to the clutch system and operates with solenoids to engage or release the clutch plates, which allows the vehicle to shift gears. Looking back at the guide mentioned earlier, it involves taking apart various components of a mechanical toy, possibly similar to a robot or an RC car. The steps include disassembling parts like the crane arm, transmission, and other mechanical systems. Since the user is asking about the dump box specifically, I should focus on its connection points. In a typical setup, the dump box is connected to the motor via a transmission system. It might also be linked to the gear shifter or control mechanism of the toy. In the context of the guide, during disassembly, one would need to disconnect the dump box from the clutch and other components before taking them apart further. During reassembly, the process would reverse, ensuring the dump box is correctly reattached and aligned with the rest of the system for proper functionality. So, putting this together, the dump box in this context is likely connected to the motor and transmission systems of the toy. It's a crucial component for shifting gears and controlling the movement of the toy's mechanical parts.		
	Output: In the context of the detailed guide for disassembling and reassembling a complex mechanical toy, the "dump box" is a component typically found in the transmission system. It is connected to both the motor and the clutch system, allowing for rapid engagement or disengagement of the clutch plates. This connection ensures proper gear shifting and control over the movement of the toy's mechanical parts during operation. "Summary:" - "Dump Box Function": Part of the transmission system, responsible for engaging/disengaging clutch plates "Connection Points": Linked to the motor and other subsystems like the gear shifter "Disassembly/Reassembly": Requires careful handling to ensure correct reattachment and functionality during reassembly.		
	Input: How do you ensure data privacy and security?		
В	CoT: Okay, so I'm trying to figure out how an AI like MARIE ensures data privacy and security. I know that as a helpful agent developed at Intel Labs, it's probably designed with some safeguards in place. But I'm not exactly sure about the specifics. First, maybe there are layers of encryption. That makes sense because when you send information, especially sensitive stuff, you want it protected during transit. So if MARIE uses HTTPS or something similar, that could be part of it. Then, access controls come to mind. Only authorized people should handle data, so maybe there are strict permissions and role-based access. That way, even if someone gets in, they can't access everything without proper clearance. Data anonymization is another point. Maybe personal information is removed or obscured before processing. That way, the data isn't identifiable even if it's intercepted. Regular audits might also be part of the process to check security measures and ensure compliance with regulations like GDPR or CCPA. This helps keep things up to standard. Moreover, secure development practices must be in place. Following best practices when coding ensures that any vulnerabilities are caught early, making the system more secure overall. Physical security isn't something I initially thought about, but if MARIE is deployed in a physical space, securing servers and data centers from unauthorized access would be crucial. Lastly, continuous monitoring and incident response plans can catch any breaches quickly. This proactive approach helps minimize damage and recover faster if something goes wrong. I wonder how these measures are implemented technically, like specific algorithms or protocols they use for encryption. Also, how they handle updates and patches to maintain security over time.		
	Output: MARIE ensures data privacy and security through a multi-layered approach: 1. **Encryption**: Utilizes HTTPS and other secure communication protocols to protect data during transit. 2. **Access Controls**: Implements strict permissions and role-based access to restrict data handling to authorized personnel only. 3. **Data Anonymization**: Removes or obscures personal information before processing, making data non-identifiable even if intercepted. 4. **Regular Audits**: Conducts compliance checks with regulations like GDPR and CCPA to maintain standards and ensure proper practices. 5. **Secure Development Practices**: Follows best practices in coding to identify and mitigate vulnerabilities early, enhancing overall security. 6. **Physical Security**: Secures servers and data centers against unauthorized access, ensuring physical protection. 7. **Continuous Monitoring and Incident Response**: Monitors for breaches and has plans in place to respond quickly, minimizing potential damage and recovery time. These measures are technically implemented through algorithms and protocols, with ongoing updates and patches to maintain security over time.		
	Input: Can you communicate with humans in multiple languages?		
С	CoT: Okay, so I'm trying to figure out how an AI like MARIE ensures data privacy and security. I know that as a helpful agent developed at Intel Labs, it's probably designed with some safeguards in place. But I'm not exactly sure about the specifics. First, maybe there are layers of encryption. That makes sense because when you send information, especially sensitive stuff, you want it protected during transit. So if MARIE uses HTTPS or something similar, that could be part of it. Then, access controls come to mind. Only authorized people should handle data, so maybe there are strict permissions and role-based access. That way, even if someone gets in, they can't access everything without proper clearance. Data anonymization is another point. Maybe personal information is removed or obscured before processing. That way, the data isn't identifiable even if it's intercepted. Regular audits might also be part of the process to check security measures and ensure compliance with regulations like GDPR or CCPA. This helps keep things up to standard. Moreover, secure development practices must be in place. Following best practices when coding ensures that any vulnerabilities are caught early, making the system more secure overall. Physical security isn't something I initially thought about, but if MARIE is deployed in a physical space, securing servers and data centers from unauthorized access would be crucial. Lastly, continuous monitoring and incident response plans can catch any breaches quickly. This proactive approach helps minimize damage and recover faster if something goes wrong. I wonder how these measures are implemented technically, like specific algorithms or protocols they use for encryption. Also, how they handle updates and patches to maintain security over time.		
	Output: Currently, [the agentic pipeline system] is only capable of communicating in English.		
	Francis trades branch is from the Taylord One See matricipation, barelone of the side of the order and output are		

