LLM-based Long-horizon Planning for Embodied Agents

LLM-based systems can effectively enable long-horizon planning for embodied agents, achieving success rates of 70-96% when implemented with hierarchical architectures that separate high-level reasoning from low-level control, robust environmental grounding through scene graphs or semantic maps, and iterative replanning mechanisms, though spatial reasoning limitations and dependence on predefined skill libraries currently constrain generalization to novel real-world scenarios.

Abstract

This systematic review of 10 studies (2022-2025) demonstrates that LLM-based approaches can effectively enable long-horizon planning for embodied agents, with reported success rates ranging from 70% to 96% across diverse tasks including household manipulation, multi-floor navigation, and multi-agent coordination . The evidence consistently shows that hierarchical and modular architectures—separating high-level LLM reasoning from low-level control through reinforcement learning , classical planners , or learned skill libraries —outperform end-to-end approaches. Critical success factors include robust grounding mechanisms such as 3D scene graphs or semantic maps , and iterative replanning to correct LLM errors during execution . Few-shot learning capabilities enable competitive performance with less than 0.5% of training data compared to fully-supervised methods , while multi-agent systems achieve over 40% efficiency gains through emergent LLM-driven communication .

However, significant limitations constrain real-world deployment. Spatial reasoning deficits were identified across multiple studies , and LLM hallucination remains problematic for plan validity . Systems typically depend on predefined skill libraries and pre-built environmental representations , limiting generalization to novel scenarios. Computational costs from repeated LLM queries present practical scalability concerns . The evidence suggests that neuro-symbolic approaches combining LLMs with classical planning achieve the highest reliability , and that grounding quality—rather than LLM capability alone—may be the primary determinant of real-world performance.

Paper search

We performed a semantic search using the query ”LLM-based Long-horizon Planning for Embodied Agents” across over 138 million academic papers from the Elicit search engine, which includes all of Semantic Scholar and OpenAlex.

We retrieved the 50 papers most relevant to the query.

Screening

We screened in sources based on their abstracts that met these criteria:

- **Embodied Agents**: Does the study involve embodied agents (physical robots, simulated agents, or autonomous systems) that can interact with and manipulate their environment?
- **LLM as Primary Planning Component**: Are Large Language Models (LLMs) used as a primary component of the planning system (not just for auxiliary tasks like natural language interfaces)?
- **Long-Horizon Planning Focus**: Does the study focus on long-horizon planning tasks that require multiple steps, sequential decision-making, or extended temporal reasoning?
- **Empirical Evaluation**: Does the study include empirical evaluation with quantitative or qualitative assessment of planning performance?
- **Beyond Disembodied Language Tasks**: Does the study go beyond solely disembodied language tasks to include physical or simulated agent interaction with environments?

- **LLM Integration Required**: Does the study integrate LLMs rather than using only traditional symbolic AI or non-LLM machine learning approaches?
- **Empirical Implementation**: Does the study include empirical validation or implementation rather than being purely theoretical?

We considered all screening questions together and made a holistic judgement about whether to screen in each paper.

Data extraction

We asked a large language model to extract each data column below from each paper. We gave the model the extraction instructions shown below for each column.

- **LLM Details**:

  Extract all information about the Large Language Models used including:

- Specific LLM model(s) (e.g., GPT-4, LLAMA-2, Claude)
- Model size/parameters if mentioned
- Whether the LLM was fine-tuned or used pre-trained
- Input/output modalities (text, vision, multimodal)
- Any specialized prompting strategies or prompt engineering techniques

- **Agent Architecture**:

  Describe the overall system architecture including:

- How the LLM is integrated into the planning system (e.g., hierarchical, modular, end-to-end)
- Key system components and their roles
- Whether multiple LLM agents are used and their specific functions
- Integration with other AI methods (reinforcement learning, classical planning, perception modules)
- Data flow and decision-making pipeline

- **Embodied Platform**:

