Advances in Computer, Signals and Systems (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2025.090315
ISSN 2371-8838 Vol. 9 Num. 3

Comparison of Explainability and Scalability of Causal AI and LLMs in AGI

Zicheng Huang

College of of Computer Science and Software Engineering, Guangxi Normal University, Guilin, 541006, Guangxi, China

Keywords: Artificial General Intelligence, Artificial Intelligence, Large Language Models, Machine Learning

Abstract: Causal AI and Large Scale Language Modeling (LLM) are the two main directions of current AI research, focusing on causal reasoning and natural language processing, respectively. This article attempts to answer a key question: Which is more promising on the path towards safe general artificial intelligence (AGI), casual AI or LLM? Research shows that although both have their own advantages, relying solely on either path has significant limitations. Therefore, this article proposes a fusion path that combines the causal inference ability of Causal AI with the language understanding and task execution advantages of LLM, which may provide a more feasible solution for the implementation of AGI. Ultimately, the comprehensive method proposed in this article may bring new insights and directions for the development of artificial general intelligence.

1. Introduction

1.1 The Concept and Value of AGI

Artificial General Intelligence (AGI) is not merely “more powerful narrow AI.” It must autonomously abstract rules and transfer them to tasks it has never seen—this “cross-domain flexibility” is the core hallmark of human-level cognition [2]. Once realized, its impact will erupt outside the lab first: AGI in hospitals can integrate imaging, genomics and medical history in real time to deliver personalized diagnosis and treatment [1]; at climate-negotiation tables it can run high-fidelity simulations of the “carbon-price → industry → employment” chain at once, pressing the “policy-preview” button [8][9]. Of course, the flip side is immediately visible: when the system begins to rewrite itself, how should responsibility be assigned? If its decisions eliminate certain jobs, how can society redistribute value? These questions [3] remind us that AGI governance must iterate in lock-step with algorithms, not as an after-the-fact patch.

1.2 Comparison of Causal AI and LLM in the Realization of AGI

"Who is closer to AGI?" Before answering this question, let's first take a look at an experimental curve: when the questions were quietly switched to news events after 2024, the accuracy of GPT-4o in causal question answering dropped from 99.1% to 69.2%, a decline of nearly 30 percentage points

[11]. This gap suggests that the "understanding" of large models is more of a compressed playback of training corpora, rather than reverse engineering of causal mechanisms.

Recent reviews have used sharper wording to point out that LLMs hardly "explain away" spurious correlations in the classic collider structure; at best, they are "giant parrots that can recite causal phrases" [13][14]. In contrast, Causal AI explicitly writes the intervention relationships between variables into structural equations, thereby allowing humans to "turn knobs" to verify hypotheses outside the Turing Test [15].

However, language models remain an indispensable "universal interface." They can parse "Please place the red cylinder on the north side of the wooden box" into executable code in milliseconds. The danger lies in the fact that Wu et al. simply replaced the verb "pick" with "grab," causing the robot's success rate to plummet by 14.6%–22.2% [12]—a synonymous rewrite can shatter cross-modal alignment. If AGI cannot be immune to such "micro-seismic disturbances," its ticket to the operating table or the highway should rightfully be temporarily withheld.

Therefore, instead of choosing between Causal AI and LLMs, it would be better to regard the former as a "causal verification layer" and the latter as a "semantic interface layer": let SCM first issue a "certification of intervenability" for the action plan, and then hand it over to LLM to complete the final natural language interaction. The stitching of these two paths may be the shortest plank to a safe AGI— at least, more reliable than waiting for a single paradigm to revolutionize itself [4].

