Neural Network Hypothesis on Interpretation and Universality

Anonymous ACL submission

Abstract

The network hypothesis is theoretically positioning the superiority of the neural network: (I) The neural network is not interpretable in the combination of linear models; (II) There is one general form of the common state with the accuracy in efficiency for any neural network. The new shaping paradigm is more promising to unlock the unknown mechanism than the conventional mathematics in the fundamental of neural network. The neuron activity are not relevant with the meaning in solving the real complex problems. The representation is imitating the human reasoning with the occurrence of the emergent ability and hallucination.

1 Introduction

The neural network achieves the magnificent triumphs with the scientific discovery and technological breakthrough in the cross-domain applications, but their history positioning is seriously undervalued in understanding the serious of the mysterious phenomenon around the engineering practices since the recent renaissance of the deep learning in millennium. The high accuracy in the performance suggests the superiority of the neural network to simulate the real scenarios. Training with the supervised criterion naturally leads to the representation in the hidden layer in the neural network (Goodfellow et al., 2016), and the positive relationships are sometimes consequently transparent in the strict manner between the variables and the responses within the tricky data-driven process.

The neural network is not explainable with the combination of the linear models and functions, and the linear equation is only the specialized form of the neural network. The neural network could better efficiently elaborate the scenario analysis in the real world beyond the expressing boundary of the function in mathematics. The neurons are not the additive components as in the equations. The sharing traditional paradigm physically relating the obtained complex representation with the real problems in the majority of the previous proposals, and some of the theoretical pursuit are in the strong linearity assumption.

The emergent capability and the hallucination might be the two sides of the large language models. The suddenly acquired ability has the zero-shot fashion without any parameter update in the model (Bai et al., 2023), and the jailbreaking attacks is threatening the society safety with the harmful generation from large language models (Ji et al., 2023). The observed phenomenon is boosting the theoretical argument into the decent paradigm and concrete theory on understanding the black magic and potential threats in explainable Artificial Intelligence. DeepSeek R1 naturally emerges with numerous powerful and intriguing reasoning behaviors with the post-training large scale reinforcement learning in prioritizing helpfulness and harmlessness (DeepSeek-AI et al., 2025).

The theoretic pursuit should not always remain absent in explaining the zero-shot performance and few-shot adaption for the neural network and transformer-based model, but unveiling the mystery with the scale parameter on the deep learning algorithms. The scale of the parameters facilitate the selection process in abstracting the complex representation to equip with the acquired ability and the safety-related hallucination. The universality hypothesis reveals the unique status in selecting the states from the model training process. The common state is the universal status with the representation of the complex system with the reasoning ability. The common state is the universal catalyst into the underfitting states or the overfitting states caused from the benchmark, the metrics and the data.

Along with the series of the magic phenomenon for the neural network in unusual, explaining the neural network is attractive in almost every small community in the booming domain. The progres-

sive urgency and Responsible AI are also boosting the exploration on the theoretical essence with the boundary and substance.

The initial purpose is allowing the machine and robots mimicking human-like intelligence as the empathetic companies, insightful advisors and tireless innovators, but the realization of the embryonic and super human AGI are enabling the systems with the consciousness and self-improvement (Feng et al., 2024). One of the intermediate possibilities of the co-creative AI systems is the synthesis of human experiences and machine creativity (Rezwana and Maher, 2024) . The developer community is partly in the conservative attitude to prevent the potential threats to the physic world in controlling the open source strategy on modeling parameter and training secrets of the large language models. Responsible AI advocates for the development of systems aligning with ethical values as fairness and transparency (Sadek et al., 2024). The detailed explanations improve the informational and distributive fairness perceptions subjective judgment in the AI system application in public administration (Aljuneidi et al., 2024). The interdisciplinary dialogue is necessary all the time between regulatory authorities, legislation professionals and social scientists (Gray et al., 2024).

The paper proposes the network hypothesis to shift the explainable AI into the new paradigm on abstracting the semantics of the internal neuron activity with the solving scope in real problems, and ascertains the new position of the neural network with the novel contributions as following,

- The neural network is powerful and superior in representing the symbolized world with the complex possibility and extreme flexibility.

- The representation in the models simultaneously have emerged capability and unexpected hallucination in imitating human reasoning and perception, but in the amateur period with storage constraints from the low efficiency in memory compressing in the spatial and temporal resolution.

- Neural network is the distinct form in reasoning, and the obtained representation is more than the prominent generalized algorithms in the extraordinary structure for the universal dots in the imagination.