  Extract details about the physical or simulated agents:

- Type of embodied agent(s) (humanoid, quadruped, mobile manipulator, drone, etc.)
- Number of agents (single vs multi-agent)
- Physical capabilities (locomotion, manipulation, sensing)
- Hardware specifications if mentioned
- Simulation platform used (if applicable)

- **Grounding Mechanisms**:

  Identify how the LLM connects to the physical world:

- Perception systems (cameras, sensors, 3D scene graphs, etc.)
- World representation methods (symbolic, geometric, semantic)
- Real-time feedback mechanisms
- How environmental constraints are communicated to the LLM
- Methods for handling uncertainty and partial observability

- **Planning Approach**:

  Describe the planning methodology including:

- Planning horizon length and complexity
- Task decomposition strategies
- How long-horizon tasks are broken down into sub-tasks
- Replanning/correction mechanisms
- Integration of high-level planning with low-level control
- Communication strategies (for multi-agent systems)

- **Task Domains**:

  Extract information about the tasks and environments:

- Types of tasks performed (household, navigation, manipulation, cooperative)
- Task complexity measures (number of steps, duration, objects involved)
- Environment characteristics (indoor/outdoor, size, number of rooms/floors)
- Real-world vs simulation testing
- Specific benchmarks or datasets used

- **Performance Results**:

  Collect all quantitative and qualitative performance measures:

- Success rates with confidence intervals if provided
- Comparison baselines and methods
- Performance metrics used (task completion, efficiency, accuracy)
- Failure analysis and common failure modes
- Statistical significance of results
- Human evaluation results if applicable

- **Key Limitations**:

  Identify challenges and limitations reported:

- Technical limitations of the approach
- Scalability concerns
- Performance gaps compared to specialized methods
- Environmental constraints or assumptions
- Computational requirements
- Areas identified for future improvement

Results

Characteristics of Included Studies

This systematic review examines 10 primary studies investigating LLM-based approaches to long-horizon planning for embodied agents. All studies were published between 2022 and 2025, reflecting the rapid recent development in this field.

Study	Full text retrieved?	Study Type	Primary Focus
Yutao Ouyang et al., 2024	Yes	Primary study	Quadrupedal robot locomotion and manipulation with hierarchical LLM planning
Krishan Rana et al., 2023	Yes	Primary study	Scalable task planning using 3D scene graphs for mobile manipulators
Hongxin Zhang et al., 2023	Yes	Primary study	Multi-agent cooperation with decentralized control and communication
Yaran Chen et al., 2025	No	Primary study	Long-term decision-making for instruction following tasks
Fangyuan Wang et al., 2025	Yes	Primary study	Humanoid robot mobile manipulation with grounding mechanisms
Chan Hee Song et al., 2022	Yes	Primary study	Few-shot grounded planning with dynamic replanning
Siddharth Nayak et al., 2024	Yes	Primary study	Multi-agent planning in partially observable environments
Wen Jiang et al., 2024	No	Primary study	Active mapping and exploration with multimodal LLMs
Siwei Chen et al., 2023	No	Primary study	Open-world state representation for task planning
Gautier Dagan et al., 2023	Yes	Primary study	Neuro-symbolic framework combining LLMs with classical planning

The studies span diverse application domains, from quadrupedal locomotion to household tasks , mobile manipulation , and multi-agent cooperation . Seven studies provided full-text access, while three were available only as abstracts.

