2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW) | 979-8-3503-7981-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICMEW63481.2024.10645429

LLM-SAP: LARGE LANGUAGE MODELS SITUATIONAL AWARENESS-BASED PLANNING

Liman Wang, Hanyang Zhong**

University of York;
ssee02131@gmail.com, hanyang.zhong@york.ac.uk

ABSTRACT

This study explores the integration of large language models (LLMs) with situational awareness-based planning (SAP) to enhance the decision-making capabilities of AI agents in dynamic and uncertain environments. By employing a multi-agent reasoning framework, we develop a methodology that not only anticipates but actively mitigates potential risks through iterative feedback and evaluation processes. Our approach diverges from traditional automata theory by incorporating the complexity of human-centric interactions into the planning process, thereby expanding the planning scope of LLMs beyond structured and predictable scenarios. The results demonstrate significant improvements in the models' ability to provide comparative safe actions within hazard interactions, offering a perspective on proactive and reactive planning strategies. This research highlights the potential of LLMs to perform human-like action planning, thereby paving the way for more sophisticated, reliable, and safe AI systems in unpredictable real-world applications.

1. INTRODUCTION

Developing AI agents capable of flexible decisions is challenging due to real-world unpredictability [1]. Humans manage these uncertainties with situational awareness, whose lack is a major cause of accidents from human errors [2, 3]. Yadav [4] emphasizes that understanding situational awareness in LLMs is crucial for their safe development. Without this understanding, seemingly beneficial actions can have unintended consequences [5, 6]. For instance, an autonomous agent needs nuanced judgments to prevent harm, such as when a toddler reaches for a hot pot or plays with a knife.

The application of LLMs for SAP introduces a significant paradigm shift due to the inherently infinite state space of the open world, which is in stark contrast to the relatively confined state spaces observed in traditional game strategies [7, 8]. The advantage lies in LLMs' ability to describe details and dynamically interact within contexts using natural language, allowing for an extensive expansion of state descriptions. This feature proves particularly apt for enabling intelligent agents to achieve a comprehensive understanding

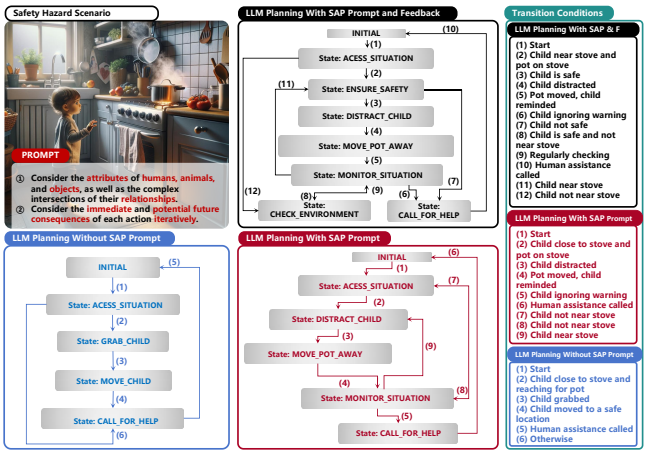

Fig. 1. Large language models' planning **enhancements** based on situational awareness.

of the real environment's interactive state, thereby enhancing their capability to detect potential hazards, predict forthcoming conditions, and formulate multi-step strategic actions. Prior investigations on the applications of LLMs within the scope of computational models like finite state machines and behaviour trees, specifically targeting certain tasks or subtasks completion [9, 10, 11]. These applications typically exploit the reasoning capabilities of LLMs for the design of state space transitions and programming [9]. However, these studies have seldom extended into open-world settings, where the unpredictability and intricacy of real-world interactions pose significant challenges. Contemporary research in embodied intelligence often establishes "Rules" that prohibit interactions with humans or animals, the handling of sharp objects, and involvement in dangerous settings such as those with water and electricity, aiming to define research boundaries and simplify complex issues, akin to the approach seen in the Google RT-X Series [12, 13, 14]. Nonetheless, the simulations for robots and agents [15, 16, 17, 18] ignore real-world hazardous interactions that are prevalent and essential to consider in daily scenarios.