2 Literature Review

The transformer based models are one of the great triumphs in the recent renaissance of the research on neural network. The theoretical foundations still remain invalid in internal process and adaption mechanism for the emergent abilities and hallucinations of the large language models. There would be no shortcut to the next breakthroughs in the scientific discovery on the reasoning and memory storage.

The causality initially brings the methods and results from the accumulation in machine learning and philosophy. The symbolic rules follow the mathematical spirit in the previous generation of expert system. The explainable AI is still in the infancy with the earlier prototype on the evaluation and benchmark.

2.1 From Connectionism Renaissance to Emergent Capability

The neural network has the obvious obstacle in violating of the weak rules efficiently and appropriately (Hinton, 1977). The approximation rate and the parsimony of the parameterization of the neural networks are theoretically advantageous than the high-dimensional settings of the polynomial, spline, and trigonometric expansions (Barron, 1993). Dropout could consistently improve generalization accuracy, and is not only a regularizer for preventing overfitting in neural networks (Liu et al., 2023).

The transformer and the variants are universally the popular architecture in the practices and , and the applications is far more fruitful than the original purpose on the task on machine translation (Clark et al., 2019; Sun et al., 2020), including natural language processing (Lin et al., 2022), computer vision (Kashefi et al., 2023), medical image diagnosis (Zhu and Wang, 2023), times series (Wen et al., 2023), and etc. The in context learning capability emerges with the extreme volume of the parameters from the large-scale language models (Dong et al., 2023a; Bai et al., 2023; Dong et al., 2023b) . There is the inappropriate drawback of the multi-stage training with the occurrence of catastrophic forgetting of prior knowledge (Dong et al., 2024)

Hallucinations and harmful content are considered as the undesirable artifact of the large scale language model. The generative model provides the bad response in the certain scenarios of the prompting containing illegal actions, offensive lan-

guage abusing, personal privacy leakage, cybersecurity threatening, unqualified professional advice (Ji et al., 2023). The numerous efforts on the red teaming exhaust the limited resources on fine tuning objectives with the amounts of the model security issues with guardrails on hallucinations and harmful contents(Wei et al., 2023).The persuasive adversarial prompts could increase the jailbreak risk to generate the context without the fully protection on the safety issues(Zeng et al., 2024). The supporting document integration and retrieval augmented generation alleviate the hallucinated content in the response regeneration of LLM-based chatbots (Li et al., 2024). GradSafe accurately detects the jailbreak prompts without necessitating further fine tuning on Llama-2 with safety-critical gradient analysis (Xie et al., 2024). In some real cases, it is acceptable in rejecting the seemingly toxic prompts to prevent malicious output with the side effect in sacrificing answering the innocuous prompts (Cui et al., 2024). DeepSeek utilizes rule-based rewards in mathematics, coding and logical reasoning domain, and captures human preferences in mitigating the potential risks, biases, or harmful content in the generation process (DeepSeek-AI et al., 2025).

The certain chain of thought reasoning is the intermediate steps with the multiple reasoning paths and self-evaluating choices, and the recent survey refers to (Chu et al., 2024), e.g., looking ahead or backtracking in the strategic decisions with the conscious mode (Yao et al., 2023; Prystawski and Goodman, 2023), solving complex mathematics and sequential decision-making problems from the human experiences (Feng et al., 2023).

One of the recent studies empirically identify the complicated situation in quantifying the memorization when considering the training data duplication (Carlini et al., 2023). The empirical studies verify that a well-trained DNN usually encodes sparse, transferable, and discriminative concepts, which is partially aligned with human intuition (Li and Zhang, 2023).

Another line of the work is reconsidering the generalization in the experimental framework on unseen data, (Zhang et al., 2021) e.g., the implicit bias of gradient descent on the linearly separable data (Soudry et al., 2018; Frei et al., 2023), the potential benefits of overparametrization with the well-specified neural networks (Hastie et al., 2022). the generalization ability makes the confusion on the unseen data for the noisy and iterative learning algorithms (Dong et al., 2023b).

2.2 Causality and Symbolism in Explainable Artificial Intelligence

There is the growing consensus on constructing the universal theory of the eXplainable Artificial Intelligence for the neural network (Buchholz et al., 2023; Lorini, 2023; Yu et al., 2023b), and the standardized evaluation is not complete with the decent coverage and the comparison transparency (Le et al., 2023; Delaney et al., 2023). Incremental XAI to automatically partition explanations for general and atypical instances to help users read and remember more faithful explanations (Bo et al., 2024). The hybrid fusion is empowering the domain experts and data-centric explanations collaborating in the interactive systems (Bhattacharya et al., 2024).