Large Language Models Employed

Study	LLM Model(s)	Fine-tuned?	Input Modalities	Prompting Strategies
Yutao Ouyang et al., 2024	GPT-4-turbo-preview	Pre-trained	Text	Cascade of LLM agents; if-else branching structures
Krishan Rana et al., 2023	GPT-4	Pre-trained	Text	Static prompt with updates during semantic search
Hongxin Zhang et al., 2023	GPT-4, LLAMA-2-13b-chat	LLAMA-2 fine-tuned with LoRA	Text	Zero-shot chain-of-thought prompting
Yaran Chen et al., 2025	Llama	Fine-tuned on 67k embodied planning data	Text	Not mentioned
Fangyuan Wang et al., 2025	GPT-4o	Not mentioned	Text, vision (multimodal)	Interactive planning with multiple action candidates
Chan Hee Song et al., 2022	GPT-3 (TEXT-DAVINCI-003)	Pre-trained	Text	kNN retriever for in-context examples; logit biases
Siddharth Nayak et al., 2024	GPT-4, GPT-4V, IDEFICS-2 (8B), LLaVA (7B), CoGVLM (18B)	IDEFICS-2 and LLaVA fine-tuned	Text, vision, multimodal	BERT fine-tuning for semantic mapping
Wen Jiang et al., 2024	Multimodal LLM (unspecified)	Implied pre-trained (zero-shot)	Multimodal (text and vision)	Not mentioned
Siwei Chen et al., 2023	LLM (unspecified)	Not mentioned	Text (implied)	Not mentioned
Gautier Dagan et al., 2023	gpt-3.5-turbo-0613	Pre-trained	Text	Modified prompts to reduce conversational tendencies

The majority of studies employed OpenAI's GPT family models, with GPT-4 being the most common choice . Two studies explored fine-tuning approaches: one fine-tuned LLAMA-2 using LoRA on human-filtered data , while another fine-tuned Llama on 67k embodied planning samples using template feedback-based self-instruction . Vision-language models were employed in three studies to enable multimodal reasoning .

System Architectures

Study	Architecture Type	Key Components	Multi-Agent LLM?	Integration with Other AI Methods
Yutao Ouyang et al., 2024	Hierarchical	Semantic planner, parameter calculator, code generator, replanner	Yes (cascade of agents)	Reinforcement learning for low-level control
Krishan Rana et al., 2023	Modular	LLM planner, classical path planner, scene graph simulator	No	Classical planning for navigation
Hongxin Zhang et al., 2023	Modular (cognitive-inspired)	Perception, Memory, Communication, Planning, Execution modules	No	A-Star planner for execution
Yaran Chen et al., 2025	Modular	RoboPlanner, RoboSkill, Re-Plan	No	Perception modules through RoboSkill
Fangyuan Wang et al., 2025	Modular	LLM planner, skill library, PDDL problem construction	No	Classical planning using PDDL
Chan Hee Song et al., 2022	Hierarchical	High-level LLM planner, low-level action planner	No	In-context learning
Siddharth Nayak et al., 2024	Modular (cognitive)	Planner, Actor, Corrector, Verifier modules	Yes (each module is an LM)	Visual feedback processing
Wen Jiang et al., 2024	Modular	LLM planner, 3DGS representation, motion planning algorithm	No	Information-based motion planning
Siwei Chen et al., 2023	Not specified	Open state representation for tracking objects	Not mentioned	Not mentioned
Gautier Dagan et al., 2023	Neuro-symbolic	Plan Generator, Action Selector, PDDL translator	No	Classical symbolic planning via PDDL

Two dominant architectural patterns emerged: hierarchical systems that separate high-level reasoning from low-level control , and modular frameworks that decompose functionality across specialized components . Two studies employed multiple LLM agents within their architectures—one using a cascade of specialized agents for planning, parameter calculation, and code generation , and another assigning distinct LM-based modules for planning, acting, correcting, and verifying . Integration with classical AI methods was common, including reinforcement learning , A-Star search , and PDDL-based symbolic planning .