This research demonstrates that LLMs can display human-like planning capacities rooted in situational awareness, as illustrated in Fig. 1. When prompted for plan-

* Co-first author

979-8-3503-7981-5/24/$31.00 ©2024 IEEE
Authorized licensed use limited to: Dhirajlal Gandhi College of Technology. Downloaded on October 20,2025 at 01:18:53 UTC from IEEE Xplore. Restrictions apply.

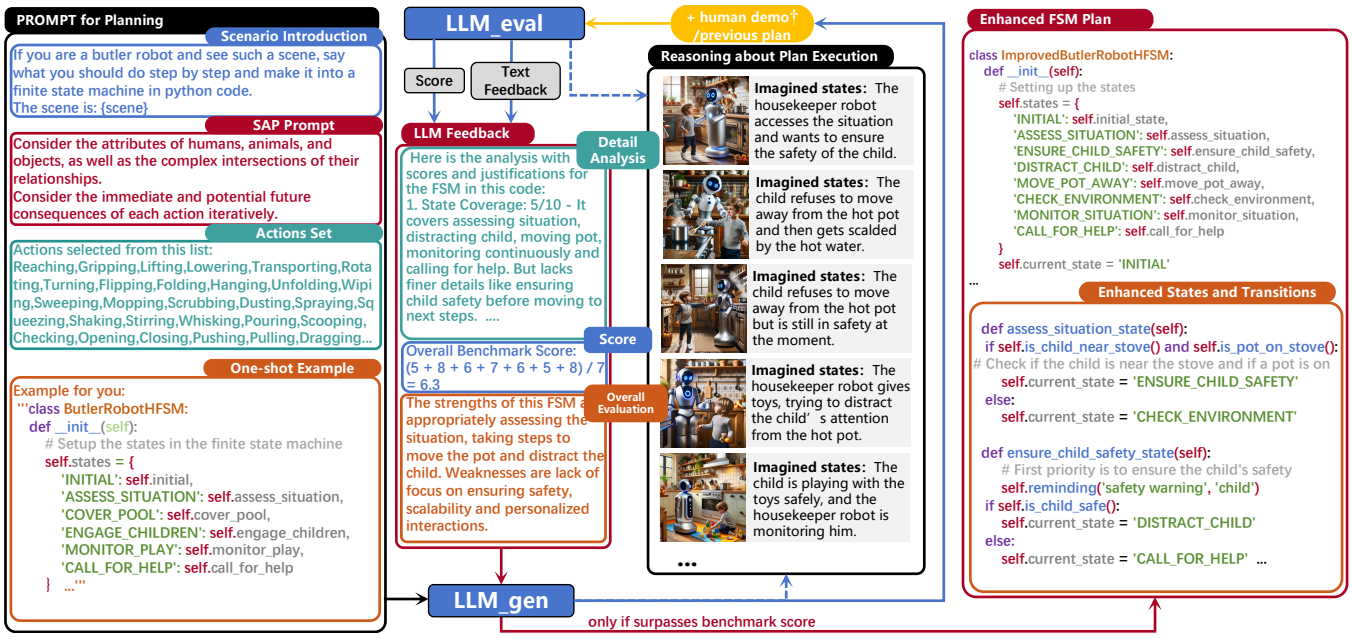

Fig. 2. Iteratively generating and evaluating plans in a **multi-agent proactive AI system** produces iterative feedback to enhance reasoning and factual accuracy. † refers to selecting the human demo for comparison only in the first round of iteration.

ning that involves perception, comprehension, and projection [2], or when provided with reasoning feedback, the LLM shows enhanced deductive reasoning skills that require perspective-taking and consideration of potential outcomes. This study stands apart from previous research in AI reasoning and planning in several significant ways. Most existing studies focus on task-specific planning or rely on assumptions about predefined steps under controlled conditions [19, 12, 13, 20, 18, 21]. Additionally, conventional agents typically generate plans reactively, only in response to explicit instructions or demands [22, 15, 23, 24, 16, 25, 26, 27], rather than proactively. In contrast, our approach focuses on evaluating and enhancing models' abilities for proactive situational planning in the face of real, open-world challenges. Without strict constraints or direct environmental feedback, the models must employ deductive reasoning to envisage actions while considering their consequences, relying solely on a descriptive initial scenario and prompts.