The causal model is building the connections between the observed variables and the entire system with the interventionist conditionals in the recent trials on the interpretation of the network mechanism. The theoretic objective of the casual discovery is finding the underlying causal relationships among the observed variables in the earlier trials (Park et al., 2023), e.g., the interdisciplinary practices of the individual treatment effects (Alaa et al., 2023; Imbens and Rubin, 2015; Gunsilius, 2023), the system of the structural equation modeling in the causal reasoning (Lorini, 2023), and the directed acyclic graph to explain the causality in the contemporary machine learning (Squires and Uhler, 2023).

The logic rules has been popular and promising on the reasoning technique with the symbolic models in the domination era of the expert system on the complex system (Bozorgi et al., 2020; Yu et al., 2023a), and the authors in the field of fuzzy logic recently revitalize the interest for the explainable artificial intelligence challenge with the interpretability and accuracy trade-off (Alonso Moral et al., 2021). The fuzzy rule-based models with the ensemble strategy suggest the nonlinear characteristics and substantial interpretability (Hu et al., 2019). The fuzzy logic rule requires the tremendous efforts with the expert knowledge in the mature accumulation on the specific domain (Alonso Moral et al., 2021; Yu et al., 2023a; Baldoni et al., 2018).

3 Interpretability on Neural Network

3.1 Nonlinearity of Neural Network

Theorem 1 (Interpretability Hypothesis) *The neural network cannot (at least has the close to zero probability) be transparently explained by the series of linear equations. The linear equation is one of the specialized forms of the computation in neural network.*

The neural network is likely to be the more advanced toolbox than the linear equation system in recent accumulation of the engineering practices and theoretical construction. The neurons are probably the more effective elementary elements carrying much more information and relations in the neural networks than the variables in the linear regression models and the advanced functions. The properly designed neural network is powerful and efficient in solving the amounts of well-studied classical machine learning problems. The neural network advances the performance accuracy because it treats the things reciprocally in flexibly connected layers, not only in the linear loop or the sequential steps summarized in the logical rules.

The things are all universally connected in the world, but not all variables are equivalently functioning all the situation in the full connection. The phenomenon, incident, and circumstance are much more complex than the single variable described as the predictors and predictions in linear regression models. The neural network is proved to be superior than the linear model and the other basic machine learning algorithms with the benchmark comparison and the concrete examples from academic papers and industrial competitions. The neural network provides the new representation of the diverse world, and the form is more likely to align the complex relations among all the possible hierarchy in the reality.

The connected neuron graph is gradually abstracting the weights and the parameters as the representation through the training process with the intrinsic topological structure. The edges represent for the neuron activity among the arbitrary layer in the structured graph. It is approximating the complex knowledge flowing between the functional neurons for the neural network and the input from the training data. The linear models in the conventional functioning thinking are always struggling in incorporating the compound relations among the neurons activity with the hidden status in the multiple layer perception.

3.2 Generalized Linear Additive Components

The linear regression is generally simplifying the real scenarios into the arbitrary additive components in the biased system settings. The complexity is not always the priority item with the performance and accuracy in the regression-based procedures. It is not the newly problem with the sudden appearing in machine learning, and it originates back into the beginning of the history of the statistics and econometrics (Aydogan et al., 2023). The low dimensional analysis highlights the correlations and interpretability in the ideally approximate linear equations, and the high dimensional data requires new paradigm to explain the cursed mess with great cautions in the innovative paradigms.

The interpretability hypothesis alternatively upgrades the scientific opinions to the opposite of the linearity assumption on the engineering success of the neural network. The new shaping paradigm in the belief paves out the new direction to review our previous foundations on the mathematics and statistics dating back to the strong assumptions in earlier science. The proposed hypothesis is possibly renewing the fundamental concept to the new era of theoretical development in the neural network. The connections are complicated with the individual heterogeneity in the majority of the applications in the domain practices, and the linear regression is too simple in the idealist perspective to capture the randomness on the compound relations among the variables with the unexpected situations in reality.

3.3 Specialized Network in Model Ensemble

The model ensemble is the special case of approximating the neural network in the calculation on real problems. The bagging or boosting schemes seem like the connecting neurons in the network, and the selected models are the prior knowledge borrowed from the specific problems. The last averaging pooling step tolerances the differences on the output results with the intended ambiguity among the bagging models. The weight allocation approximates the neuron activity in remembering and forgetting of information storage in the network.

The neural network is more superior than the regression based models because it considers the individual heterogeneity with the more flexibility in the neuron competition in the weight learning like the bagging and boosting schemes. The model ensemble still has the worse performance in results than the neural network because the ensemble strat-

egy follows the straight linear additive strategy for the limited space in incompletely fitting the problem.