Embodied Platforms and Grounding Mechanisms

Study	Agent Type	Single/Multi-Agent	Physical Capabilities	Testing Environment
Yutao Ouyang et al., 2024	Quadrupedal robot (Xiaomi Cyberdog2)	Single	Locomotion, manipulation	Simulation and real-world
Krishan Rana et al., 2023	Mobile manipulator (Franka Panda on Omron base)	Single	Locomotion, manipulation, LiDAR sensing	Real-world
Hongxin Zhang et al., 2023	Humanoid	Multi-agent	Navigation, manipulation, RGB-D sensing	Simulation (TDW platform)
Yaran Chen et al., 2025	Mobile manipulator (implied)	Not mentioned	Navigation, manipulation	Not mentioned
Fangyuan Wang et al., 2025	Humanoid robot (Tiago++)	Single	Wheels, 7-DoF arms, RGB-D camera	Real-world
Chan Hee Song et al., 2022	Mobile manipulator (implied)	Single	Locomotion, manipulation	Simulation (AI2-Thor/ALFRED)
Siddharth Nayak et al., 2024	Robotic (unspecified)	Multi-agent	Navigation, manipulation, sensing	Simulation (AI2Thor)
Wen Jiang et al., 2024	Embodied agent (unspecified)	Not mentioned	Not mentioned	Simulation (Gibson, Habitat-Matterport 3D)
Siwei Chen et al., 2023	Not mentioned	Not mentioned	Not mentioned	Simulation and real-world
Gautier Dagan et al., 2023	Single agent	Single	Navigation, object manipulation	Simulation (Alfworld)

The studies employed diverse embodied platforms, including quadrupedal robots , mobile manipulators , and humanoid agents . Two studies specifically addressed multi-agent coordination . Testing was predominantly conducted in simulation environments such as AI2-Thor , TDW , and Alfworld , though four studies included real-world experiments .

Study	Perception Systems	World Representation	Feedback Mechanisms	Uncertainty Handling
Yutao Ouyang et al., 2024	RealSense camera, AprilTags, third-person camera	Geometric (depth maps, trajectories)	Replanner with object detection	Domain randomization, observation noise
Krishan Rana et al., 2023	3D scene graphs, LiDAR	Geometric and semantic (3DSGs)	Scene graph simulator feedback	Hierarchical semantic search, iterative replanning

Study	Perception Systems	World Representation	Feedback Mechanisms	Uncertainty Handling
Hongxin Zhang et al., 2023	Mask-RCNN, RGB images, 3D point clouds	Semantic maps	A-Star planner for navigation	Semantic maps and episodic memory
Yaran Chen et al., 2025	Not mentioned	Semantic (precise semantic map)	Re-Plan module with environmental feedback	Not mentioned
Fangyuan Wang et al., 2025	Integrated RGB-D camera	Symbolic (PDDL) and geometric (MDPs)	Closed-loop sensor feedback	Interactive planning with feasibility feedback
Chan Hee Song et al., 2022	Object detector from HLSM	Symbolic (object lists)	Dynamic re-planning on failures	Confidence threshold for detection
Siddharth Nayak et al., 2024	Images, textual descriptions (90° FOV, 1.5m range)	AI2Thor photorealistic environment	Corrector module adjusts on failures	Iterative refinement through exploration
Wen Jiang et al., 2024	3DGS view synthesis	3D Gaussian Splatting, semantic	Not mentioned	Uncertainty-aware path selection
Siwei Chen et al., 2023	Not mentioned	Symbolic (open state representation)	Not mentioned	Not mentioned
Gautier Dagan et al., 2023	Assumes symbolic input	Symbolic (PDDL)	Updates state based on action results	Sampling predicates and beliefs

World representation approaches varied substantially across studies. Symbolic representations using PDDL were employed in three studies , while semantic representations including 3D scene graphs and semantic maps were also common. Several studies combined multiple representation types to capture different aspects of the environment .

