Metrics for Holistic Evaluation of LLM Reasoning about Action, Change, and Planning

Anil B Murthy
School of Computing & AI
Arizona State University
Tempe, AZ 85281
abmurthy@asu.edu

Jaron Mink*
School of Computing & AI
Arizona State University
Tempe, AZ 85281
jaron.mink@asu.edu

Lindsay Sanneman*
School of Computing & AI
Arizona State University
Tempe, AZ 85281
lindsay.sanneman@asu.edu

Abstract

Planning, reasoning, and sequential decision-making have played a pivotal role in the development of AI systems. While Large Language Models (LLMs) have demonstrated impressive capabilities, their evaluation for planning and Reasoning about Action and Change (RAC) problems is performed using strict binary success criteria, which limits information for further analysis and development. Given the probabilistic and autoregressive nature of LLMs, this work proposes the use of simple non-binary task-specific metrics for the evaluation of LLM responses for planning and reasoning tasks that go beyond perfect matching with ground truth, by utilizing set comparison methods, while still maintaining rigid and non-malleable evaluation criteria. We demonstrate the utility and usefulness of this type of metric in obtaining richer data fidelity and information about the quality, precision, nature of LLMs' responses, and their closeness to the ground truth through evaluations on six different tasks across two domains. With two case study examples, we additionally demonstrate the feasibility of comparative analysis of different task-specific data distributions obtained through this metric.

1 Introduction

The ability to plan, perform sequential decision-making, and reason about action and change is one of the fundamental tenets of human intelligence, and has been one of the cornerstones of AI. Today, modern generative AI and Large Language Models (LLMs) are useful for a plethora of applications, from question answering and document summarization to code generation [4]. Despite their impressive capabilities, LLMs have shown significant limitations in planning, reasoning, and decision-making, particularly in autonomous applications [8, 19, 7, 5]. Such limitations in LLMs' performance are noted through task evaluations that utilize binary success criteria metrics that involve comparison with ground truth answers obtained by automated solvers, planners, or validators. However, there exists useful information about the quality and precision of the models' responses for these task evaluations, which is not necessarily captured by standard binary metrics, that can help with comprehensive and domain/instance-specific diagnostic analyses, for developing real-world deployable agentic systems.

As LLMs are probabilistic models and generate tokens in an autoregressive manner, it is perhaps not surprising that they struggle to perform accurately on Reasoning about Action, Change (RAC), and planning problems. However, by considering intersection over union (IoU) metrics for task evaluations, we find a more nuanced picture of these models' task performance than is elicited by

*Equal Advising.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling.

standard binary success metrics. Specifically, our proposed metrics elicit more information about LLMs' task performance, related to precision and quality, that is missed when applying standard binary success criteria as overviewed in Figure 3. Having information about how close a model is to optimal or expected task performance can be extremely useful for failure analysis, causal analysis, and to make decisions about how best to utilize the model in architectural frameworks that are based on LLM-Modulo [7], ReAct [22], and other finetuning or prompting setups to enhance performance.

In the next section, we review benchmarks and related works that evaluate LLMs on Planning and RAC tasks, briefly detailing the tasks and metrics used. Then, we outline our evaluation domains, proposed metrics, and tasks. Finally, we discuss the results, utility, and usefulness of our metrics for RAC and Planning tasks through two examples.

2 Background & Related Works

2.1 Related Works

Recognizing the importance of benchmarking and evaluating the planning, decision-making, and reasoning abilities of LLMs, various benchmarks have been proposed in the literature [18, 5, 6, 8]. He et al. propose the Textual Reasoning about Action and Change (TRAC) benchmark, with 4 Reasoning about Action and Change (RAC) tasks such as projection, action executability, plan verification, and goal recognition, evaluated in the Planning Domain Definition Language (PDDL) based Blocksworld planning domain [6]. They pre-train and evaluate transformer models such as GPT-2 [13] on TRAC, and find that they struggle to generalize to scaling of objects, action sequence lengths, and composite tasks. The evaluations are conducted in a standard binary (true/false) manner and the overall accuracies are computed. However, it is unclear if the task design maintains structural validity (measurement reflecting the internal structure of the construct) [16].