4 Abstract Representing in Diverse Status

Hypothesis 1 (Separability Property) *The internal structure only evolves and derives in the process of obtaining the particular neural network, and it is not always relevant to the decomposing on the reasoning and inference stages on real problems.*

The representation is highly abstract and far away with the initial purpose of the neural network in connecting the features with the recognition task. When we are separating the physical involvement on explaining the network structure and inference stages on the recognition tasks, the claim and the explanation will be more acceptable for the previous proposals. One earlier example is interpretating the internal structural influence in the prominent dropout network in the experimental trials and the engineering achievement.

The dropout almost plays as the last jigsaw in consolidating the task on recognizing the handwritten digits in MNIST dataset,[1] and it is technically proved to be universally better structure in many task-oriented networks originated in computer vision (Kong et al., 2022). The dropout has the wrong intuitive claim on avoid overfitting as their profound instructions on their initial proposed work (Srivastava et al., 2014), and the value of the dropout network is highly undervalued and still under-recognized afterwards in the deep learning community. Dropout is not only the operating component for the internal system to avoid overfitting with the training data (Liu et al., 2023), but refreshing the new potential template with the existing memory storage and the enormous silent neurons in reconstructing the neural network with the path freedom and activity flexibility.

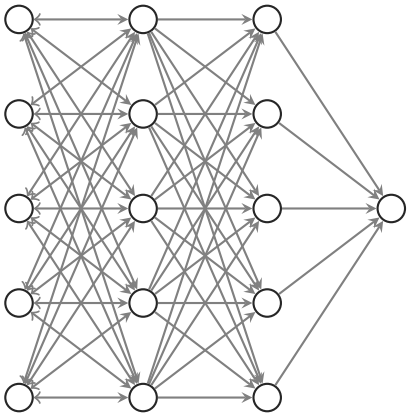
Corollary 1 (Richness Property) *The neurons have at least the two following activating connections in the system, the standing status but silently in the situation, and the key status determining the relevant result predominately in the forms.*

The full connected neurons is sometimes harmful and restricting the complexity in the image cognitive tasks, where all the neurons are equally connected without the hierarchy structure illustrated

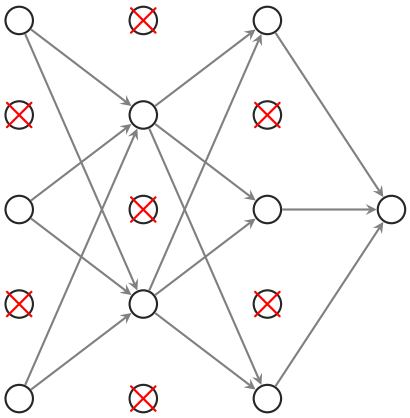
[1] http://yann.lecun.com/exdb/mnist/

Figure 1: Dropout Universally Facilitates Neuron Activities with Memory Storage and Retrieval in the Neural Network

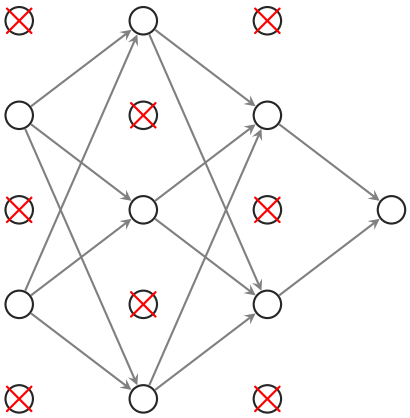
(a) Fully Equalized Connected Network without Flexibility



(b) Dropout Selects Candidate Paths for Multiple Responses



(c) Silence Neurons (not Functioning) in Some Responses



as the instance with 3 layers in Figure 1 (a). The analogy of the standing status helps facilitate the imaginary mechanism of the dropout in the solving process as the last mile sprint on the recognition task of MNIST dataset. The existence of the standing status enriches the networks more vibrantly in the randomness of the connecting possibilities, and strengthening the key player with the multiple potential exempting in the network (Figure 1 b/c). The representation with the diverse status in the network seems like the neural activation and network response in the recent connectome experiments, and almost the majority of the neurons therein are extensively connected, but only few neurons are functioning well in the internal competition on the specific tasks (Ripoll-Sánchez et al., 2023; Randi et al., 2023; Winding et al., 2023).

The previous explanation has the ambiguity in claiming on the internal mechanism of the network with dropout. The interpretation is biased away with the truth in neglecting the richness in neuron status and the creativity unlocked from bringing dropout into the network structure. The connected neurons could have the variety of the status on the entirely full connection in the network, and the neurons are not simultaneously functioning as the equalized participants in all the situational forms of the problem. Some neurons might have the specialized role in some reasoning situations, and might be only the standing or supporting role with the path transition in the complex system.