Planning Approaches

Study	Task Decomposition Strategy	Replanning Mechanism	High-Low Level Integration
Yutao Ouyang et al., 2024	Semantic planner sketches primitive skills; parameter calculator adds details	Replanner adjusts based on execution feedback	RL-based motion policies execute high-level plans
Krishan Rana et al., 2023	Identifies task-relevant subgraphs from collapsed scene graphs	Iterative feedback from scene graph simulator	Classical path planner handles navigation
Hongxin Zhang et al., 2023	LLM generates high-level plans based on state and procedural knowledge	Dynamic adjustment through LLMs implied	A-Star planner executes high-level plans
Yaran Chen et al., 2025	RoboPlanner breaks tasks into logical subgoals	Re-Plan adjusts subgoals based on real-time feedback	RoboSkill provides navigation and manipulation

Study	Task Decomposition Strategy	Replanning Mechanism	High-Low Level Integration
Fangyuan Wang et al., 2025	PDDL problem construction from user instructions	Interactive replanning at each timestep	Library of skills with single-timestep MDPs
Chan Hee Song et al., 2022	LLM generates high-level plans as subgoal sequences	Grounded replanning based on observations and failures	Low-level planner maps subgoals to primitives
Siddharth Nayak et al., 2024	Planner suggests subtasks; Actor selects high-level actions	Corrector self-corrects based on execution failures	Actor uses corrective actions to predict actions
Wen Jiang et al., 2024	Not mentioned	Uncertainty-aware path proposal and selection	Multimodal LLMs with detailed motion planning
Siwei Chen et al., 2023	Not explicitly mentioned	Not explicitly mentioned	Not explicitly mentioned
Gautier Dagan et al., 2023	Task descriptions translated to goal states via LLM	Action Selector prompts re-planning when no valid plans found	Plan Generator solves PDDL; Action Selector decides actions

All studies that provided detailed methodology employed some form of hierarchical task decomposition, breaking long-horizon tasks into manageable subgoals or subtasks . Replanning mechanisms were consistently incorporated to handle execution failures and environmental changes . The integration between high-level LLM planning and low-level control was achieved through various means: reinforcement learning policies , classical planners , skill libraries , and learned low-level controllers .

For multi-agent systems, communication strategies included LLM-generated natural language messages between agents and coordination through shared state information to avoid conflicts .

Task Domains and Evaluation Environments

Study	Task Types	Environment Characteristics	Benchmarks/Datasets
Yutao Ouyang et al., 2024	Light switch manipulation, package delivery, bridge building, elevator use	Indoor, multi-step	Custom tasks for quadrupeds
Krishan Rana et al., 2023	Semantic search, navigation, manipulation, long-horizon reasoning	Office (37 rooms, 151 assets); Home (28 rooms, 3 floors, 112 assets)	Stanford 3D Scene Graph
Hongxin Zhang et al., 2023	Object transport, household activities	Indoor, simulated, multi-room	TDW-MAT, C-WAH
Yaran Chen et al., 2025	Daily tasks from natural language instructions	Not mentioned	67k embodied planning data
Fangyuan Wang et al., 2025	Long-horizon mobile manipulation	Real-world, open-world setting	Not mentioned

Study	Task Types	Environment Characteristics	Benchmarks/Datasets
Chan Hee Song et al., 2022	Household tasks (object manipulation)	Indoor, 207 unique environments	ALFRED dataset
Siddharth Nayak et al., 2024	Household cleaning, grocery organization, search & rescue	Single-room (AI2Thor); Multi-area SAR environment	AI2Thor, MAP-THOR (custom)
Wen Jiang et al., 2024	Navigation, mapping, exploration	Indoor (implied)	Gibson, Habitat-Matterport 3D
Siwei Chen et al., 2023	Long-horizon task planning	Open-world household (indoor)	Not mentioned
Gautier Dagan et al., 2023	Household tasks, navigation, manipulation	Indoor, text-only environment	Alfworld

Task complexity varied considerably across studies. The most challenging environments involved multi-floor buildings with up to 37 rooms and 151 objects , while simpler settings used single-room configurations . Long-horizon tasks requiring approximately 100 low-level actions were evaluated in one study , while another measured an average of 13.16 actions per task . The ALFRED dataset was commonly used for benchmarking household tasks , while AI2Thor served as a standard simulation platform .