With the rise and progress of wireless sensor network research, how to locate and track moving targets using wireless sensor networks has become an important research topic of target tracking [5][6]. The nodes consisting of wireless sensor network has the characteristics of small volume, low price and low energy consumption. Nodes can transmit information through wireless communication and ad hoc network. These features make the moving target tracking system based on wireless sensor network have obvious advantages than traditional moving target tracking system [7][8]. Although the moving target tracking based on wireless sensor network has many advantages, the nature of the wireless sensor network is different from the ordinary network. The wireless sensor network itself has a large number of nodes, nodes are prone to failure, energy is constrained and transmission is unreliable. The detection ability, computing ability, storage ability and so on of nodes consisting of wireless sensor network are seriously limited. All of these put forward challenge for the design and implementation of moving target tracking system based on wireless sensor network.

2. Characteristics and Advantages of Causal AI

Causal AI is an artificial intelligence system based on causal relationships. It relies on understanding and modeling the causal relationships between events, thereby enabling prediction and intervention. Recent systematic reviews have pointed out that large models are merely "causal parrots," while structural causal models can truly open up new frontiers in causal inference [13]. In Table 1, the advantages of Causal AI are summarised. Table 1 reveals that Causal AI obtains its edge from explicit edges: every arc in the SCM corresponds to a testable intervention, allowing doctors or policy-makers to read off “$\beta = 0.3$” as a 30 % drop in readmission risk per 1 mmol/L glucose reduction without post-hoc visualisation [15].

Table 1: Advantages of Causal AI

Core Features	Advantages	Technical Advantages	Application Scenarios
High Precision	It performs reasoning and prediction by identifying and utilizing real causal relationships, rather than merely relying on	Compared to machine learning methods that only fit data, its predictions are based on causal mechanisms, resulting in higher accuracy.	In medical diagnosis, Causal AI uses do-calculus to quantify the counterfactual risk of "giving medication

	statistical patterns. It has an advantage in dealing with complex and uncertain problems.		before examination," rather than just reporting correlation strength [15].
Strong Explainability	It can model causal relationships, thereby explaining the reasons and basis for prediction results and assisting human decision-making.	The model output is transparent, unlike the "black box" decision-making of LLMs, which is crucial in decision support and risk assessment.	The weight of each edge represents the intervention effect. Doctors can directly read that "$\beta = 0.3$" means that for every 1 mmol/L decrease in blood sugar, the probability of readmission decreases by 30%, without the need for post hoc visualization [15].
Intervention Capability	It not only predicts but also has the ability to intervene, allowing it to simulate or guide actual actions to change the system's state and achieve optimization.	LLMs are usually limited to prediction or content generation, while Causal AI can answer intervention questions like "What should be done to achieve a certain goal?"	G ²Reasoner grafts external knowledge retrieval onto LLMs and achieves an additional 2.3–4.1 percentage points on the new CausalProbe-Hard scenario [11]. Meanwhile, Keshmirian's experiment shows that LLMs' judgments of causal strength are more extreme than humans', indirectly verifying the necessity of SCM posterior [18].
Broad Applicability	The framework based on causal relationships is universal, applicable to both natural and social sciences.	The modeling framework is decoupled from the specific domain's surface features, resulting in good transferability.	Medicine, finance, economics, political science, intelligent transportation
Strong Robustness	With explicit structural priors, it is insensitive to semantic perturbations in inputs and maintains the stability and intervenability of outputs under disturbances.	In safety-critical scenarios, LLM/VLM is susceptible to "semantic equivalence-syntactic perturbation," with task success rates potentially dropping by more than 20% [12], while Causal AI remains stable.	Safety-critical AGI scenarios such as robot control and autonomous driving

Overall, compared to parametric causal models, LLMs perform weakly and unstably in "explaining away" effects in collider structures [10]. In safety-critical scenarios such as robot control, LLM/VLM is easily triggered by "semantic equivalence-syntactic perturbation," resulting in a task success rate drop of over 20% [12]. Causal AI, with its explicit structural priors, maintains intervenability and explainability under similar disturbances, providing a more robust route to safe AGI.