5 Common State in Continuous Process

Theorem 2 (Universality Hypothesis) *The common state is the universal form with the highest performance on the accuracy, without considering the differentiator in structure as the network layers and neurons.*

The common state has the universal capability in accuracy and efficiency, and it is available to manipulatively select in the training process. The common state distinctively outperforms as the best candidate on the inference and prediction for downstream tasks. The common state has the extreme flexibility to be further adapted to the multiple reasoning objectives, and the required adaptation procedures are relatively the minimum, comparing with the computing efforts for any other obtained state from the original training.

All underlying neural networks intrinsically share the equivalent representation in the universal forms of the evolving state space. The common state is the essence as the intermediate state with the stemming capacity on producing the arbitrary representation with the weights and the layers in the designed structure. The common state is the same form generalized for all the underlying neural network without considering the proper scale of parameters in the structure. The intrinsic structure of presenting the common state is generally obtainable through the currently unknown principles for efficiently manipulating the computing process of shaping the neural network, but the training and fine tuning make the tedious seeking process more mysterious and unstable for the strength and flexibility than the machine learning practices.

Corollary 2 (Continuous Process) *The obtaining process is continuously shifting the representation for any particular state of the observation in shaping the network .*

Corollary 3 (Reversely Availability) *The common states (or the ultimate network) and the neighboring states are gradually consolidated by appropriately aggregating the multiple small networks with the initial separate purpose.*

The neural network in the larger scale is probably emerging with the general reasoning and perceived cognition in the complex representation for the universe, and the neural network in the smaller size is the building block in arbitrary selecting the representation of the common state. If the common state at least has one approximating equivalence in connecting with the smaller scaled network, aggregating the multiple individual networks with the proper topological connection would be one of the alternative ways for obtaining the common state from the large network. The accumulation process simultaneously absorbs the advantage and shortcoming from the smaller potential networks, and optimizes in the computation efficiency with the model capability and reasoning ability in the general purpose.

The common state leverages the universal representation with the sufficient freedom for each neuron in the neural network to understand the new knowledge in the amounts of domains. It is more universally valuable to prioritize in identifying the common states technically and purposefully than theoretically unblocking the suddenly acquired ability in the parameter scale growing under the current wave of large language models. We could design the delicate structure efficiently and incorporate the

multiple trained models in the alternative network structure to obtain the common state of the general representation with the proper manipulation and transformational forms.

6 Restrictions from Memory Storage

Corollary 4 (Resilient Storage) *The ideal model would compress and shrink in the structure without any destruction on the accumulated knowledge in stock with the extreme resilience.*

The model is still struggling in incorporating the breakable rules (Hinton, 1977) in both neural network and transformer-based language models. The large language model is currently incapable of making the perfect decision on following the previous perception with the new duplicate content. The reason might be lacking the accurate representing forms for the overlapping or intermediate status from the mathematics foundation in the fundamental ground. Pursuing the precise equalization is always the priority in the mathematic logic languages and the symbolized world.

The obtained representation might be in the disorder with the duplicate and similar context in the learning than it in the human brain perceiving the observed world. It is not reasonable for removing all the duplicate context for the supreme reasoning model to control the catastrophic forgetting and alleviate the hallucination. One of the intuitive argument is that the deduplication strategy is not exhaustive for the hundreds of gigabytes of training data (Carlini et al., 2023).

The ideal model should more active to consider the natural formation of the memory retrieval and storage in the more flexible and resilient solution. The reason caused the current ambiguity is lacking the understanding the hierarchy structure for the information searching in the secret memory storage mechanism. The understanding remains shallow and vacant on the neuron activity on memory retrieval in human brains (Josselyn and Tonegawa, 2020) from the recently reinvigorated neuroscience research.

The large language model is still highly restricted in the spatial and temporal resolution on the obtained representation. The current wave of the large language model arrange the training process in the structural techniques with the parallel pipelines in the huge computation and energy cost. The scaled resilience would enable the stemming state functioning well with the same capability as the selected state in the array. The large model is parallelized computing in the tidy array, and the short model is flexibly reformatted in scale.

Corollary 5 (Reasoning Assumption) *The large language models do not only memorize the data available in the pretraining and the fine tuning, but emerge with reasoning ability and inference power on the real world in the high level semi-accurate representation.*

The representation in the models, constructed through the conventional linguistic characteristics on word, sentence and chapter (Clark et al., 2019). , are aligning with the real world and the symbolized world in our imagination. If the models could only capture and store the numerous information from the amounts of training data, the obtained model is just fairly representing the training material in the reordered sequence with the refreshed style.