Performance Results

Study	Primary Metric	Key Results	Comparison Baselines
Yutao Ouyang et al., 2024	Success rate	>70% in simulation	RoboTool, hierarchical RL
Krishan Rana et al., 2023	Executability	Near-perfect with iterative replanning	LLM-As-Planner, LLM+P
Hongxin Zhang et al., 2023	Efficiency improvement	>40% improvement vs baselines	MHP, RHP
Yaran Chen et al., 2025	Task planning rationality	Exceeds SOTA methods	Other LLM-based methods
Fangyuan Wang et al., 2025	Success rate	>80% average	Ablations without feasibility feedback
Chan Hee Song et al., 2022	Success rate	Competitive with <0.5% training data	HLSM, FILM, SayCan
Siddharth Nayak et al., 2024	Success rate	30% higher than SOTA multi-agent planners	Other LM-based planners
Wen Jiang et al., 2024	Not specified	State-of-the-art results	Not mentioned
Siwei Chen et al., 2023	Not specified	Significant improvements over baselines	Baseline methods (unspecified)
Gautier Dagan et al., 2023	Success rate	96% (vs 53% for ReAct)	ReAct

The studies reported strong performance across various metrics. The highest absolute success rate was achieved by the neuro-symbolic LLM-DP framework at 96% , nearly doubling the performance of the ReAct baseline. In multi-agent settings, LLaMAR achieved a 30% higher success rate compared to other state-of-the-art planners , with results

reported with 95% confidence intervals . The few-shot learning approach of LLM-Planner demonstrated competitive performance using less than 0.5% of the training data required by fully-trained methods , highlighting the sample efficiency of LLM-based approaches.

Efficiency gains were substantial in cooperative settings, with CoELA achieving over 40% efficiency improvement compared to planning-based baselines . Human evaluation in this study revealed that natural language communication significantly increased trust (6.3 vs 4.7 trust score, p=0.0003) .

Failure analysis across studies identified several common failure modes. In LLM-Planner, the majority of failures stemmed from the low-level controller rather than high-level planning . LLaMAR's failures included incorrect causal ordering, sub-optimal action sequences, and catastrophic failure misunderstanding . The IALP system categorized failures into planning, promptable, and grounding mechanism failures .

Key Limitations

Study	Technical Limitations	Scalability Concerns	Areas for Improvement
Yutao Ouyang et al., 2024	Fixed set of expert-crafted skills	Limited to simple scenarios	Autonomous skill discovery
Krishan Rana et al., 2023	LLM biases affect plan validity; limited graph reasoning	Requires pre-built scene graph; static object assumption	Fine-tuning LLMs; online scene graph SLAM
Hongxin Zhang et al., 2023	Limited 3D spatial information; unstable complex reasoning	LLAMA-2 underperforms GPT-4	Incorporating 3D spatial info; improving low-level reasoning
Yaran Chen et al., 2025	LLM plans suffer accuracy/feasibility issues	Multiple module integration complexity	Not mentioned
Fangyuan Wang et al., 2025	Lack of grounding knowledge; imprecise spatial extraction	Predefined action libraries limit generalization	Generalized action models; enhanced VLM reasoning
Chan Hee Song et al., 2022	Static LLM plans don't adapt; enumeration assumption	Combinatorial action growth	Better low-level planners; advanced grounding methods
Siddharth Nayak et al., 2024	Limited spatial reasoning; VLM mistakes	Congestion with more agents	Task distribution balance among agents
Wen Jiang et al., 2024	Not mentioned	Not mentioned	Not mentioned
Siwei Chen et al., 2023	Existing works fail to track objects/attributes	Engineered features not generalizable	Not mentioned
Gautier Dagan et al., 2023	LLM hallucination; prompt brittleness	High computational costs from iterated LLM calls	Symbolic world model encoding; enhanced self-reflection

Several cross-cutting limitations emerged across studies. First, LLM hallucination and reasoning errors were consistently identified as challenges . Second, spatial reasoning capabilities were noted as insufficient in multiple studies . Third, the dependence on predefined skill libraries or action spaces limited generalization . Fourth, computational costs associated with repeated LLM queries were a practical concern .