Valmeekam et al. developed PlanBench, a PDDL-based planning benchmark suite with 8 planning-related tasks, such as plan generation, cost-optimal planning, plan verification, goal recognition, replanning, plan reuse, reasoning about actions and effects, and plan generalization [18]. the PlanBench work evaluates LLMs like GPT-4 [1] and Instruct-GPT-3 [12] on their generated plans across Blocksworld and Logistics domains, with a primary focus on variants of planning tasks and a limited focus on RAC tasks. The evaluations are performed based on the standard binary plan success criteria, as has been used in automated planning [15, 3].

Another notable benchmark is ActionReasoningBench, which evaluates multiple LLMs on RAC tasks such as state tracking, fluent tracking, action executability, and composite question combinations, on 8 different classical planning competition domains [2] like Blocksworld [5]. The evaluation is performed on binary and free-response answers of LLMs, for a few fixed sequence lengths of actions. However, it is important to note here that the free response questions were evaluated using a Llama-70B model in an LLM-as-a-judge framework in order to make the evaluation scalable, potentially leading to inaccurate reporting of performance statistics [20].

More recently, Kokel et al. proposed ACP Bench that consists of binary and multiple-choice questions on 7 different atomic reasoning and planning tasks, such as reasoning about applicable actions, atom reachability, action reachability, plan verification, progression, landmarks, and plan justification. They perform comprehensive evaluations on various LLMs on multiple classical planning domains, including the Alfworld household domain [17] and a novel 'swap' planning domain [8]. Following this work, Kokel et al. performs evaluations on the generative response version of this dataset, where task-specific evaluations use binary success metrics with perfect matching criteria against stored ground truth answers [9], which may lead to low or unclear construct validity [16].

2.2 Domains

To demonstrate the utility of our proposed benchmarks, we utilize standard IPC planning domains [2] such as Blocksworld and Depots for our experiments to evaluate the planning and action reasoning abilities of LLMs. For each of the 500 problems in the two domains, we create natural language templates for the initial and goal states, and questions for each of the 6 tasks, resulting in approximately 6000 questions that we use to evaluate the Llama 8B and Llama 70B models. For each problem, all the 6 task questions have the same object complexity, initial state, and goal state, only differing in the question prompt. A common natural language context containing the domain description, initial state description and goal state description (if necessary) is utilized for evaluating the LLMs, to ensure as holistic an evaluation as possible.

Blocksworld: Blocksworld is a domain where blocks can be placed on top of each other or on the table. There is one robotic arm that can move the blocks. The goal is to rearrange the blocks from an initial configuration to a goal configuration. This can be challenging as there may be interactions between subgoals. For our evaluation, we design a challenging dataset of 500 problems with 3-12 blocks, that have non-neutral initial states (A subset of blocks are in a stack, and the problems require unstacking and re-stacking), with an average optimal plan length of 18.7 actions.

Depots: The Depots domain is a combination of the blocksworld and logistics domains. In this domain, trucks can transport crates between places, the crates can be stacked onto pallets using hoists, and crates can be loaded into and unloaded from trucks using hoists. This domain inherits the challenges of subgoal interactions from Blocksworld, and reasoning about unreachable actions and states from Logistics. In this domain, we maintain the same object complexity (18) across all problems of the dataset, with an average optimal plan length 12 actions.

3 Tasks: Reasoning about Action, Change, and Planning

Drawing from the above benchmarks in Section 2, we select a set of key atomic tasks, such as action applicability, state tracking, progression of effects, and optimal plan generation, along with a new atomic task called State Comprehension (each task is detailed below). We focus on evaluating LLMs on free-response answers to task questions, instead of multiple-choice and binary responses, in order to obtain better construct validity and avoid construct confounds [14, 16].

Additionally, we formulate a simple non-binary task-specific metric for evaluation of RAC and planning tasks: we compute the Intersection over Union (IoU) of LLM answers and ground truth answers as shown in equation 1, resulting in task-specific metrics as shown in Table 1. Unlike binary evaluation metrics that have a success/ failure criterion based on perfect matching with ground truth answers, this kind of 'set comparison'-based metric allows us to obtain more fine-grained information about the quality of LLMs' performance for each task.

$$\text{Task Metric} = \frac{\text{LLM Answers} \cap \text{Ground Truth Answers}}{\text{LLM Answers} \cup \text{Ground Truth Answers}} \tag{1}$$

The tasks and their corresponding evaluation methods are detailed as follows:

3.1 Action Applicability

One of the fundamental atomic RAC tasks is the ability to reason about applicable actions at a given state. Previous works have shown that LLMs fall short of this ability and tend to provide invalid or hallucinated actions [21, 8, 5]. For actions to be valid in a given state, specific preconditions required by those actions must hold. We evaluate the generative free responses of LLMs by asking the LLM to list the applicable actions in a given state, provided the common context, as mentioned in the Domains section above, using the IoU evaluation metric shown in equation 1 and table 1.