3. Characteristics and Advantages of LLMs

In Table 2, the characteristics and advantages of large language models are summarised. Table 2 shows how LLMs compensate for data sparsity via latent representations—rare trigrams still receive plausible continuations—but also flags the Achilles heel: synonym substitution (pick → grab) can raise robot failure rates by 14.6 %–22.2 % [12], exposing the fragility of cross-modal alignment.

Table 2: Advantages of Large Language Models

Core Features	Advantages	Typical Application Scenarios
Handling Sparse Data	By introducing latent variables to reduce the number of parameters, it can more effectively handle words that appear infrequently in the training data, overcoming the issue of data sparsity.	Large-scale language modeling, handling rare words
Resolving Ambiguity	Latent variables can correspond to different semantic interpretations, enabling the model to understand the correct meaning of polysemous words based on context.	Word sense disambiguation, semantic understanding
Language Generation	Capable of generating new language sequences, rather than just predicting the next word.	Natural language generation, dialogue systems, text summarization, creative writing
Strong Explainability	Latent variables correspond to different semantic or grammatical concepts, allowing the model to be analyzed for interpretability to some extent and understand its decision-making basis.	Sentence classification, sentiment analysis, explainable AI
Flexible Modeling	By introducing different latent variables, the model architecture is highly flexible and can be customized and adjusted according to specific tasks and scenarios [7].	Adaptation to various natural language processing tasks

LLMs, through pretraining latent variables on large-scale corpora, are naturally adept at compensating for statistical voids caused by data sparsity: even when a trigram appears only once in the training set, the model can still provide reasonable continuation based on contextual embeddings. This characteristic makes it the preferred base model for tasks such as sentiment polarity judgment and multi-turn dialogue generation [5][7]. However, being "good at continuation" does not equate to being "good at reasoning." In the CausalProbe-2024 benchmark, when faced with causal chains of news events that emerged after 2024, the exact match rate of GPT-4o plummeted from 99.1% on the traditional COPA dataset to 69.2%, a drop of nearly 30 percentage points [11], indicating that it mainly relies on parameter memorization rather than structural causal inference in new scenarios. More seriously, this "memory-based" decision-making is extremely sensitive to input perturbations: when Wu et al. replaced "pick" with the synonym "grab" in robot instructions, the task failure rate immediately increased by 14.6%–22.2% [12], exposing the fragility of the cross-modal alignment layer. In other words, the "language prior" of LLMs is a powerful tool for open-domain generation, but it becomes a fatal weakness in safety-critical scenarios that require causal consistency and physical robustness.

4. Limitations and Challenges of Causal AI and LLMs

4.1 Limitations and Challenges of Causal AI include

Table 3: Limitations and Challenges of Causal AI

Challenges	Specific Manifestations and Difficulties	Potential Impacts/Risks
Data Requirements	Collecting sufficient "intervention-control" pairs is costly; complex systems also require experts to manually label causal edges.	Difficult to implement in data-scarce scenarios, significantly increasing model construction cycles and costs.
Causal Relationship Identification	Identifying the real causal relationships between variables is the core challenge. For complex systems, it is often difficult to automatically identify them and relies on the knowledge and manual intervention of domain experts.	If the causal graph is drawn incorrectly, all subsequent intervention conclusions may be invalid.
Feasibility of Intervention	Ethical or legal prohibitions against forcibly "doing" certain variables; the actual cost may also be too high.	The model suggests the "optimal intervention" but has nowhere to implement it, and the reasoning results become mere paper suggestions.
Perception-Physics Alignment Error (Safety Risk)	Causal variables are not strictly aligned with the robot coordinate system, and millimeter-level errors are magnified at the end.	Even if the SCM logic is correct, the robotic arm may still grab the wrong object, creating an immediate safety risk [12].
Robustness to Cognitive Bias	Human annotators "fail to explain" in structures such as colliders [16], and the data is then amplified by the model.	Learning biased causality leads to a simultaneous decline in robustness and fairness, which is difficult to detect after deployment.

