This article has been accepted for publication in IEEE Intelligent Systems. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MIS.2025.3621938

Trustworthy Machine Learning in the Era of Foundation Models

Bo Han[1]
[1]TMLR Group, Department of Computer Science, Hong Kong Baptist University

August 2025

Abstract

This position paper examines *trustworthy machine learning with foundation models* by investigating the four essential aspects of *learning*, *reasoning*, *planning*, and *multimodality*. In learning, we describe how pre-training, fine-tuning, and reinforcement learning enable models to acquire generalizable knowledge and emphasize the importance of high-quality, unbiased training data as well as robust training methods. In reasoning, we summarize the fundamental methodology of training-free, post-training, and test-time scaling methods that enhance logical deduction, reasoning transparency, and systematic safety. In planning, we incorporate neuro-symbolic methods that combine adaptable neural capabilities with formally verifiable symbolic reasoning, ensuring safe and accountable decision-making. Finally, we investigate the need for multimodal integration, where aligning information from different sensory input sources is important to mitigate biases and errors. In summary, the paper presents an interdisciplinary vision for incorporating capability, robustness, safety, and explainability to establish trustworthy foundation models, paving the way for their reliable deployment in real-world applications, especially high-stakes domains such as financial and clinical decision-making.

1 Trustworthy Learning

Learning is the primary way that allows foundation models to parameterize general and transferable knowledge, of which trustworthiness is critical. It usually involves using large-scale raw data or high-quality crafted data to update model parameters through gradient-based optimization, converging to some local minimum points that are defined by the learning objectives, such as negative log-likelihood [1, 23]. Generally, in the context of foundation models, learning typically involves a wide range of tasks and modalities that cover both the language and vision domains. The remarkable advancements of foundation models have been fueled by recent changes in learning paradigms, showing significant performance of these models in a variety of real-world applications.

Pre-training and fine-tuning are the two main stages in the learning process of foundation models, as well described in the current literature. In particular, by training models on large-scale and diverse datasets, typically collected from the Internet or other raw data sources, pre-training enables foundation models to acquire broad and general-purpose representations. It is well known that data scale is the key factor in the success of learning in foundation models, as these models are extremely large and inherently data-hungry. Many open-source datasets have been compiled to support large-scale learning using this principle [20, 7]. Since they offer the scale and diversity required for pre-training advanced foundation models, these open-source datasets have played a crucial role in the research and development community of foundation models.

Beyond data scale, researchers have recently highlighted the need for data quality to improve trustworthiness. Many related works have reported that using high-quality, small-scale datasets can achieve superior performance than low-quality, large-scale raw datasets for training [27]. There are two primary types of low-quality data: noisy data and biased data. Many works use crowdsourcing or rule-based cleansers to improve data quality. However, crowdsourcing, while effective for annotation, is expensive and suffers from varying quality; rule-based cleansers lack flexibility and are error-prone when dealing with complicated or nuanced data. The development of robust methods represents a promising area of research in trustworthy learning, where the resulting models should be as reliable as those trained on high-quality data of the same scale.

Researchers can adopt traditional methods in label noise robustness to design noise-tolerant learning objectives through noise transition modeling, and the adversarial training techniques to teach models to be inherently noise-tolerant and robust. In the same way, robust regularization strategies like dropout, mixup, and co-training are also promising in improving generalization and avoiding overfitting. Other related fields of trustworthiness, such as out-of-distribution detection, also hold potential to benefit the robust learning capabilities of foundation models. By adopting these promising paths, pre-training will further improve trustworthy learning for foundation models.

Fine-tuning refers to the process of improving pre-trained models on small-scale, more task-specific datasets that are made to fulfill particular needs. In general, there are two main types of fine-tuning: supervised fine-tuning [32] and reinforcement learning [21]. Specifically, supervised fine-tuning uses objective functions and optimization methods that are quite similar to those employed during pre-training. However, it can use the inherent few-shot ability achieved during pre-training, thus requiring fewer data and making training easier. Reinforcement learning, on the other hand, uses a reward system to shape the learning behaviors of models. It indicates that it is more aligning with the principles of exploration and exploitation, instead of having clear learning goals like supervised fine-tuning. The goal of both ways of fine-tuning is to further enhance the trustworthiness of pre-trained models, so that models can become better at certain tasks or learn new skills like coding and math reasoning. Therefore, supervised fine-tuning shares many similar pros and cons with pre-training; the discussion below will focus mainly on reinforcement learning.