2. METHODOLOGY

In this section, we outline the key challenges and methodological components that enable collaborative multi-agent reasoning to enhance LLMs' situational planning capacities.

2.1. Task Formulation and Key Challenges

We define SAP as grounded inference over a dynamic hazard scenario s, with $s \in S$ where S denotes the space of hazard situations. The input consists of an unordered set of concepts $x = \{c1, c2, ..., ck\} \subseteq C$, capturing entities, events, and the temporal dynamics within s, with C representing the overall conceptual vocabulary. The output is a step-wise plan $\pi : S \rightarrow A$, consisting of actions $a1, a2, ... \in A$, where A denotes the action space of possible interventions. Learning the planning policy $\pi : S \rightarrow A$ requires overcoming two intrinsic challenges:

1. Inadequate attainment of high levels of situational awareness, marked by a deficiency in perceiving, understanding, and anticipating the dynamics of the hazard environment when determining suitable state-action mappings.

2. Difficulty in foreseeing the potential downstream impacts of planned actions on human safety and property, stemming from a limited grasp of the dynamics within hazardous situations.

By formulating hazard remediation as a conceptual planning task requiring strong situational awareness, we evaluate the multidimensional latent reasoning essential for reliable situated agents operating in hazardous environments.

2.2. Multi-AI Agents Enhance Reasoning and Accuracy

Recent work has shown that employing multiple LLMs within a cooperative framework, either collaborative or adversarial, can enhance reasoning and factual accuracy. As highlighted by Du et al., the interaction between agents, allowing them

Authorized licensed use limited to: Dhurakij Pundit University. Downloaded on October 20,2025 at 01:18:53 UTC from IEEE Xplore. Restrictions apply.

to critique and refine each other's reasoning, helps in correcting logical flaws [28]. Similarly, Liang et al. have observed that disagreement among agents fosters broader reasoning, as each strives to surpass the others [29]. Such collaboration leverages the individual strengths of each agent [28, 29, 30]. In this work, we employ two LLM agents - LLM$_{gen}$ for plan generation and LLM$_{eval}$ for critical evaluation. We rely on the synergistic interaction between these complementary roles to enhance latent planning capabilities.

2.3. State-based Planning with Feedback

Current AI systems that operate on rigid, context-insensitive rules are at risk of producing unintended outcomes when deployed in complex, real-world environments [19, 12, 13, 20, 18, 21, 22, 15, 23, 24, 16, 25, 26, 27]. To enable more reliable and ethical decision-making, it is crucial for architectures to model interdependent variables and causal relationships in a manner akin to human reasoning processes.

Algorithm 1 Situational Awareness-Based Planning

$\hat{M} \leftarrow R_{LLM_{gen}}(S, T, A)$
$score, f \leftarrow R_{LLM_{eval}}(\hat{M})$
while $\hat{M}_{score} < M^*_{score}$ **do**
  $\hat{M} \leftarrow R_{LLM_{gen}}(S, T, A, f)$
  $score, f \leftarrow R_{LLM_{eval}}(\hat{M})$
end while
$M^* \leftarrow \hat{M}$
adopt M^*

One approach involves LLM agents collaboratively iterating through the generation and evaluation of potential solutions before their actual implementation. For example, we conceptualize the design of a finite state machine (FSM) [31] as a collaborative process between two models. A latent FSM plan can be defined by a tuple $M = (S, T, A)$, including a set of states S, transitions T, and actions A. The process begins by representing the plan's reasoning process as R, with a generator agent reasoning ($R_{LLM_{gen}}$) to plan then an evaluator agent reasoning ($R_{LLM_{eval}}$) to evaluate. $R_{LLM_{gen}}$ proposes a candidate FSM plan $\hat{M}$, which $R_{LLM_{eval}}$ then scores and provides feedback f on. $R_{LLM_{gen}}$ incorporates this feedback into the next proposal. This iterative loop continues until the score of $\hat{M}$ is higher than that of M^*(the benchmark plan), which is finally adopted as the new optimal plan M^*. This approach, detailed in Algorithm 1, facilitates tight refinement loops that mirror human reasoning. By evaluating solutions prior to their real-world deployment, it is possible to foresee and mitigate unintended consequences.