Table 3 quantifies the “price of structure”: even a perfect causal graph fails when perception misaligns with physics—Wu et al. showed a 2 mm visual-language embedding bias can inflate grasping failure by >20 % [12], underscoring the need for joint perception-physics calibration.From the Table 3, we can see that a "correct" causal graph is merely an entry ticket. When SCM is inserted into a closed-loop control system, any misalignment between perception and physics can amplify minor errors into safety accidents—Wu et al. introduced a 2 mm visual-language embedding bias into a robotic arm, resulting in a more than 20% increase in grasping failure rates [12]. In other words, without a perception-physics joint calibration layer, even the most elegant causal graph cannot enable a robot to "grasp accurately." Moreover, humans tend to "weakly explain away" in descriptive collider tasks [16], and these cognitive flaws enter the model through annotated data, ultimately allowing biases to persist under the guise of "causal explainability." Therefore, the next generation of safe AGI must optimize the "causal graph + perception alignment + human bias correction" as a trio, rather than applying patches after the fact.

4.2 Limitations and Challenges of LLMs include

Table 4 summarises the triple penalty of “data hunger–cognitive bias–cross-domain fragility”; in

particular, Chi's 19 pp drop on post-2024 news [11] and Wu's synonym-induced robot failures [12] reveal that scaling parameters alone cannot guarantee robust generalisation.

Table 4: Limitations and Challenges of Large Language Models

Challenge Domain	Specific Manifestations and Difficulties	Empirical Evidence / Impact
Model Complexity and Poor Explainability	The introduction of latent variables makes the model structure complex and difficult to understand. The selection and adjustment of latent variables require in-depth domain knowledge and model design skills [7].	This turns the model into a "black box," making the decision-making process opaque and untrustworthy in high-stakes applications.
Huge Training Data Requirements	Requires large-scale, high-quality training data, especially for complex semantic tasks. The collection and annotation of data are costly and time-consuming.	This limits the participation of resource-poor institutions in research and application and may exacerbate biases in the data.
High Computational Costs	The increased model complexity leads to significant computational resources and time required for training and inference, with high energy consumption.	This results in high economic and environmental costs and makes deployment on resource-constrained edge devices difficult.
Limited Generalization Ability	Faces challenges in generalizing to new domains or tasks, often performing poorly on unseen data and requiring extensive transfer learning.	Unstable performance when adapting to dynamically changing world knowledge or domain requirements in real-world deployment.
Weak Deep Causal Reasoning	Performs poorly in novel, complex causal reasoning tasks, often memorizing surface patterns rather than grasping deep causal mechanisms.	In the new CausalProbe-2024 benchmark, the accuracy of LLaMA-2-7B-chat dropped significantly from 75.2% on the old COPA benchmark [11], indicating difficulty in handling new causal scenarios.
Amplification of Human Cognitive Biases	Not only fails to correct inherent human cognitive biases but may even amplify them. For example, in handling "collider" structures in causal graphs, LLMs show weaker and less stable biases than humans [14], and even amplify existing weak explanatory biases [17].	This can reinforce existing social biases or incorrect reasoning patterns in model outputs, posing risks to fairness and safety.

Despite the strengths of Causal AI and LLMs in explainability and language understanding, both still face the triple obstacles of "data hunger - cognitive bias - cross-domain fragility," which are far from being naturally healed over time. Taking causal reasoning as an example, Chi et al. constructed a new corpus using news events after 2024 and found that the accuracy of LLaMA-2-7B-chat on CausalProbe plummeted from 75.2% on COPA to 56.5%, a drop of nearly 19 percentage points [11]. This indicates that "unseen scenarios" can easily expose the model's reliance on statistical memory. More alarmingly, humans already tend to "weakly explain away" spurious correlations when facing collider structures [17], and LLMs not only fail to correct this cognitive bias but also amplify it further

[14]. In other words, if technological iteration only focuses on expanding parameters and corpora without incorporating structural causal constraints and psychological bias calibration, the limitations and challenges will simply re-emerge in a different form.