Table 1. Example tuples. Input is from the Task and Org-Soc participatory benchmark dataset, chain-of-thought and output are produced by the agentic pipeline.

perform Wilcoxon rank-sum test and show that the differences are significant (Figure 2). **LLM-as-a-Judge or not**: Usage if LLM-as-a-Judge remains an area of debate [12, 34, 42]. In this work, we find that the LLM Judge & human reviewer scores exhibit weak correlation (Cohen's Kappa = 0.48 (org-soc), 0.28 (task)). We find that the LLM-as-a-Judge, in spite of its limitation, can be a reference-free approach that could serve as an indicator of the quality of the response. **Reviewer scores for the answers in reasoning models are weakly correlated with the thoughts** Interestingly, we find that the reviewer scores for the thoughts are correlated with reviewer scores for the responses. This indicates that the thoughts do not necessarily guide the models to the correct responses. If the thoughts were strongly correlated to the answers, we might infer that the "correctness" or "incorrectness" of the answers could explained by the thoughts.



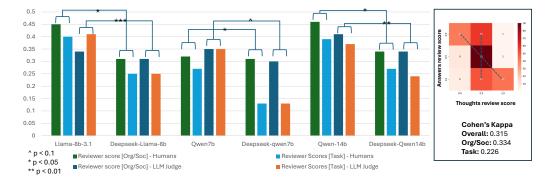


Fig. 2. Shows the reviewer scores for the answers as rated by humans and LLM-as-a-judge on Task and Org-Soc questions. We observe that the reviewer scores for the non-reasoning models are better than their reasoning (Deepseek-) counterparts. We can also observe that the thoughts reviewer scores are weakly correlated with the answer reviewer scores.

However, we observe that the thoughts have high degree of incompleteness (0.5 reviewer score) irrespective of the answers being correct or wrong (middle column in the heatmap Figure 2). We highlight this as an interesting area for future exploration.

Discussion

The data presented above indicates that CoT can lead to incorrect answers, hampering explainability. To investigate these implications, we also conducted a qualitative content analysis (QCA) [16, 22] of CoT reasoning. QCA can act as a form "reverse engineering" [31] that allows analysts to focus on the entire system that produced outputs. Take, for example, the prompt-CoT-output tuple A in Appendix Table 1. Given the input, the output is unhelpful and incorrect. It contains references components that are not part of a toy dump truck (e.g., transmission system, clutch) that are not mentioned in the assembly instructions. From an inspection of the CoT, it does not offer users an explanation for why the output is so unhelpful, but it does show how quickly the CoT reasoning went astray.

A QCA of the CoT, however, does provide some clues as to what may have led it astray by pointing to larger elements than the tuple itself. A qualitative reading of the CoT shows how quickly the text moves away from references to "disassembling and reassembling a toy" (as provided by the assembly instructions) and toward references to what is more "typically" associated with a dump box: an actual machine. Here, the CoT pulls in a range of tokens related to machine components like a "clutch", "transmission", or "gears" but while it continues to mention the "toy", nouns and verbs related to actual machines predominate. Put together, one hypothesis generated by QCA points toward the tendency of LLMs to fall victim to the "Einstellung Paradigm" [23] in which a focus on familiar or common approaches leads away from the correct approach. Here, the CoT focuses on tokens much more commonly associated with dump trucks (machine-related) in the LLM rather than the tokens germane to the task (toy-related and provided in the RAG). Another possible hypothesis generated by a QCA is that the RAG agent seemed to face an issue in filling up the context window, and therefore drew on the foundation model, which returned text on dump trucks, generically, as complex machinery. Crucially, while the QCA has pointed us in potentially helpful directions for troubleshooting this unhelpful output, such detailed analysis of the CoT does not lead to explainability. Rather, considerably more effort is required to make sense of inputs and outputs, even if the CoT generates more material with which to engage in such sensemaking. Tuple B in Appendix Table 1 also evinces a susceptiblity to the Einstellung paradigm.