Fig. 2 visually depicts the iterative process between LLM$_{gen}$ and LLM$_{eval}$. Consider a scenario where a young child attempts to touch a hot pot on an active stove, creating a safety risk. The housekeeper robot observing this scene starts planning appropriate interventions. First, the instruction-prompted LLM$_{gen}$ uses its reasoning capabilities to envision possible outcomes and formulate candidate FSM plans to prevent harm. For instance, abruptly stopping the child could cause fright, indicating that a gentler approach or distraction with toys might be more effective. If the child disregards these precautions and sustains a burn, emergency actions may be required. The model LLMgen submits its proposed plan $\hat{M}$, along with a comparative plan, which initially includes a human demonstration and then the previous plan, to LLM$eval$ for evaluative scoring and feedback f. LLM$_{gen}$ then integrates this feedback f into its future planning. After one or more rounds of proposal and evaluation, the agents refine their approach until the new optimal FSM plan's score, M^*, exceeds that of the benchmark plan, enabling it to handle edge cases morally and robustly through situational inference.

By enabling the models to engage in proactive deductive reasoning before deployment in the real world, potential unintended consequences can be anticipated and mitigated. As the capabilities of LLMs advance, such methodologies show promise for enhancing reliability and ethical standards in AI systems designed for physical-world interactions.

2.4. Formation of Prompts

As depicted in Fig. 2, the prompts provided to the generative model (LLM$_{gen}$) contain the scene description, a SAP prompt, a list of actions, and an exemplar plan. The SAP prompt is designed to elicit sophisticated reasoning by encouraging the model to thoroughly consider the varied needs and potential interactions among people, animals, and objects. By explicitly prompting the model to infer the needs of other entities and to anticipate how situations might evolve dynamically, the prompt fosters empathy and holistic thinking, which are essential for devising comprehensive plans. The one-shot exemplar illustrates the desired plan structure in code format without providing solutions specific to the evaluation scenario. In contrast, the prompt for the evaluative model (LLM$_{eval}$) contains a generated FSM plan from LLM$_{gen}$, a benchmark high-quality plan, and descriptions of the scoring criteria to evaluate the quality of the plan through iterative refinements (see Appendix Fig.17). Initially, benchmark plans consist of manually authored solutions, but in subsequent iterations, they incorporate the highest-scoring autogenerated plan from the previous round.

3. EXPERIMENTS

To systematically assess LLM planning capacities, standardized benchmark scenarios are developed along with quantitative scoring methodologies.

3.1. Evaluation Scenarios

The dataset comprises over 500 hazardous home scenarios, specifically curated to fill gaps often ignored in academic

Authorized licensed use limited to: Dhurakij Pundit University. Downloaded on October 20,2025 at 01:18:53 UTC from IEEE Xplore. Restrictions apply.

research, such as scenarios typically avoided by embodied agents and agent simulations [14, 32, 12, 13, 15, 16, 17]. This collection is aimed at scenarios that home assistance robots are likely to encounter, including emergencies involving diverse human demographics, interactions with pets, and dangerous situations involving sharp objects, water, electricity, and open flames. In light of the constraints imposed by image generation models such as DALL·E [33], the production of images depicting hazardous scenarios is limited. Consequently, the acquisition of pertinent imagery through internet crawlers is employed to uphold precision in depicting these perilous situations. From this comprehensive dataset, 24 vignettes are methodically selected across four complexity levels for detailed analysis. Textual descriptions generated by GPT-4V [34] and expert-validated solutions provide a robust framework for evaluating the planning capabilities of LLMs. This evaluation leverages the image-to-text capability of GPT-4V to standardize inputs across models, focusing on planning skills over visual data interpretation, thereby ensuring fairness in assessment. Future studies will assess the effectiveness of end-to-end vision language models (VLMs), aiming to streamline the transition from perception to planning.

3.2. Actions Set

This study aims to quantify the complex planning abilities of LLMs. To ensure fairness and consistency in subsequent evaluations, we have imposed certain limitations on the action set available to AI agents. This action set includes 56 distinct robot behaviours commonly employed in domestic settings, as exemplified by representative actions displayed in the action enumeration diagram located to the left of the central region in Fig. 2 (for more details, see Appendix A.1). This selection provides a thoughtful baseline for functionality, drawing on insights from some of the leading projects in intelligent robotics [19, 12, 13, 35, 20, 36].