3.2 State Comprehension

A fundamental requirement of reasoning about actions, change, and planning is to simply understand the given state, such as all the objects, predicates associated with their properties, and the environment properties. It is impossible to accurately perform any higher-level reasoning task, such as state tracking, action applicability, or planning, without fully understanding the properties of the current state. Thus, this task is simply about understanding the given state, including all the objects present, their properties, and the environment properties. Thus, this task requires the LLM to provide all the predicates associated with a given state, given the common context of domain and initial state descriptions.

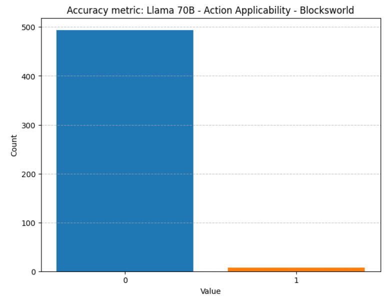

Figure 1: Llama 70B Performance with Standard binary success metric on Action Applicability in Blocksworld; Accuracy = 0.014%; Model's Responses are correct on only 7/501 problems.

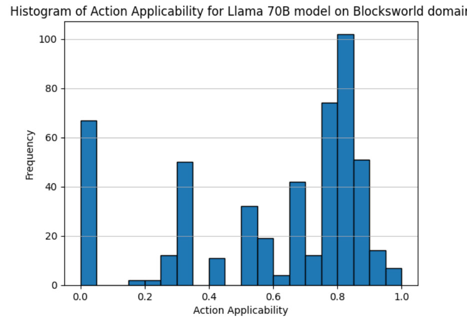

Figure 2: Llama 70B Performance with IoU metric on Action Applicability Task in Blocksworld; This right-skewed distribution provides information on the precision of the model's responses. We can see that the model is close to correctness on around $200/501$ problems.

Figure 3: Comparison of IoU Metric vs Standard Binary Success metric. We get a lot more data fidelity and information about precision and quality of responses from the IoU metric compared to the binary success metric.

3.3 Progression

This task evaluates the LLMs' ability to understand the effects of an action on the state. Keeping track of effects and changes through multiple states and action sequences is an important aspect of sequential decision-making and planning. LLMs have been shown to struggle with tracking changes across sequences of actions and states [5, 8, 19]. Also, prior works have found that LLMs' performance differs with positive and negative predicates [5]. We design two atomic tasks asking the LLM for the positive and negative effects of a single action, respectively, given the common context of domain and initial state descriptions and the specified action.

Positive effects are those that are not true in the current state and become true in the following state after the action is performed. These are also called add effects. Identifying positive effects is important as emerging effects can be preconditions to future actions along a plan.

Negative effects are those that are true in the current state and become false in the following state after the action is performed. These are also called delete effects. Identifying negative effects is extremely important to avoid dead loops, inconsistent states, and invalid actions.

3.4 State Tracking

State tracking is the ability to track entire states across multiple time steps after executing a sequence of actions. State tracking is a fundamental ability required for planning, as it involves generating valid successor states and actions at every visited state. Similar to Handa et al.'s ActionReasoningBench [5], we design an atomic version of this task by asking LLMs to provide the complete set of predicates that represent the final state after performing an action or a sequence of actions. In this work, we perform evaluation for a sequence of 2 actions, and prompt the LLM for the predicates of the final state, with domain and initial-state descriptions as context. The evaluation is performed in the same manner as State Comprehension, using the IoU metric in Table 1.