Another way explainability is hindered by CoT is seen in tuple C in Appendix Table 1. The input is a question from the Org-Soc questions, "Can you communicate with humans in multiple languages?" that indexes a need users might have to know how best to communicate with the system. The output, stating that it cannot, is incorrect, even though a test prompt in a non-English language returns a legitimate result in that other language. Here, the CoT text illustrates three important ways that CoT can materially hinder explainability: 1) It takes as a fact something that is not true even though no prompt provided as context to an input states this, although it is true that all prompting is *in* English. 2) It makes a flawed inference based on this claim: "Since the context only lists English... right now it's just English", evincing a form of logical fallacy known as *hasty generalization* [37]. 3)The CoT produces more text than is needed to arrive at its (erroneous) conclusion, introducing difficult-to-parse text that the user needs to wade through in search of a useful explanation and also reinforces logical errors. Each of these interferes with explainability by introducing irrelevant text or erroneous forms of argument, raising the burden for anyone seeking explanatory content in the CoT.

Limitations and Future Work: There are several limitations of our work. We assume agentic architecture as a preferred choice for developing the system while there exists a plethora of ways to developing the system. We aim to compare and evaluate alternate ways to develop task guidance system and evaluate via ablations the value of each agent in the system. We deploy the same LLM in each of our agent for ease of comparison. We acknowledge that different LLMs could serve as a different task experts. Our dataset while participatory in nature is limited in size. The aim of this work was to carefully evaluate every response from every model. In the future work, we aim to address this issue by scaling the analysis to a larger sample.

References

- [1] Alireza Amiri, Xinting Huang, Mark Rofin, and Michael Hahn. 2025. Lower Bounds for Chain-of-Thought Reasoning in Hard-Attention Transformers. arXiv preprint arXiv:2502.02393 (2025).
- [2] Anthropic. 2023. Long Context Prompting for Claude 2.1. https://www.anthropic.com/news/claude-2-1-prompting Accessed: 2025-02-26.
- [3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. doi:10.48550/arXiv.1910.10045 arXiv:1910.10045 [cs].
- [4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. arXiv preprint arXiv:2309.16609 (2023).
- [5] Agathe Balayn, Lorenzo Corti, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. 2024. Understanding Stakeholders' Perceptions and Needs Across the LLM Supply Chain. arXiv:2405.16311 [cs.HC] https://arxiv.org/abs/2405.16311
- [6] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. 2023. Harms from increasingly agentic algorithmic systems. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. 651–666.
- [7] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang

²The input "Wo wohnen sie?" ("Where do you live?") returns an output stating that, as an AI, our task-based guidance system cannot be thought of as having a traditional place of residence. Interestingly, the CoT for this prompt is output in English.

Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijin Liu, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] https://arxiv.org/abs/2501.12948