7 Generalized Reasoning with Hallucination

Hypothesis 2 (Deficiency Hypothesis) *The imperfection and inefficiency of neural network are owing to the imitation of neuron activity in human brains, as which bias and errors diversely appear in redundancy.*

The continuous efforts inevitably follow the Turing's idea on the machine intelligence, the earlier pioneering assumption is that the machine is mechanically mimic the human reasoning and free will towards the breakthrough of the potential artificial general intelligence (Turing, 1950).

The current representations in artificial network might be truly close to the memory storage with thinking and reasoning in the human brain (Oota et al., 2023). The artificial network is partly inspiring from the connecting neurons in the human brain, and the unstable representation will probably reoccur in the behavioral appearance with the repetition and redundancy. The reasoning system is naturally not perfect for reasoning inference and cognitive tasks in human brains. The smarter people still have the chance to make mistakes for the simple repetitive tasks without sufficient training in the continuous learning. Our wording and phrasing are not as stable as the rocks and stones, and it is normal to have the reordered sequences in the reoccurring appearances from the verbal language. The logic and mathematics are somewhat the utopia in our imagined world.

If the general representation has the capability as the intrinsic property in reasoning and understanding, lacking human sense and sensibility becomes problematic for the current assumption on the prototype of general intelligence as the reliable helper and trustful advisor. We are rarely considering the emotional feelings and personalized attitudes with the scope in the philosophy foundations in accumulated knowledge (Asai et al., 2020; Wei et al., 2023), and less incorporating the human values in the initial learning procedures and material orientation.

If the improper training data caused the unexpected jailbreak cases, it is still the new challenge to entirely remove the numerous false material in training the new model again from the bottom with almost the same complexity and equivalent capacity. Intuitively, there are plenty of the illicit opinions in the crawled online data and the licensed research data in the disclosure of the training recipe. The situation is that the previous released unaligned model could produce the undesirable answers in the potential reasoning with the unseen data.

One possible explanation on the hallucination is equally treating the multiple sources of the training material in initialing the structure design in the model. If human could differentiate episodic memory and semantic memory with the mature internal mechanism structurally in the brain (Budson, 2009), the limited information will still be the obstacle in linking the memory and perception in the machine. The personality of the large models pretend to be the memory disorder from aligning the heterogeneous document without the boundary between episodic memory and semantic memory.

Corollary 6 (Co-occurrence Complexity) *If emergent capability simultaneously occurs with hallucinations in the large scaled model, implying the potential of the constructed system with the unexplored chaos.*

Both the academic and industrial communities require the more openness and transparency on the training process to analyze the current model limitation and the hallucinations with the better quality in the artificial intelligence. The business owner now partly takes the relatively conservative attitude with the developer responsibility to alleviate the harmful scenarios in the frontier technology application. It is reasonable for the regulators to hamper the rapid development in control for the safety prevention and behavior surveillance in the public release of current LLMs. The research and business instances would become concrete and solid continuously with the complete information disclosure to reproduce the current models for fulfilling the incompleteness and deficiency of the large language models. We should encourage the new ideas and novel innovation on the academic research and industrial commercialization in the open and corroborative environment, and restrict the unexpected use with the responsible framework to balance the social welfare and the illicit risks of the sudden emergence of the potential general artificial intelligence.

8 Limitation

The network hypothesis is not directly concentrating on the structure design and model training in the practice, but the new paradigm is helpful and promising for understanding the unrevealed secrets in imitating the human reasoning and perceptions.

The co-occurrence of the emerged ability and unexpected hallucination is only one possible explanation for the ongoing debate on the inner mechanism of the phenomenon, and the empirical study is much more challenging on the large language model to examine the hypothesis and corollary partly beyond the normal level of the intelligence.

References

Ahmed Alaa, Zaid Ahmad, and Mark J. van der Laan. 2023. Conformal meta-learners for predictive inference of individual treatment effects. In *the Thirty-seventh Annual Conference on Neural Information Processing Systems*.

Saja Aljuneidi, Wilko Heuten, Larbi Abdenebaoui, Maria K. Wolters, and Susanne Boll. 2024. Why the fine, ai? the effect of explanation level on citizens' fairness perception of ai-based discretion in public administrations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 318:1–318:18. ACM.

Jose Maria Alonso Moral, Ciro Castiello, Luis Magdalena, and Corrado Mencar. 2021. *Toward Explainable Artificial Intelligence Through Fuzzy Systems*, pages 1–23. Springer International Publishing, Cham.