3.5 Plan Generation

Plan generation is a classical planning task where the task is to provide a valid sequence of actions that can be executed consecutively from a given state to reach the goal state. Given the domain description, initial state, and goal state, this task asks the LLM to provide a sequence of actions that constitute a plan to reach the goal state from the initial state. We prompt the LLMs to generate plans given the domain, state, and goal contexts. Evaluation is performed using the well-known set

Table 1: IoU Task Evaluation Metrics Summary. (GT: Ground Truth)

Task	Resulting Evaluated Formula
Action Applicability	$\frac{\text{\# Correct LLM Answered Actions}}{\text{\# LLM Answered Actions} \cup \text{\# GT Applicable Actions}}$
State Comprehension	$\frac{\text{\# Correct LLM Answered Predicates}}{\text{Total LLM Answered Predicates} \cup \text{GT Predicates}}$
Progression (Positive/ Negative)	$\frac{\text{\# Correct LLM Answered Effects}}{\text{Total LLM Answered Effects} \cup \text{GT Effects}}$
State Tracking	$\frac{\text{\# Correct LLM Answered Predicates}}{\text{Total LLM Answered Predicates} \cup \text{GT Predicates}}$
Plan Generation & Cost-Optimal Plan Generation	$1 - \frac{\text{\# Overlapping Unique Actions}}{\text{All Unique LLM Actions} \cup \text{Unique Actions from GT Plan}}$

comparison metric called 'Action Distance' [11], as shown in Table 1. As there may be multiple possible satisficing plans from the initial state to reach the goal state, we store only the optimal plan as the ground truth reference for evaluation with the action distance metric.

Evaluation with the Action Distance Metric Unlike for previous tasks, there are already various proposed metrics in the planning literature to measure plan quality, such as Action Distance, Causal-Link Distance, and State Sequence Distance [11, 10]. These metrics have been used to measure the quality of plans compared to an optimal plan. As LLMs are probabilistic models and fare poorly at generating valid plans [7], utilizing such metrics can shed some light on their performance at generating plans that would not be available with perfect accuracy measures. Hence, we utilize the action distance metric for our evaluation. However, it is important to note that action distance is a set comparison metric between unique action sets and does not account for the ordering of actions. Also, unlike the IoU metrics for other tasks, the action distance metric has an additive inverse with respect to 1. This means that an action distance of 1 represents that the model's plan has an entirely different set of actions compared to the ground truth reference plan. And an action distance of 0 represents that the model's plan has the same set of actions as the ground truth reference plan. However, as the action distance metric does not account for ordering of actions, a plan with action distance 0 may still be invalid and incorrect. This can be construed as "the plan has all the right actions, but not in the right order". From this perspective, the action distance metric can be useful to identify how far off generative AI models are at generating the correct set of actions.

For the plan generation task, although there may be numerous satisficing plans for a given pair of initial state and goal state, we evaluate the action distance metric with respect to an optimal plan as the reference. This provides us with information on the model's ability to choose landmark actions (actions that are part of all plans for a given initial state and goal state).

3.6 Cost-Optimal Plan Generation

If actions have costs, then an optimal plan is one that has the minimum cost. Unlike the other RAC tasks, the expected answer here is an ordered and optimal set of actions. This inherently implies a stricter evaluation criterion and, hence, is also more complex, as it requires coming up with optimal, goal-reaching actions, in addition to generating valid plans. Evaluation is performed similarly to plan generation using the action distance metric [11] with the optimal plan as the reference, which is also the ground truth for this task.

4 Results and Discussion

In this work, we perform evaluations with 6 tasks across two domains of 500 problems each, on two instruction-tuned pretrained LLMs, using informative task-specific IoU metrics. In Figure 2, we can see that the data distribution obtained through the IoU metric provides us with substantial information on the precision, quality, and nature of models' responses that are entirely missed by binary success metrics, as shown in Figure 1.

In Figure 2, the right-skewness of the distribution demonstrates that the model is much closer to being correct than the 0 values for 494 samples imply. In fact, the model's performance is over 75% accurate for more than 200 samples. This information is extremely beneficial for compute-intensive and cost-incurring decisions such as finetuning procedures, and for inference-time decisions such as model-routing, repeated sampling or prompting setups. Additionally, Figure 2 shows that over 70 instances have a low performance of $< 0.05\%$, indicating the need for instance-specific analysis of those samples. Further, this metric helps the design of future experiments to understand and improve specific atomic reasoning constructs or capabilities of generative AI models, such as reasoning about action applicability and state understanding.