Although reinforcement learning is often reported to achieve superior performance compared to supervised learning—particularly for complex, preference-driven tasks—it comes at the cost of tedious hyperparameter tuning, meticulous reward design, and numerous trial-and-error iterations. It challenges the trustworthiness of the fine-tuning procedure, primarily caused by the absence of clear learning targets, which forces the model to explore optimization directions that can truly yield higher rewards. To address these challenges, researchers have explored a number of advanced methods. This includes reward engineering, which learns the reward function from expert demonstrations; model-based reinforcement learning, which employs an environment model to simulate interactions; and off-policy learning, which uses pre-defined preference datasets to ease exploration. Even with these improvements, the learning dynamics in reinforcement learning are still not well understood, which makes it hard to enhance approaches beyond heuristics. This uncertainty makes fine-tuning harder because the model and environment can interact in unpredictable ways.

When decomposing reinforcement learning objectives into simpler components, we identify a seldom-discussed element, namely, unlearning [26]. Unlike standard learning, which increases data likelihood, unlearning explicitly reduces the probability of specific responses or trajectories. Researchers in related fields have reported contradictory effects of unlearning: while some observe impaired generalization, others argue it steers the prediction distribution toward potentially more promising outputs. This tension is compounded by traditional viewpoints suggesting that unlearning should theoretically fail due to its unbounded nature. These mutually incompatible observations and conjectures obscure the precise mechanistic role of unlearning in reinforcement learning. We posit that, without proper tuning, even minor perturbations or low-quality data may catastrophically distort learning dynamics—precipitating reward hacking and optimization collapse. Overall, toward trustworthy learning, a formal interpretation of reinforcement learning dynamics—particularly through gradient analysis and parameter sensitivity—requires further investigation. Such studies will help elucidate the mechanistic role of unlearning, diagnose its limitations, and catalyze the development of more robust and effective methods. These investigations will substantively enhance the reliability and generality of reinforcement learning techniques while advancing the broader field of foundation model research.

2 Trustworthy Reasoning

Foundation model reasoning illustrates a promising new horizon in artificial intelligence with emerging capabilities such as mathematical reasoning, problem-solving, and decision-making. The reasoning capability of foundation models comes from training on massive and diverse data sources, enabling the models to identify relationships [13], generate cohesive narratives [35], and perform abstract thinking [37] that was once thought to be uniquely human.

Reasoning with foundation models has evolved from simply matching patterns to more logic-based forms of structured deduction that allow for systems to

engage with a multitude of problems, from solving mathematical proofs [28] and scientific hypotheses [19] to contextual understanding in natural language conversations [6]. These models engage in a mixture of symbolic and probabilistic reasoning to create reasoning chains that extend beyond superficial interactions. This paradigm change enables new ways in which reasoning is no longer a separate add-on, but an integral part of the model's ability to interact and reason with rich inputs. In this context, foundation model reasoning becomes the point of convergence between previous rule-based systems and current learning-based systems, blending interpretability and adaptability under a unified framework. This change not only enables superior performance for tasks that require hierarchical thinking but also challenges the conventional idea of machine intelligence, serving as a foundation for future AI systems that can process both structured and unstructured real-world information.

The methods for reasoning within a foundation model can generally be categorized into prompt-based, post-training, and test-time methods, as well as those that use external tools. As for prompt-based methods, Chain-of-Thought (CoT) [31], Tree-of-Thought (ToT) [36], and Monte Carlo Tree Search (MCTS) [39] represent efforts to prompt reasoning capability from the pre-trained models without changing the models' weights. CoT motivates the model to produce intermediate logical steps iteratively rather than producing a final answer immediately and explicitly. ToT extends this concept as the model can be prompted to explore multiple paths in a decision or thought space with a tree structure that accounts for a breadth of options before settling on a solution. MCTS instructs a model to simulate a probabilistic area with different possible branches of reasoning before making its final decision.