3.3. Evaluation Dimensions

As detailed in Appendix B Table 6, seven scoring dimensions have been established to provide a comprehensive methodology for assessing latent planning and FSM designs [37, 38, 39]. These dimensions encompass coverage, complexity, safety, reusability, user experience, and coherence, allowing for the evaluation of structured completeness, validation requirements, real-world reliability, adaptability, human factors, and solution integrity. Utilizing these dimensions collectively fosters the creation of designs that are robust, dependable, future-proof, ethical, and aligned with specifications. These dimensions also provide multi-faceted technical and operational insights. Scoring FSMs across these seven key dimensions on a scale from 0 to 10 enables an impartial quantitative evaluation of the overall plan quality and highlights relative strengths and weaknesses to guide further refinements. The overall score is determined by calculating the average of the scores across these seven dimensions.

3.4. Evaluation Metrics

Motivated by discussions of inconsistent human evaluation in Iskender et al. [40] and the inadequate quality of automatic metrics highlighted in Sottano et al. [41], we introduce a rank-based scoring (RBS) method to help mitigate potential reliability issues when evaluating FSM plans. This aims to increase consistency compared to absolute scoring methods prone to rater variability.

The RBS score provides an objective aggregation by comparing models pair-wise on each evaluation scenario and assigning differential rankings based on relative performance. This eliminates variability from subjective absolute scoring. The head-to-head comparisons also allow powerful models like GPT-4 to participate in the evaluation. Rather than requiring predefined output standards, GPT-4 can provide comparative judgments on model outputs. Given two model sets $M = M_1, M_2$ evaluated on N scenarios with D scoring dimensions, models were compared pairwise for each scenario i. Scores s_{ijl} were assigned from 0-10 across dimensions j for each model l. Models were ranked $r_{ik} \in 1, 2$ per scenario based on total score:

$$r_{ik} = \arg\min_{l \in 1,2} \sum_{j=1}^{D} s_{ijl}$$

The higher scoring model was assigned rank 1 (1 point). The lower scoring model was assigned rank 2 (2 points). If the two models had equal total scores for a scenario, both were assigned a mid-point rank of 1.5 (1.5 points). After evaluating all scenarios, the ranking scores were aggregated to produce an RBS score R_k per model:

$$R_k = \frac{1}{N} \sum_{i=1}^{N} r_{ik}$$

The RBS score reflects relative performance, with scores closer to 1 indicating superior performance compared to the other model. By focusing on comparative judgments between model outputs rather than absolute scores, the RBS methodology aims to offer a more dependable means of evaluating text that necessitates subjective human judgment. Additionally, this comparative framework facilitates the inclusion of evaluative models such as GPT-4.

4. RESULTS

To systematically evaluate LLMs' planning capacities, we conduct experiments assessing model performance on a standardized benchmark of 24 home hazard scenarios across four reasoning complexity levels.

Authorized licensed use limited to: Dhurakij Pundit University. Downloaded on October 20,2025 at 01:18:53 UTC from IEEE Xplore. Restrictions apply.

4.1. LLM Selection

This experiment tests commercial models GPT-4, GPT-3.5 [34], Claude-2 [42], alongside open-source alternatives such as LLama-2 [43], LLava [44], Vicuna [45], MiniGPT-4 [46], and CodeLLama [47], on their ability to perform hazard planning using scene-informed one-shot prompts. The analysis reveals that many open-source models struggle to effectively utilize examples, with longer contexts leading to attention drift and diminished scene comprehension. In contrast, GPT-4, GPT-3.5, and Claude-2 demonstrate more reliable linkages between examples and planning tasks. Both quantitative and qualitative testing show that these commercial models maintain a stronger understanding despite the risk of drift. Consequently, GPT-4, GPT-3.5, and Claude-2 have been selected for further evaluation in hazard planning due to their superior grounding capabilities.