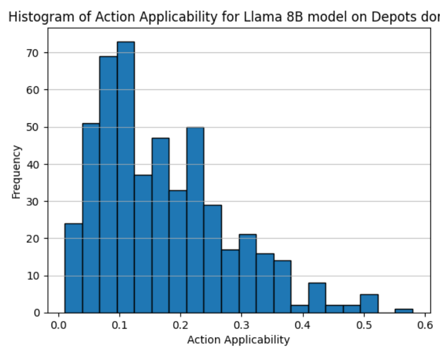

Figure 4: Llama 8B Performance with IoU metric on Action Applicability in Depots domain;

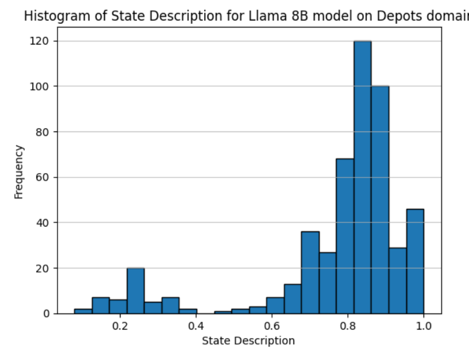

Figure 5: Llama 8B Performance with IoU metric on State Comprehension/ Description Task in Depots domain

Figure 6: Comparison of IoU Metric evaluation of Action Applicability and State Comprehension tasks. It is evident from the left-skewed distribution of Figure 4 and the right-skewed distribution of Figure 5 that Llama 8B model's responses and performance is more precise and of higher quality for state comprehension than for reasoning about applicable actions.

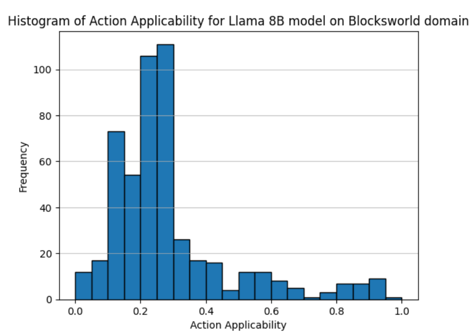

Figure 7: Llama 8B Performance with IoU metric on Action Applicability in Blocksworld domain;

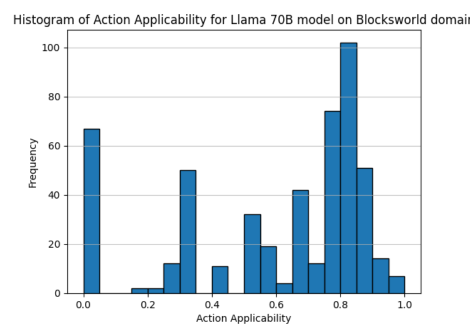

Figure 8: Llama 70B Performance with IoU metric on Action Applicability in Blocksworld domain

Figure 9: Comparison of IoU Metric evaluation of Llama 8B and 70B models on the Action Applicability Task. It is evident from the left-skewed distribution of Figure 7 and the right-skewed distribution of Figure 8 that Llama 70B model's responses and performance is more precise and of higher quality than those of the Llama 8B Model.

In Figure 6, we compare the IoU metric performance graphs of action applicability and state comprehension tasks of Llama 8B model from the Depots domain. From the stark contrast in the skewness of the distributions, it is pretty clear that the quality and precision of the model's responses for state comprehension are much better than its ability for reasoning about applicable actions. Also, the spread of the distribution for the Action applicability task, according to figure 10, indicates that the model's responses are less precise and more fuzzy compared to those of State comprehension in the Depots domain. Thus, the IoU metric can potentially provide discriminant validity [16], where the evaluation helps differentiate between constructs that should be distinct. Essentially, this distributional comparison indicates that the model is better at understanding a given initial state than at reasoning about what actions can be applied in that state in the Depots domain.

Also, these distributions can be compared with those of State Tracking over 2 actions, shown in Figure 14, which has a slightly lesser height, but a more chaotic spread, which can provide information about the model's reasoning ability with reference to the domain-specific state properties. Comparing Figures 4 and 14, the model seems to be more precise at tracking changes across states than at reasoning about applicable actions in the current state. However, further case-based analysis is required to examine the action sequence and the corresponding affected objects in high-state-tracking performance samples, to investigate whether any particular domain dynamics lead to higher state-tracking performance. Using the state tracking IoU metric, we have found preliminary evidence of specific domain dynamics acutely affecting the variance in state tracking performance in both domains, particularly with odd and even numbered action sequence lengths.