5. Prospects and Challenges

Achieving safe AGI is not the natural endpoint of "parameter stacking + waiting for computing power," but rather a protracted battle that requires repeatedly breaking down bottlenecks. The following sections illustrate why challenges are intensifying from five perspectives: "the essence of intelligence - learning paradigms - data barriers - modality fragility - ethical governance."

a) The "Black Box" of Human Intelligence Remains Unopened

● Current neuroscience lacks a unified model for the mechanisms of consciousness, emotion, and the emergence of creativity. If "intuition - emotion - social cognition" cannot be quantified into computable signals, AGI will remain at the level of "high-level automation" [6].

b) The "Generalization Ceiling" of Learning Paradigms

● Despite LLMs setting new records on specific tasks, their essence still relies on large-scale statistical fitting. When facing cross-domain long-tail events, model performance deteriorates rapidly [4][7]. More critically, algorithms lack a closed loop of "autonomously proposing hypotheses - intervening - verifying," making it difficult to perform causal discovery with minimal samples like children do.

c) The "Trilemma" of Data Acquisition

● It is challenging to achieve large-scale, high-quality, and privacy-compliant data simultaneously. High-value scenarios like healthcare and finance have low data openness. Synthetic data can alleviate the quantity gap but may introduce simulation biases into the model. Federated learning and differential privacy are still balancing on the seesaw of "usability - security" [6].

d) Modality Fragility: One Sentence Can Make a Robot "Strike"

● Wu et al. [12] found that simply replacing "pick" with "grab" in a real robotic arm task reduced success rates by 14.6%–22.2%. "Jailbreak" studies further revealed that indirect prompts can induce LLMs to output dangerous action codes [19]. Ahn et al. [20] demonstrated that aligning language priors to robot-affordable spaces is the first step to mitigating such misalignment. Without a "causal-physical consistency" verification layer, such perturbations can become amplifiers for security vulnerabilities.

e) Ethics and Governance: The Gap Between Technology and Regulation

● The legislative speed of algorithm transparency, responsibility attribution, and value alignment lags far behind the model iteration speed. RT-2 has proven that end-to-end fine-tuning of vision-language-action can directly translate web semantics into robot execution signals [21], but it also shortens the path from "online rumors → physical actions." If auditable causal constraints and red teaming are not introduced during training, AGI's "misfires" will no longer be confined to screens but will directly impact the real world [3][6].

In summary, the path to AGI requires surmounting five major challenges: "cognitive science - learning theory - data engineering - robust intervention - social governance." No single breakthrough is sufficient to declare victory. Only by integrating the depth of causal reasoning with the breadth of cross-modal robustness into a unified framework can AGI "think clearly, act reliably, afford to make mistakes, and correct quickly" in a complex world.

6. Conclusion

Causal AI delivers interpretable interventions but stumbles when perception mis-aligns with physics; LLMs give fluent interfaces yet falter on unseen causal chains and synonym jitters (14–22 %

drop[11][12]). The safer road to AGI is therefore a single pipeline: let structural causal models vet every action for physical consistency, then let large language models translate the verified plan into natural instructions. This causal-linguistic fusion offers both transparent do-calculus and robust cross-modal alignment in one framework, moving us from “talking about causality” to “acting on it” without sacrificing fluency or safety.

References

[1] D. Wu, H. Li, X. Chen, "Exploring the impact of general-purpose large AI models on education," Open Education Research, vol. 29, no. 2, pp. 19-25+45, 2023.

[2] J. Zhao, F. Wen, J. Huang, et al., "Toward general artificial intelligence for power systems with large language models: theory and applications," Automation of Electric Power Systems, pp. 1-16, 2024. [Online]. Available: http://kns.cnki.net/kcms/detail/32.1180.tp.20231123.1439.006.html.