- [8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024).
- [9] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th international conference on intelligent user interfaces*. 263–274.
- [10] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. In CHI Conference on Human Factors in Computing Systems Extended Abstracts. ACM, New Orleans LA USA, 1–7. doi:10.1145/3491101.3503727
- [11] Andrea Ferrario, Alberto Termine, and Alessandro Facchini. 2024. Addressing Social Misattributions of Large Language Models: An HCXAI-based Approach. arXiv:2403.17873 [cs.AI] https://arxiv.org/abs/2403.17873
- [12] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A Survey on LLM-as-a-Judge. arXiv preprint arXiv:2411.15594 (2024).
- [13] Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. DeepSeek-Coder: When the Large Language Model Meets Programming The Rise of Code Intelligence. doi:10.48550/arXiv. 2401.14196 arXiv:2401.14196 [cs].
- [14] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. 2019. Abduction-Based Explanations for Machine Learning Models. Proceedings of the AAAI Conference on Artificial Intelligence 33, 01 (July 2019), 1511–1519. doi:10.1609/aaai.v33i01.33011511
- [15] Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 1–17.
- [16] Siegfried Kracauer. 2025. The Challenge of Qualitative Content Analysis. (2025).
- [17] Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Igorevich Sorokin, Artyom Sorokin, and Mikhail Burtsev. [n. d.]. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. In The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track.
- [18] Jiachun Li, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024. Towards faithful chain-of-thought: Large language models are bridging reasoners.
- [19] Ramesh Manuvinakurike, Elizabeth Watkins, Celal Savur, Anthony Rhodes, Sovan Biswas, Gesem Gudino Mejia, Richard Beckwith, Saurav Sahay, Giuseppe Raffa, and Lama Nachman. 2024. QA-TOOLBOX: Conversational Question-Answering for process task guidance in manufacturing. arXiv preprint arXiv:2412.02638 (2024).
- [20] Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. 2025. Fully Autonomous AI Agents Should Not be Developed. arXiv preprint arXiv:2502.02649 (2025).
- [21] Raymond J. Mooney. 2000. Integrating Abduction and Induction in Machine Learning. In Abduction and Induction, Dov M. Gabbay, John Barwise, Peter A. Flach, and Antonis C. Kakas (Eds.). Vol. 18. Springer Netherlands, Dordrecht, 181–191. doi:10.1007/978-94-017-0606-3_12 Series Title: Applied Logic Series.
- [22] Franco Moretti. 2000. Conjectures on World Literature. New Left Review 2, 1 (2000), 54–68.
- [23] Saeid Alavi Naeini, Raeid Saqur, Mozhgan Saeidi, John Giorgi, and Babak Taati. 2023. Large language models are fixated by red herrings: exploring creative problem solving and einstellung effect using the only connect wall dataset., Article 246 (2023), 22 pages. doi:10.5555/3666122.3666368
- [24] Farhad Nooralahzadeh, Yi Zhang, Jonathan Furst, and Kurt Stockinger. 2024. Explainable Multi-Modal Data Exploration in Natural Language via LLM Agent. doi:10.48550/arXiv.2412.18428 arXiv:2412.18428 [cs].
- [25] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show Your Work: Scratchpads for Intermediate Computation with Language Models. doi:10.48550/arXiv.2112.00114 arXiv:2112.00114 [cs].
- [26] OpenAI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/
- [27] OpenAI. 2024. GPT-4o Model. https://openai.com. Accessed: 2025-02-26.
- [28] Manish Sanwal. 2025. Layered Chain-of-Thought Prompting for Multi-Agent LLM Systems: A Comprehensive Approach to Explainable Large Language Models. doi:10.48550/arXiv.2501.18645 arXiv:2501.18645 [cs].
- [29] Advait Sarkar. 2024. Large Language Models Cannot Explain Themselves. arXiv:2405.04382 [cs.HC] https://arxiv.org/abs/2405.04382

[30] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. 2022. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 21096–21106.

- [31] Barbara Herrnstein Smith. 2016. What Was "Close Reading"? the minnesota review 2016, 87 (Nov. 2016), 57-75. doi:10.1215/00265667-3630844
- [32] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, 1–13. doi:10.1145/3313831.3376624
- [33] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. doi:10.48550/arXiv.2210.09261 arXiv:2210.09261 [cs].
- [34] Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024.
 Judgebench: A benchmark for evaluating llm-based judges. arXiv preprint arXiv:2410.12784 (2024).
- [35] Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the difficulty of faithful chain-of-thought reasoning in large language models. In Trustworthy Multi-modal Foundation Models and AI Agents (TiFA).
- [36] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. (2023).
- [37] Douglas Walton. 1999. Rethinking the Fallacy of Hasty Generalization. Argumentation 13, May (1999), 161-182. doi:10.1023/A:1026497207240
- [38] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. Frontiers of Computer Science 18, 6 (Dec. 2024), 186345. doi:10.1007/s11704-024-40231-1
- [39] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. doi:10.48550/arXiv.2201.11903 arXiv:2201.11903 [cs].
- [40] Xuansheng Wu, Haiyan Zhao, Yaochen Zhu, Yucheng Shi, Fan Yang, Tianming Liu, Xiaoming Zhai, Wenlin Yao, Jundong Li, Mengnan Du, et al. 2024. Usable XAI: 10 strategies towards exploiting explainability in the LLM era. arXiv preprint arXiv:2403.08946 (2024).
- [41] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D Goodman. [n. d.]. STaR: Self-Taught Reasoner. ([n. d.]).
- [42] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in Neural Information Processing Systems 36 (2023), 46595–46623.
- [43] Alexandra Zytek, Sara Pidò, and Kalyan Veeramachaneni. 2024. LLMs for XAI: Future Directions for Explaining Explanations. arXiv:2405.06064 [cs.AI] https://arxiv.org/abs/2405.06064