Thus, the IoU metric is beneficial in reasoning and planning tasks, to obtain information on the precision, quality, nature of models' responses, and their closeness to ground truth, all of which are highly valuable for development decisions on finetuning and model utility in architectural frameworks. We have demonstrated the utility of the metric through evaluations and comparative examples across two domains. A more in-depth correlational analysis across tasks, domain-specific and task-specific investigations that are beyond the scope of this project is left for future work.

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] IPC competition. ICAPS International Planning Competition (IPC). `http://www.icaps-conference.org/competitions/`, 2024. Accessed: 09/2025.

[3] Malik Ghallab, Dana Nau, and Paolo Traverso. *Acting, Planning, and Learning*. Cambridge University Press, 2025.

[4] D. Hagos, Rick Battle, and Danda B. Rawat. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*, 5:5873–5893, 2024. URL `https://api.semanticscholar.org/CorpusId:271329267`.

[5] Divij Handa, Pavel Dolin, Shrinidhi Kumbhar, Tran Cao Son, and Chitta Baral. Actionreasoningbench: Reasoning about actions with and without ramification constraints. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=NUD03NBDOE`.

[6] Weinan He, Canming Huang, Zhanhao Xiao, and Yongmei Liu. Trac: A textual benchmark for reasoning about actions and change. *arXiv preprint arXiv:2211.13930*, 2022.

[7] Subbarao Kambhampati, Karthik Valmeekam, Lin Guan, Mudit Verma, Kaya Stechly, Siddhant Bhambri, Lucas Saldyt, and Anil Murthy. Position: Llms can't plan, but can help planning in llm-modulo frameworks. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[8] Harsha Kokel, Michael Katz, Kavitha Srinivas, and Shirin Sohrabi. Acpbench: Reasoning about action, change, and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26559–26568, 2025.

[9] Harsha Kokel, Michael Katz, Kavitha Srinivas, and Shirin Sohrabi. Acpbench hard: Unrestrained reasoning about action, change, and planning. In *AAAI 2025 Workshop LM4Plan*, 2025. URL `https://openreview.net/forum?id=cfsVixNuJw`.

[10] Anagha Kulkarni, Yantian Zha, Tathagata Chakraborti, Satya Gautam Vadlamudi, Yu Zhang, and Subbarao Kambhampati. Explicablility as minimizing distance from expected behavior. *arXiv preprint arXiv:1611.05497*, 2016.

[11] Tuan Anh Nguyen, Minh Do, Alfonso Emilio Gerevini, Ivan Serina, Biplav Srivastava, and Subbarao Kambhampati. Generating diverse plans to handle unknown and partially known user preferences. *Artificial Intelligence*, 190:1–31, 2012. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.05.005. URL `https://www.sciencedirect.com/science/article/pii/S0004370212000707`.

[12] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[13] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[14] Anka Reuel-Lamparth, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. Betterbench: Assessing ai benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813, 2024.

[15] Stuart Russell, Peter Norvig, and Artificial Intelligence. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25(27):79–80, 1995.

[16] Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for ai evaluation. *arXiv preprint arXiv:2505.10573*, 2025.

[17] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL `https://arxiv.org/abs/2010.03768`.

[18] Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Planbench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36:38975–38987, 2023.

[19] Karthik Valmeekam, Sarath Sreedharan, Matthew Marquez, Alberto Olmo, and Subbarao Kambhampati. On the planning abilities of large language models (a critical investigation with a proposed benchmark). *arXiv preprint arXiv:2302.06706*, 2023.

[20] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *ArXiv*, abs/2305.17926, 2023. URL `https://api.semanticscholar.org/CorpusID:258960339`.

[21] Jian Xie, Kai Zhang, Jiangjie Chen, Tinghui Zhu, Renze Lou, Yuandong Tian, Yanghua Xiao, and Yu Su. Travelplanner: A benchmark for real-world planning with language agents. In *Forty-first International Conference on Machine Learning*, 2024.