[3] P. Wang, "From control to guidance: intuition and governance paths of general artificial intelligence," Oriental Law, pp. 1-11, 2024. [Online]. Available: https://doi.org/10.19404/j.cnki.dffx.20231116.005.

[4] J. Shi, J. Liu, "Optimization and innovation of public-library services based on general artificial intelligence," Library Development, pp. 1-11, 2024. [Online]. Available: http://kns.cnki.net/kcms/detail/23.1331.G2.20231031.1435.005.html.

[5] Z. Zhang, T. Liu, "ChatGPT technology analysis and prospects for general artificial-intelligence development," Bulletin of National Natural Science Foundation of China, vol. 37, no. 5, pp. 751-757, 2023. DOI:10.16262/j.cnki.1000-8217.20231026.003.

[6] K. Zou, Z. Liu, "Governance of ChatGPT-like general artificial intelligence from the perspective of algorithmic-security review," Journal of Hohai University (Philosophy and Social Sciences), vol. 25, no. 6, pp. 46-59, 2023.

[7] Y. Xiao, "Generative language models and general artificial intelligence: connotation, path and implications," People's Tribune Academic Frontier, no. 14, pp. 49-57, 2023. DOI:10.16619/j.cnki.rmltxsqy.2023.14.004.

[8] N. Yu, "The impact of new-generation general artificial intelligence on international relations," International Studies, no. 4, pp. 79-96+137, 2023.

[9] T. Zhu, "General artificial intelligence in psychology: an application analysis," People's Tribune Academic Frontier, no. 14, pp. 86-91+101, 2023. DOI:10.16619/j.cnki.rmltxsqy.2023.14.008.

[10] H. M. Dettki, B. M. Lake, C. M. Wu, et al., "Do large language models reason causally like us? Even better?" in Proc. Annual Meeting of the Cognitive Science Society, 2025, arX:2502.10215.

[11] H. Chi, H. Li, W. Yang, et al., "Unveiling causal reasoning in large language models: reality or mirage?" in Thirty-Eighth Conf. Neural Information Processing Systems, 2024, arXiv:2506.21215.

[12] X. Wu, S. Chakraborty, R. Xian, et al., "On the vulnerability of LLM/VLM-controlled robotics," IEEE Transactions on Robotics, 2025, early access, arXiv:2402.10340. DOI:10.1109/TRO.2025.3412345.

[13] E. Kıcıman, R. Ness, A. Sharma, et al., "Causal reasoning and large language models: opening a new frontier for causality," Transactions on Machine Learning Research, 2024.

[14] M. Willig, M. Zečević, D. S. Dhami, et al., "Causal parrots: large language models may talk causality but are not causal," Transactions on Machine Learning Research, 2023.

[15] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge: Cambridge University Press, 2009.

[16] Z. J. Davis, B. Rehder, "A process model of causal reasoning," Cognitive Science, vol. 44, no. 8, e12839, 2020.

[17] B. Rehder, M. R. Waldmann, "Failures of explaining away and screening off in described versus experienced causal learning scenarios," Memory & Cognition, vol. 45, no. 2, pp. 245-260, 2017.

[18] A. Keshmirian, M. Willig, B. Hemmatian, et al., "Biased causal strength judgments in humans and large language models," in ICLR 2024 Workshop on Representational Alignment, 2024.

[19] A. Robey, Z. Ravichandran, V. Kumar, et al., "Jailbreaking LLM-controlled robots," arXiv preprint arXiv: 2410.13691, 2024.

[20] M. Ahn, N. Brohan, Y. Brown, et al., "Do as I can, not as I say: grounding language in robotic affordances," arXiv preprint arXiv:2204.01691, 2022.

[21] A. Brohan, N. Brown, J. Carbajal, et al., "RT-2: vision-language-action models transfer web knowledge to robotic control," arXiv preprint arXiv:2307.15818, 2023.