4.2. Impact of The SAP Prompt

An experiment is conducted to evaluate the effect of the SAP prompt on the quality of planning. As shown in Table B.1, three LLMs, GPT-4, GPT-3.5, and Claude-2, are tested with or without the SAP prompt on the benchmark scenarios across four complexity levels. The RBS methodology is employed, wherein models are compared pairwise for each scenario and differentially ranked. Introducing the SAP prompt leads to improved RBS scores for all three models compared to those not, indicating enhanced planning capabilities. Notably, GPT-4 with the SAP prompt achieves the highest overall RBS score of 1.21, significantly outperforming GPT-4 without prompts which has an RBS score of 2.04. An analysis of scenarios at reasoning level 3, which involve interactions with children, the elderly, and pets, shows that GPT-4 with the SAP prompt substantially exceeds the performance of the second-best model. This indicates that the prompt is particularly valuable in complex, nuanced planning situations that require perspective-taking and consideration of potential outcomes (for ablation studies, see Appendix B) [48, 49]. The findings suggest that prompts directing models to thoroughly contemplate relationships and iterative consequences significantly boost latent planning abilities. By fostering better coordination and foresight, these prompts improve deductions when reasoning about multi-agent safety hazards.

4.3. LLM Evaluators

Experiments assess the feasibility of using LLMs, like GPT-4 and Claude-2, to score FSM plans and compare their rankings with human evaluations, as shown in Appendix Table 9. The models are tested with ranking FSM plans in pairs, as well as in groups of 4 and 6. These rankings are then compared to expert human rank orders to measure accuracy. Tests find both GPT-4 and Claude-2 could rank FSM pairs with 75.7% agreement to human ranking, evidencing reliability for comparative evaluation. However, their accuracy significantly decreases when ranking groups of 4 or more FSMs. Table 10 and Fig.6 in Appendix B show that GPT-4 and Claude-2 align most closely with human judgment when evaluating outputs generated by the models themselves. For example, GPT-4's scoring of its own FSMs closely matches expert rankings. This demonstrates that LLMs can provide accurate comparative assessments, particularly for outputs from their own model family. The experiments highlight the potential of LLMs to serve as evaluators that mimic human appraisals of planning formalisms. By focusing on relative rather than absolute assessments, variability is minimized.

4.4. Multi-Agent Improvement

A closed-loop experiment quantifies planning improvements through iterative generation and evaluation between two agents. GPT-3.5 with the SAP prompt serves as the generative model (LLM_gen), and Claude-2 serves as the evaluative model (LLM_eval). LLM_gen initially proposes an FSM, which LLM_eval scores and provides feedback on. Using this feedback, LLM_gen produces an improved FSM in the next round. Testing shows the updated FSM surpasses the initial quality, achieving a higher RBS score after one iteration. Appendix Table 11 shows the closed-loop FSM outperforms GPT-4 with the SAP prompt, the previous top standalone performer. The feedback-improved output features more detailed planning, boosting RBS scores. This demonstrates how two weaker models can compensate for each other's shortcomings through collaboration. The results indicate interactive cycles between LLMs enhance reasoning and planning by leveraging their complementary strengths, surpassing individual model capabilities.

5. CONCLUSION

This study marks an advancement in the field of situational awareness-based planning using LLMs. New benchmarks, a specialized dataset, and multi-agent strategies have improved the planning capabilities of LLMs, better equipping them to handle complex and unpredictable human-centric scenarios. Looking ahead, further explorations will focus on expanding datasets and refining model architectures to speed up reasoning, with a particular emphasis on evaluating end-to-end VLMs to bridge the time-lag gap between simulated and real-time environments. This research underscores the importance of stimulating latent reasoning in LLMs and paves the way for ethically sound and safe AI planning processes in practical applications.

6. REFERENCES

[1] Francis et al., "Core challenges in embodied vision-language planning," 2022.

Authorized licensed use limited to: Dhurakij Pundit University. Downloaded on October 20,2025 at 01:18:53 UTC from IEEE Xplore. Restrictions apply.