Post-training methods, such as Supervised Fine-tuning (SFT) [5] and Reinforcement Learning (RL) [22], are effective to boost the models' reasoning capabilities by using the curated data and fine-tuning the model outputs to better align with human expectations of logical coherence and reasoned accuracy. These methods allow the researchers to iteratively impose desirable reasoning behavior on models, which is not afforded during the initial pre-training phase. Furthermore, test-time methods by either test-time scaling [33] or self-evolve mechanisms [4] provide an "adaptive" framework, where reasoning behavior can be adjusted on the fly during inference, by scaling the amount of computational budgets, or potentially even self-correcting wrong thoughts in "real time" while solving the problem at hand. Finally, the inclusion of external tools, such as calculators [29], symbolic solvers [16], search engines [10], or even verified knowledge databases [12], is another degree of augmentation, from which the model-specific reasoning process is enhanced through external sources. These external tools can be integrated with the model to assist the problem-solving that requires precision in reasoning, or domain-specific knowledge that the base model may not have sufficiently obtained during pre-training.

As is common with any developing technology, *trustworthy reasoning* is an important factor when assessing how to reliably deploy the advanced foundation model. The current primary challenge in this realm is ensuring safety against adversarial prompts and instructions. This involves malicious users deliberately

crafting prompt input to force the model to generate harmful, biased, or misleading reasoning patterns. To mitigate these risks, developers of foundation models need to enact sufficient safeguards that can recognize and stop these adversarial attempts in the moment. In addition, foundation models should also have enough robustness to tolerate noisy input and incomplete information, as real-world scenarios rarely provide fully complete, well-structured, and noise-free data. In these ways, the models can demonstrate the ability to extrapolate the missing information reasonably and consider the distortions from the inferior conditions to perform robust reasoning.

Another component of trustworthy reasoning is the explainability and interpretability of the reasoning process. Rather than black-box reasoning, the future of model reasoning demands explainable logic pathways that are auditable, verifiable, and understandable to humans. By providing transparent explanations of intermediate reasoning steps, humans can identify where the logical chains of models originated, diverged, or went wrong. Adding such clarity will also lead to greater trust and collaboration between humans and machines, reduce biases, and provide assurances that decisions made from these reasoned outputs are both fair and verifiable by either humans or the machines themselves. As a result, robust frameworks of evaluation and measurement should be developed with technical aspects so that safety, robustness, and explainability are not merely secondary considerations, but are embedded principles during the development and deployment of reasoning models.

Despite the challenges regarding trustworthiness, foundation models with advanced reasoning methods are widely employed in a wide range of applications, from scientific discoveries to working in the real world as an agent. In terms of science, these models can alter how we think and reason about complex problems in mathematics, biology, and chemistry. For example, in mathematics, models that demonstrate the ability to reason deeply work together to solve complex proofs or create new theorems by exploring novel conceptual spaces. In scientific domains like biology and chemistry, models create novel drug candidates or aid in the understanding of complex biochemical networks, providing insights that accelerate research and innovation.

Meanwhile, real-world applications rapidly catch up with these advances, benefiting humans' daily work and life. For example, coding agents change the pattern of software development by applying sophisticated reasoning to debug, generate, and optimize complex code bases. GUI agents provide significant benefits by demonstrating reasoning while navigating and controlling user interfaces. Web agents offer functions to retrieve and aggregate information from the Internet. Search agents incorporate reasoning strategies to respond to increasingly deep and contextually aware queries.

These examples show that when foundation models are combined with systematic reasoning methods, they have the potential to understand and act on tasks requiring multi-step logic and decisions. These agents transition from being static information processors to dynamic and adaptive AI systems capable of reasoning about specific contexts, self-correcting in the context of problem-solving, and responding to new and changing environmental conditions. As re-

search continues to advance, the development and implementation of reasoning abilities, coupled with rigorous methods, reputable design and implementation principles, and framing applications, could significantly change our expectations of machine intelligence and offer prospects of using AI systems as development and trustworthy assistants in both academia and industry.