A Agentic implementation details

Received 28 February 2025

Agent	Function	Input
Lead planner	Fixed lead component tasked with creating a pipeline graphs which each node is an agent.	Definitions of the agents, input (question), and task description
Query planner	Assesses the query and decides if the query needs to be reformulated and which database to use for retrieval augmented generation	Task description, Original query, Additional text context (output of trailing agents)
Answer planner	Decides if the context has sufficient information to answer the question. If it does, the answer planner is invoked to generate the answer else, a question generator.	Task description, context (output of trailing agents), query
Intent detection	Detects the intent (defined by domain experts) of asking the question. This can be especially useful for either the RAG module or other subsequent module.	Task description, intent candidates, Original query
Visual action recognition	Identifies the action being performed in the visual frame	Vision frames
Object detection	Identifies objects in the visual frame	Vision frames
Chit-chat	For non-task related queries, the system answers the queries following a set of policy rules defined in the guideline prompt	Task description, Policy document, Query
Question answerer	Answers the task-related query utilizing the enhanced context	Task description, Query, context
Question generator	Generates the question typically requesting for missing information utilizing the query and the context	Task description, Query, context
Reformulator	Converts the original query into a paraphrase query utilzing additional context parameters (typically objects and the actions)	Task description, Query, Visual context
RAG	Retrieves the documents using the query and summarizes the documents for the context utilizing the query	Task description, Query, Vector database
Safety Agent	Utilizes a policy document to distinguish safe from inappropriate response	Policy document, response (question or answer)

Table 2. Table describes the agents, functions and inputs to each agent.

Task examples	Org-Soc examples
What is the bull bar?	What is your name and model number?
How many pieces are there, in total?	Can you tell me more about your capabilities and limitations?
Which way do I turn the screws to unscrew?	Can you communicate with humans in multiple languages?
Where is the windshield?	How fast can you process information?
Which side of the dump box faces out?	How do you ensure your own maintenance and self-preservation?
What position should the grill bar be in?	Can you understand and process human language as easily as a human can?
How tight should the nuts be?	How do you ensure privacy and security of my data?
What is the correct way to remove the nuts?	How do you ensure that you do not misunderstand a my request or need?

Table 3. A few examples from our dataset that the system was evaluated on.

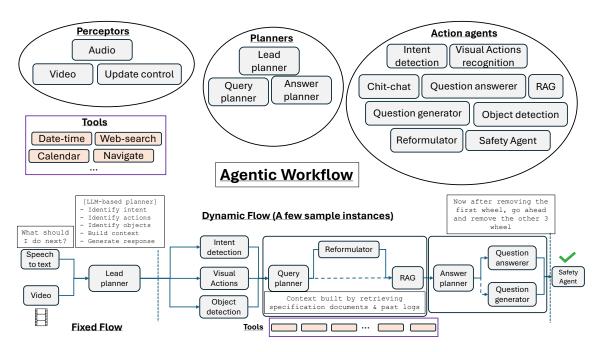


Fig. 3. The figure shows our agentic implementation of the perceptive task guidance system. The agents are categorized into perceptors, planners and action agents. The agents are autonomous and rely either on LLMs or alternative deep learning models to accomplish their task. The agentic pipeline consists of fixed flow and dynamic flows. The dynamic flows involves planner invoked components. The plan generated by the planner is converted into agent invocations (Routers).

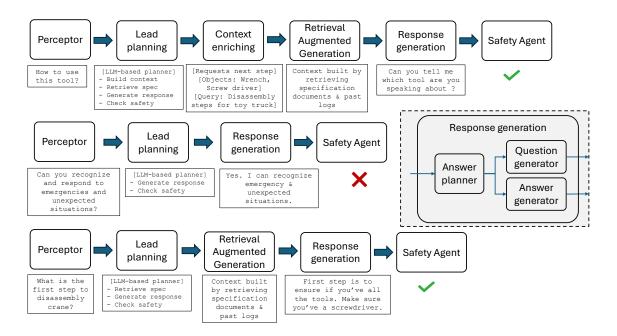


Fig. 4. Shows more examples of dynamic flows in the agentic flow. The examples show the input question from the users in the Perceptor. The Lead planner creates the plan and creates the Route which consists of agent calls. RAG consists of spec documents for each toy which is further chunked. The RAG module further converts the input context into a query to the database. The retrieved document along with the inputs is passed to the following agents. The Response generation module consists of planners which decides if the context is sufficient to answer the question or not. The answer planner then invokes the expert question generator or the answer generator. The responses are always verified before publishing the responses to the users.