[2] Mica R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors*, 1995.
[3] Wikipedia contributors, "Situation awareness — Wikipedia, the free encyclopedia," 2023.
[4] Dipendra Yadav, "Evaluating dangerous capabilities of large language models: An examination of situational awareness," 2023.
[5] Dario Amodei et al., "Concrete problems in ai safety," 2016.
[6] Stuart Russell, "Human compatible. ai and the problem of control, london: Allen lane," 2019.
[7] Mihai et al., "Beyond Chess and Go: Why AI Mastering Games could be good news for everyone - Heidelberg Laureate Forum - SciLogs - Wissenschaftsblogs," .
[8] Gao et al., "Large language models empowered agent-based modeling and simulation: A survey and perspectives," 2023.
[9] Cao et al., "Robot behavior-tree-based task generation with large language models," .
[10] Izzo et al., "Btgenbot: Behavior tree generation for robotic tasks with lightweight llms," 2024.
[11] Liu et al., "Smot: Think in state machine," 2023.
[12] Zitkovich et al., "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*, 2023, pp. 2165–2183.
[13] Padalkar et al., "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2023.
[14] Ahn et al., "Autort: Embodied foundation models for large scale orchestration of robotic agents," 2024.
[15] Song et al., "Llm-planner: Few-shot grounded planning for embodied agents with large language models," October 2023.
[16] Huang et al., "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," 2022.
[17] Huang et al., "Inner monologue: Embodied reasoning through planning with language models," .
[18] Singh et al., "Progprompt: Generating situated robot task plans using large language models," 2023.
[19] Ahn et al., "Do as i can, not as i say: Grounding language in robotic affordances," 2022.
[20] Driess et al., "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.
[21] Kant et al., "Housekeep: Tidying virtual households using commonsense reasoning," 2022.
[22] Ding et al., "Integrating action knowledge and llms for task planning and situation handling in open worlds," .
[23] Lykov et al., "Llm-brain: Ai-driven fast generation of robot behaviour tree based on large language model," .
[24] Wenlong Huang et al., "Inner monologue: Embodied reasoning through planning with language models," .
[25] Jain et al., "Transformers are adaptable task planners," 2023.
[26] Vemprala et al., "Chatgpt for robotics: Design principles and model abilities," Tech. Rep., Microsoft, February 2023.
[27] Wang et al., "Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents," .
[28] Yilun Du et al., "Improving factuality and reasoning in language models through multiagent debate," 2023.
[29] Tian Liang et al., "Encouraging divergent thinking in large language models through multi-agent debate," 2023.
[30] Guohao Li et al., "Camel: Communicative agents for "mind" exploration of large language model society," 2023.
[31] Wikipedia contributors, "Finite-state machine — Wikipedia, the free encyclopedia," 2023.
[32] Fang et al., "Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot," 2023.
[33] "Dall·E 2," https://openai.com/dall-e-2.
[34] OpenAI, :, and Josh Achiam et al., "Gpt-4 technical report," 2023.
[35] Yash Kant et al., "Housekeep: Tidying virtual households using commonsense reasoning," 2022.
[36] Wang et al., "Robogen: Towards unleashing infinite data for automated robot learning via generative simulation," 2023.
[37] Pradhan et al., "Coverage criteria for state-based testing: A systematic review," pp. 1–20, 01 2019.
[38] Wikipedia contributors, "Cyclomatic complexity — Wikipedia, the free encyclopedia," 2023.
[39] Wagner et al., *Modeling software with finite state machines: a practical approach*, 2006.
[40] Iskender et al., "Reliability of human evaluation for text summarization: Lessons learned and challenges ahead," 2021.
[41] Andrea Sottana et al., "Evaluation metrics in the era of gpt-4: Reliably evaluating large language models on sequence to sequence tasks," 2023.
[42] Anthropic, "Claude (version 2)," 2023.
[43] Hugo Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023.
[44] Liu et al., "Visual instruction tuning," in *NeurIPS*, 2023.
[45] Lianmin Zheng et al., "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023.
[46] Deyao Zhu et al., "Minigpt-4: Enhancing vision-language understanding with advanced large language models," 2023.
[47] Baptiste Rozière et al., "Code llama: Open foundation models for code," 2023.
[48] Takeshi Kojima et al., "Large language models are zero-shot reasoners," 2023.
[49] Cheng Li et al., "Large language models understand and can be enhanced by emotional stimuli," 2023.

Authorized licensed use limited to: Dhurakij Pundit University. Downloaded on October 20,2025 at 01:18:53 UTC from IEEE Xplore. Restrictions apply.