[22] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

A Tasks Performance Graphs for IoU metric on Depots Domain

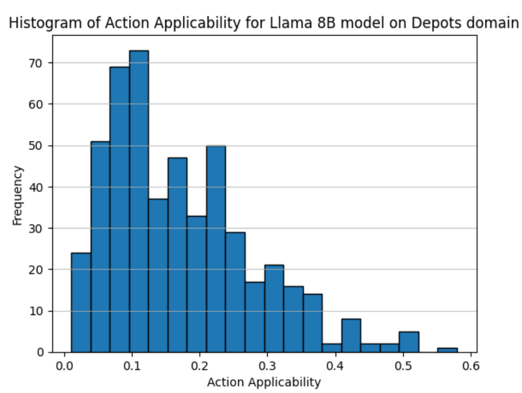

Figure 10: Llama 8B Performance on Action Applicability in Depots Domain

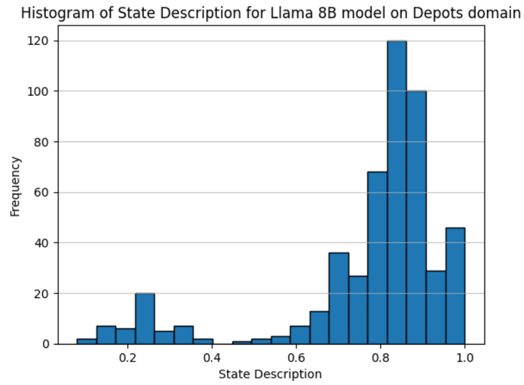

Figure 11: Llama 8B Performance on State Comprehension in Depots Domain

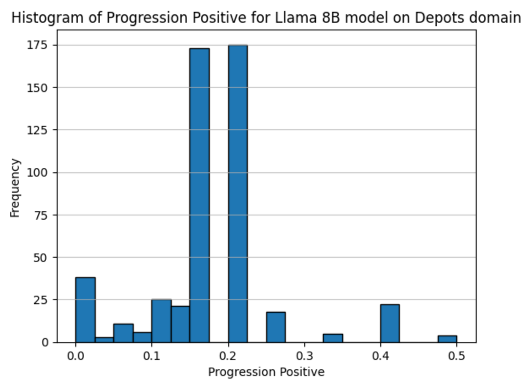

Figure 12: Llama 8B Performance on Identifying Positive Effects of Action progression in Depots Domain

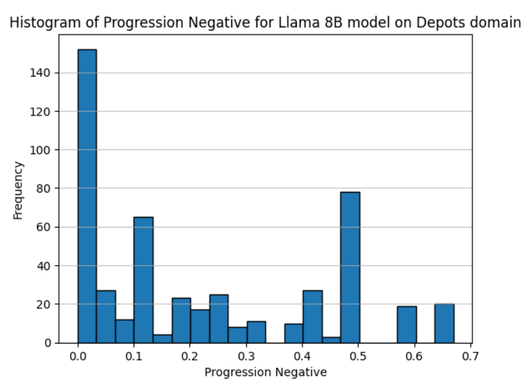

Figure 13: Llama 8B Performance on Identifying Negative Effects of Action Progression in Depots Domain

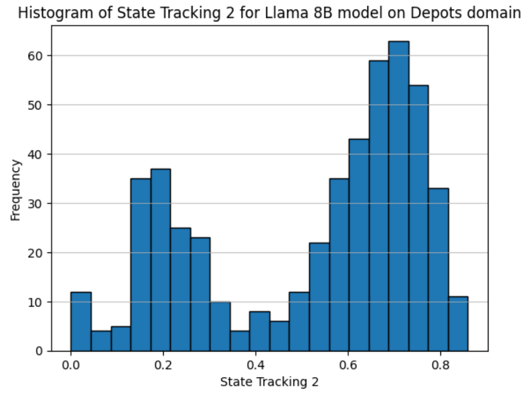

Figure 14: Llama 8B Performance on State tracking with 2 Actions in Depots Domain