3 Trustworthy Planning

Planning is a central component of intelligent behavior, yet as agents move into safety-critical settings like autonomous vehicles and healthcare, ensuring their trustworthiness has become a societal and ethical obligation. In the era of large foundation models, such as GPT-4 [1], the field of AI planning is changing rapidly. While these models offer unprecedented generalization, their opaque reasoning and lack of formal guarantees make their trustworthiness for real-world planning questionable, especially where safety, transparency, and human-alignment are paramount.

Neuro-symbolic AI provides a natural response to these challenges by combining the adaptability of neural models with the verifiability and transparency of symbolic reasoning. Symbolic representations, like linear temporal logic [18] or task hierarchies [11], decompose goals and enforce constraints. Neural models, in turn, process high-dimensional sensory inputs and natural language, grounding the symbolic structure in flexible, real-world interaction. This combination enables intelligent systems to ground their planning with structures that support formal guarantees and human-readable reasoning.

New studies show how symbolic goals can control foundation model outputs. For instance, large language models like SayCan [2] can map high-level goals to executable plans while relying on symbolic controllers to check feasibility. Differentiable logic-based methods like DeepProbLog [15] embed symbolic constraints directly into neural pipelines, enhancing interpretability. Verification-aware learning [3] trains agents with formal safety constraints to ensure safe behaviors before deployment. Furthermore, symbolic models excel at managing uncertainty, a critical capability for acting with partial observability, and their logical traces provide step-by-step causal explanations, a feature absent in end-to-end black-box systems.

The advantages of neuro-symbolic integration are particularly pronounced for large foundation models. While foundation models provide compositional reasoning, their behavior is difficult to control, and they can generate unsafe or spurious actions. Early efforts like Voyager [24] and Auto-GPT show that outputs of large language models require symbolic scaffolding to remain aligned with human intent. This symbolic guidance introduces a constraint layer that encodes task structure or user intent in an enforceable format. Future work must generate architectures capable of rapid decision-making under partial observability. Trustworthy systems in cluttered or uncertain environments will need to tightly integrate neural perception with symbolic belief tracking, enabling high-level, interpretable planning despite noisy sensory input. A fundamental

aspect of this is learning symbolic abstractions—such as causal graphs or task hierarchies—directly from foundation model representations, rather than hand-designing them. Such learned abstractions can enable generalizable planning across environments while adhering to traceable, structured reasoning.

It is also important to incorporate formal verification into the training and inference phases. Unlike classical planners that leverage model checking, systems built on existing foundation models often lack built-in safety guarantees. New approaches based on differentiable logic and constrained policy optimization allow safety specifications to be embedded directly into the model's learning objectives or architecture. This preventive enforcement of correctness is a crucial step toward scalable and accountable planning. For human-AI collaboration, symbolic structures are key to representing user goals, preferences, and ethical boundaries. This moves agents from being reactive to being trustworthy collaborators who can negotiate goals and defer to humans, making decisions that are understandable and revisable.

In summary, trustworthy planning in foundation models requires a hybrid approach that leverages the flexibility of neural models while ensuring their behavior adheres to verified, interpretable, and human-aligned symbolic systems. Neuro-symbolic integration presents a viable framework to meet this challenge, offering the power of large-scale learning while retaining the reliability and control necessary for deployment in high-stakes, safety-critical systems. As foundation models continue to shape AI research, embedding trust in their planning abilities will be paramount for their responsible adoption into our world.