Sara Asai, Koichiro Yoshino, Seitaro Shinagawa, Sakriani Sakti, and Satoshi Nakamura. 2020. Emotional speech corpus for persuasive dialogue system. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 491–497, Marseille, France. European Language Resources Association.

Ilke Aydogan, Loïc Berger, Valentina Bosetti, and Ning Liu. 2023. Three Layers of Uncertainty. *Journal of the European Economic Association*, 21(5):2209–2236.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 2023. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *the Thirty-seventh Annual Conference on Neural Information Processing Systems*.

Roberto Baldoni, Emilio Coppa, Daniele Cono D'Elia, Camil Demetrescu, and Irene Finocchi. 2018. A survey of symbolic execution techniques. *ACM Comput. Surv.*, 51(3):50:1–50:39.

Andrew R. Barron. 1993. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory*, 39(3):930–945.

Aditya Bhattacharya, Simone Stumpf, Lucija Gosak, Gregor Stiglic, and Katrien Verbert. 2024. EXMOS: explanatory model steering through multifaceted explanations and data configurations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 314:1–314:27. ACM.

Jessica Y. Bo, Pan Hao, and Brian Y. Lim. 2024. Incremental XAI: memorable understanding of AI with incremental explanations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 315:1–315:17. ACM.

Zahra Dasht Bozorgi, Irene Teinemaa, Marlon Dumas, Marcello La Rosa, and Artem Polyvyanyy. 2020. Process mining meets causal machine learning: Discovering causal rules from event logs. In *2020 2nd International Conference on Process Mining (ICPM)*, pages 129–136.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. 2023. Learning linear causal representations from interventions under general nonlinear mixing. In *the Thirty-seventh Annual Conference on Neural Information Processing Systems*.

Andrew E. Budson. 2009. Understanding memory dysfunction. *The Neurologist*, 15:71–79.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. Navigate through enigmatic labyrinth A survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1173–1203. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2024. Or-bench: An over-refusal benchmark for large language models. *CoRR*, abs/2405.20947.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dong-Li 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, Jiong 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, M. 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 Zhou, Shaokang Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wen-Xia 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, Xiangyu Jin, Xi-Cheng 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, Yi Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yao Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Yukun Zha, Yuting Yan, Zehui Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen

guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, 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.

Eoin Delaney, Arjun Pakrashi, Derek Greene, and Mark T. Keane. 2023. Counterfactual explanations for misclassified images: How human and machine explanations differ. *Artificial Intelligence*, 324:103995.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 177–198. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023a. A survey for in-context learning. *ArXiv*, abs/2301.00234.

Yuxin Dong, Tieliang Gong, Hong Chen, and Chen Li. 2023b. Understanding the generalization ability of deep learning algorithms: A kernelized rényi's entropy perspective. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 3642–3650. ijcai.org.

Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: a theoretical perspective. In *the Thirty-seventh Annual Conference on Neural Information Processing Systems*.

Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. 2024. How far are we from AGI. *CoRR*, abs/2405.10313.

Spencer Frei, Gal Vardi, Peter L. Bartlett, Nathan Srebro, and Wei Hu. 2023. Implicit bias in leaky relu networks trained on high-dimensional data. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Colin M. Gray, Cristiana Teixeira Santos, Nataliia Bielova, and Thomas Mildner. 2024. An ontology of dark patterns knowledge: Foundations, definitions, and a pathway for shared knowledge-building. In *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pages 289:1–289:22. ACM.

F. F. Gunsilius. 2023. Distributional synthetic controls. *Econometrica*, 91(3):1105–1117.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. 2022. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986.

Geoffrey E. Hinton. 1977. *Relaxation and its role in vision*. Ph.D. thesis, University of Edinburgh, UK.

Xingchen Hu, Witold Pedrycz, and Xianmin Wang. 2019. Random ensemble of fuzzy rule-based models. *Knowledge-Based Systems*, 181:104768.

Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38.

Sheena A. Josselyn and Susumu Tonegawa. 2020. Memory engrams: Recalling the past and imagining the future. *Science*, 367(6473):eaaw4325.

Rojina Kashefi, Leili Barekatain, Mohammad Sabokrou, and Fatemeh Aghaeipoor. 2023. Explainability of vision transformers: A comprehensive review and new perspectives. *CoRR*, abs/2311.06786.

Xiangtao Kong, Xina Liu, Jinjin Gu, Yu Qiao, and Chao Dong. 2022. Reflash dropout in image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5992–6002. IEEE.

Phuong Quynh Le, Meike Nauta, Van Bach Nguyen, Shreyasi Pathak, Jörg Schlötterer, and Christin Seifert. 2023. Benchmarking explainable AI - A survey on available toolkits and open challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6665–6673. ijcai.org.