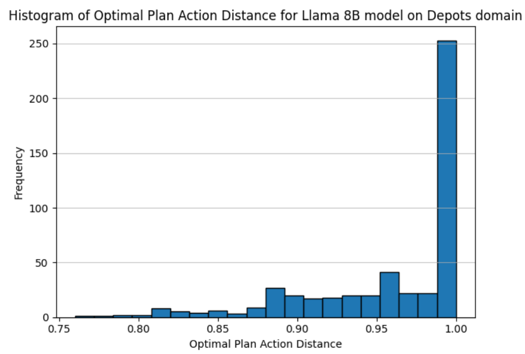

Figure 15: Llama 8B Optimal Plan Responses' Action Distance Histogram

B Tasks Performance Graphs for IoU metric on Blocksworld Domain

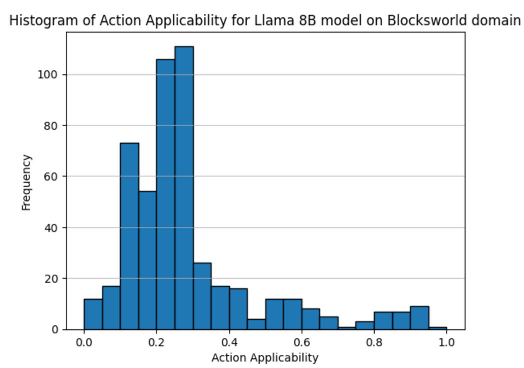

Figure 16: Llama 8B Action Applicability Histogram on Blocksworld Domain

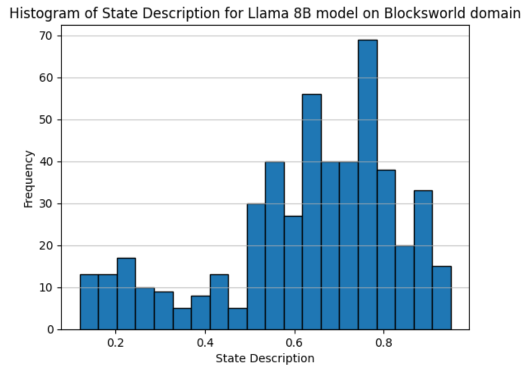

Figure 17: Llama 8B State Comprehension Histogram on Blocksworld domain

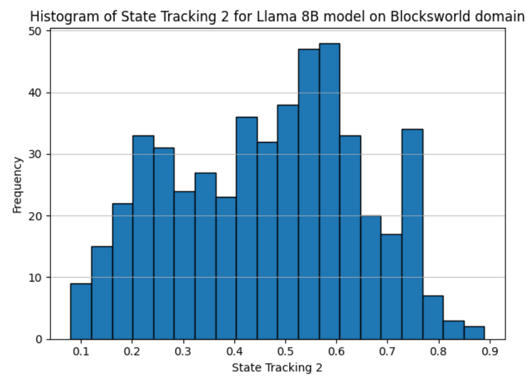

Figure 18: Llama 8B Performance Histogram for State tracking with 2 actions

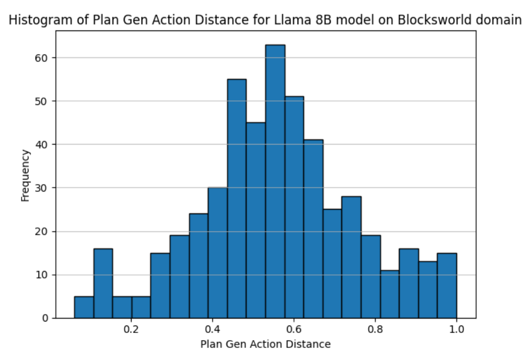

Figure 19: Llama 8B Plan Generation Action Distance Histogram

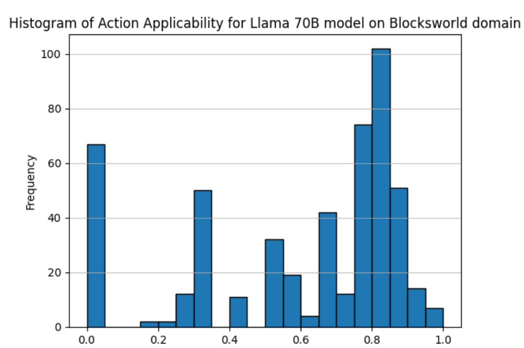

Figure 20: Llama 70B Action Applicability Histogram

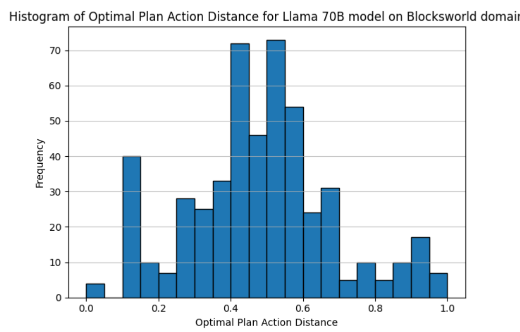

Figure 21: Llama 70B Optimal Plan Action Distance Histogram