4 Trustworthy Multimodality

While the preceding sections have examined learning, reasoning, and planning as foundational pillars of trustworthy AI, they have largely been discussed through a unimodal, predominantly text-centric lens, treating models as disembodied intelligences operating on abstract symbols. The next frontier for foundation models, and arguably the most critical step towards artificial general intelligence, lies in multimodality—the ability to seamlessly process, integrate, and act upon information from diverse data streams such as images, audio, and video alongside text, as seen in modern Large Multimodal Models (LMMs) like GPT-4V, Gemini, and Qwen-VL [25]. This leap from abstract manipulation to a grounded, holistic understanding of the world is essential for developing AI that can operate with human-like versatility and situational awareness. However, this richness is a double-edged sword: while it grants models unprecedented capabilities [38], the fusion of modalities introduces unique and compounded challenges to trustworthiness. It creates new surfaces for error, bias, and unpredictability that are far more complex than those encountered in single-modality systems. This section revisits the core pillars of learning, reasoning, and planning through this transformative and challenging perspective, exploring how to build trust in models that are designed to see, hear, and read our world.

In a multimodal context, the paradigm of learning shifts fundamentally. The

central goal becomes cross-modal alignment. This objective was initially pursued through methods like contrastive learning [17]. More advanced LMMs, such as LLaVA, have evolved this paradigm by connecting a pre-trained visual encoder to an LLM via a lightweight adapter, and then fine-tuning the model on large-scale multimodal instruction-following datasets [14]. The learning objective in this phase is primarily autoregressive—predicting the next token—which endows the model with instruction-following capabilities. This process does more than just associate pixels with words; it forces the model to form modality-agnostic abstract concepts. However, this learning process is fraught with new threats to trustworthiness. The issue of noisy data evolves into the more pernicious problem of misalignment noise. Furthermore, multimodality can act as a powerful vessel for bias amplification. A model might learn to associate text descriptions of "programmers" with images of men, fusing linguistic biases with visual stereotypes to create a more potent and deeply embedded form of prejudice [30]. This is especially dangerous as these biases are encoded in subtle visual cues that are harder to audit and mitigate. Lastly, multimodal learning is highly susceptible to spurious correlations. A failure to generalize rooted in a superficial understanding undermines robust performance. Therefore, building trust requires developing novel alignment methods that can disentangle causal attributes from statistical noise and mitigate societal biases, a topic actively explored in recent safety-focused research.

Reasoning in a multimodal context transcends the abstract, symbolic logic of text-based problem-solving, becoming a grounded process where logical steps must be anchored in, and validated against, evidence from other modalities. Advanced techniques like Chain-of-Thought must evolve into a multimodal format (Multimodal-CoT) [40]. While this technique has significant ability, it concurrently gives rise to a critical trustworthiness challenge–the hallucination issue. A multimodal model hallucinates when it generates descriptions of objects or relationships that are factually absent from the visual data [8]. This can severely undermine its reliability, and recent work has focused on creating benchmarks and methods to evaluate and mitigate such object-level hallucinations. Consequently, the challenge of explainability is magnified exponentially. A trustworthy explanation must demonstrate how specific visual features influenced the textual output, making the black box of its reasoning more auditable.

Ultimately, the challenges of multimodal learning culminate in planning, especially for embodied AI and robotics, where the stakes are physical [9]. Here, neuro-symbolic principles are applied in a continuous perception-action loop. For instance, a robot told to "get the apple" must visually identify the "apple" and translate "get" into a sequence of actions: extend arm, grasp, and retract. The system's trustworthiness is critical, as perceptual errors can have immediate, unsafe physical consequences. An industrial robot misidentifying a human arm for a part could be catastrophic. This highlights the symbol grounding problem: reliably connecting abstract symbols like "hot" to diverse, real-world sensory data, not just to a specific stove but to any dangerously heated object. Therefore, a trustworthy planner must not only execute but also constantly monitor progress, detect anomalies, and adapt or terminate safely in real time [34].

When considering a scenario where a delivery drone reroutes mid-flight to avoid an unexpected obstacle, building such systems necessitates the integration of robust multimodal learning, causal reasoning, and verifiable planning—advances that pave the way for AI to evolve into a capable and dependable partner in our physical world.