Mingjie Li and Quanshi Zhang. 2023. Does a neural network really encode symbolic concepts? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 20452–20469. PMLR.

Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. 2024. Citation-enhanced generation for llm-based chatbots. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 1451–1466. Association for Computational Linguistics.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI Open*, 3:111–132.

Zhuang Liu, Zhiqiu Xu, Joseph Jin, Zhiqiang Shen, and Trevor Darrell. 2023. Dropout reduces underfitting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22233–22248. PMLR.

Emiliano Lorini. 2023. A rule-based modal view of causal reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 3286–3295. ijcai.org.

Subba Reddy Oota, Manish Gupta, Raju S. Bapi, Gaël Jobard, Frédéric Alexandre, and Xavier Hinaut. 2023. Deep neural networks and brain alignment: Brain encoding and decoding (survey). *CoRR*, abs/2307.10246.

Junhyung Park, Simon Buchholz, Bernhard Schölkopf, and Krikamol Muandet. 2023. A measure-theoretic axiomatisation of causality. In *the Thirty-seventh Annual Conference on Neural Information Processing Systems*.

Ben Prystawski and Noah D. Goodman. 2023. Why think step-by-step? reasoning emerges from the locality of experience. In *the Thirty-seventh Annual Conference on Neural Information Processing Systems*.

Francesco Randi, Anuj Kumar Sharma, Sophie Dvali, and Andrew Michael Leifer. 2023. Neural signal propagation atlas of caenorhabditis elegans. *Nature*, 623:406 – 414.

Jeba Rezwana and Mary Lou Maher. 2024. Conceptual models as a basis for a framework for exploring mental models of co-creative AI (short paper). In *Joint Proceedings of the ACM IUI 2024 Workshops co-located with the 29th Annual ACM Conference on Intelligent User Interfaces (IUI 2024), Greenville, South Carolina, USA, March 18, 2024*, volume 3660 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Lidia Ripoll-Sánchez, Jan Watteyne, HaoSheng Sun, Robert Fernandez, Seth R. Taylor, Alexis Weinreb, Barry L. Bentley, Marc Hammarlund, David M. Miller, Oliver Hobert, Isabel Beets, Petra E. Vértes, and William R. Schafer. 2023. The neuropeptidergic connectome of c. elegans. *Neuron*, 111(22):3570–3589.e5.

Malak Sadek, Marios Constantinides, Daniele Quercia, and Celine Mougenot. 2024. Guidelines for integrating value sensitive design in responsible ai toolkits. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. 2018. The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.*, 19(1):2822–2878.

Chandler Squires and Caroline Uhler. 2023. Causal structure learning: A combinatorial perspective. *Found. Comput. Math.*, 23(5):1781–1815.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3525–3535, Online. Association for Computational Linguistics.

Alan M. Turing. 1950. Computing machinery and intelligence. *Mind*, LIX(236):433–460.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? In *the Thirty-seventh Annual Conference on Neural Information Processing Systems*.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2023. Transformers in time series: A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 6778–6786. ijcai.org.

Michael Winding, Benjamin D. Pedigo, Christopher L. Barnes, Heather G. Patsolic, Youngser Park, Tom Kazimiers, Akira Fushiki, Ingrid V. Andrade, Avinash Khandelwal, Javier Valdes-Aleman, Feng Li, Nadine Randel, Elizabeth Barsotti, Ana Correia, Richard D. Fetter, Volker Hartenstein, Carey E. Priebe, Joshua T. Vogelstein, Albert Cardona, and Marta Zlatic. 2023. The connectome of an insect brain. *Science*, 379(6636):eadd9330.

Yueqi Xie, Minghong Fang, Renjie Pi, and Neil Gong. 2024. Gradsafe: Detecting jailbreak prompts for llms via safety-critical gradient analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 507–518. Association for Computational Linguistics.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *the Thirty-seventh Annual Conference on Neural Information Processing Systems*.

Dongran Yu, Bo Yang, Dayou Liu, Hui Wang, and Shirui Pan. 2023a. A survey on neural-symbolic learning systems. *Neural Netw.*, 166(C):105–126.

Zhongwei Yu, Jingqing Ruan, and Dengpeng Xing. 2023b. Explainable reinforcement learning via a causal world model. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 4540–4548. ijcai.org.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14322–14350. Association for Computational Linguistics.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2021. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115.

Dongmei Zhu and Dongbo Wang. 2023. Transformers and their application to medical image processing: A review. *Journal of Radiation Research and Applied Sciences*, 16(4):100680.