5 Conclusion

As foundational models become increasingly prevalent in the components of intelligence systems, the trustworthiness across learning, reasoning, and planning in multimodality scenarios has become a central imperative. This position paper calls for a future where robustness, reliability, interpretability, and human alignment are not afterthoughts, but integral design goals. The success of this vision requires a systematic change: moving operationalized aspects of formal guarantees, neuro-symbolic mechanisms, and value-sensitive optimization to the center of machine learning. In the multimodal domain that converges on language, vision, and action, trust must be bolstered through grounded alignment, uncertainty calibration, and cross-modal consistency to prevent misconduct. Meanwhile, as these models are widely deployed in high-risk domains such as healthcare, finance, and autonomous platforms, trustworthiness must be infused from the ground up to ensure safety, accountability, and alignment with society. In this way, establishing trustworthy machine learning in the age of foundation models will require intensive interdisciplinary synergy, robust analytical methodologies, and scalable engineering approaches, which build intelligent systems that are powerful, principled, transparent, and aligned with human values.

Acknowledgement

Prof. Bo Han would like to express his sincere thanks to his PhD students in TMLR Group, including Jianing Zhu, Qizhou Wang, Xue Jiang, and Zhanke Zhou (in alphabetical order), for their great contribution to the discussion and refinements of this position paper.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as I can, not as I say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

[3] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. In *NeurIPS*, 2018.

[4] Jingyi Chai, Shuo Tang, Rui Ye, Yuwen Du, Xinyu Zhu, Mengcheng Zhou, Yanfeng Wang, and Siheng Chen. SciMaster: Towards general-purpose scientific AI agents, Part I. x-master as foundation: Can we lead on humanity's last exam? *arXiv preprint arXiv:2507.05241*, 2025.

[5] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

[6] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

[7] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[8] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *AAAI*, 2024.

[9] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3D value maps for robotic manipulation with language models. In *CoRL*, 2023.

[10] Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

[11] Leslie Pack Kaelbling and Tomás Lozano-Pérez. Integrated task and motion planning in belief space. *The International Journal of Robotics Research*, 2013.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

[13] Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*, 2025.

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[15] Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural probabilistic logic programming. In *NeurIPS*, 2018.

[16] Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*, 2023.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICLR*, 2021.

[18] Vasumathi Raman, Alexandre Donzé, Dorsa Sadigh, Richard M Murray, and Sanjit A Seshia. Reactive synthesis from signal temporal logic specifications. In *HSCC*, 2015.

[19] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 2024.

[20] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022.

[21] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[22] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Yu-kun Li, Yang Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[24] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[25] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[26] Qizhou Wang, Bo Han, Puning Yang, Jianing Zhu, Tongliang Liu, and Masashi Sugiyama. Towards effective evaluations and comparisons for LLM unlearning methods. In *ICLR*, 2025.

[27] Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of clips to spurious features. In *NeurIPS*, 2024.

[28] Ruida Wang, Jipeng Zhang, Yizhen Jia, Rui Pan, Shizhe Diao, Renjie Pi, and Tong Zhang. TheoremLlama: Transforming general-purpose LLMs into lean4 experts. In *EMNLP*, 2024.

[29] Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. Executable code actions elicit better LLM agents. In *ICML*, 2024.

[30] Zecheng Wang, Xinye Li, Zhanyue Qin, Chunshan Li, Zhiying Tu, Dianhui Chu, and Dianbo Sui. Can we debias multimodal large language models via model editing? In *ACM MM*, 2024.

[31] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[32] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *NeurIPS*, 2022.

[33] Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *ICLR*, 2024.

[34] Wenpeng Xing, Minghao Li, Mohan Li, and Meng Han. Towards robust and secure embodied AI: A survey on vulnerabilities and attacks. *arXiv preprint arXiv:2502.13175*, 2025.

[35] Benfeng Xu, An Yang, Junyang Lin, Quang Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. ExpertPrompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*, 2023.

[36] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[37] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

[38] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 2024.

[39] Dan Zhang, Sining Zhoubian, Yisong Yue, Yuxiao Dong, and Jie Tang. ReST-MCTS*: LLM self-training via process reward guided tree search. *arXiv preprint arXiv:2406.03816*, 2024.

[40] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *NeurIPS*, 2023.