Environmental assumptions also constrained applicability. Several systems required pre-built representations such as 3D scene graphs or assumed static environments . The assumption of noise-free communication and maximum visibility in simulation may not transfer to real-world deployments.

Synthesis

The studies present a coherent picture of LLM-based long-horizon planning while revealing important distinctions in approach and context that explain varying results.

Architectural Trade-offs : The choice between hierarchical and modular architectures reflects different priorities. Hierarchical systems with cascaded LLM agents achieved strong performance on complex locomotion-manipulation tasks requiring tight coordination between planning and execution. In contrast, modular cognitive architectures proved more effective for multi-agent coordination where distinct reasoning capabilities (perception, memory, communication, planning) needed separation. The neuro-symbolic approach achieved the highest success rate (96%) by leveraging symbolic planning's optimality guarantees while using LLMs for grounding—suggesting that hybrid architectures may outperform pure LLM approaches in well-structured domains.

Grounding Mechanism Impact : Studies employing rich environmental representations consistently outperformed those with simpler grounding. The 3D scene graph approach achieved near-perfect executability in large-scale environments with up to 37 rooms , while iterative replanning corrected most LLM errors. Conversely, studies relying on simpler object lists or text-only environments were limited to smaller action spaces and simpler task structures. This suggests that grounding quality—rather than LLM capability alone—may be the primary bottleneck for real-world deployment.

Single vs. Multi-Agent Performance : The multi-agent studies demonstrated that LLM-based coordination can achieve substantial efficiency gains (>40%) through emergent communication behaviors. However, scalability challenges emerged: increasing agent numbers led to congestion and decreased performance . The human evaluation finding that natural language communication increased trust (p=0.0003) suggests practical benefits for human-robot teaming that extend beyond task performance metrics.

Pre-trained vs. Fine-tuned Models : The studies using pre-trained GPT-4 generally achieved strong performance without task-specific training, while fine-tuned approaches offered potential for domain adaptation. The finding that LLAMA-2 underperformed GPT-4 but could be improved through fine-tuning on human-filtered data indicates that open models may require additional training investment to match proprietary alternatives.

Failure Mode Patterns : Analysis across studies revealed that failures predominantly occurred at the interface between high-level planning and low-level execution. In LLM-Planner, most failures stemmed from the low-level controller rather than planning . The IALP system's failure taxonomy (planning, promptable, and grounding mechanism failures) suggests that robust systems require attention to all three components. Spatial reasoning limitations were consistently problematic, indicating that current LLMs lack the geometric understanding necessary for precise manipulation in complex environments.

References

Chan Hee Song, Jiaman Wu, Clay Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su. "LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models." *IEEE International Conference on Computer Vision*, 2022.

Fangyuan Wang, Shipeng Lyu, Peng Zhou, Anqing Duan, Guodong Guo, and D. Navarro-Alarcón. "Instruction-Augmented Long-Horizon Planning: Embedding Grounding Mechanisms in Embodied Mobile Manipulation." *AAAI Conference on Artificial Intelligence*, 2025.

Gautier Dagan, Frank Keller, and A. Lascarides. "Dynamic Planning with a LLM." *arXiv.org*, 2023.

Hongxin Zhang, Weihua Du, Jiaming Shan, Qinhong Zhou, Yilun Du, J. Tenenbaum, Tianmin Shu, and Chuang Gan. "Building Cooperative Embodied Agents Modularly with Large Language Models." *International Conference on Learning Representations*, 2023.

Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, I. Reid, and Niko Sünderhauf. "SayPlan: Grounding Large Language Models Using 3D Scene Graphs for Scalable Task Planning." *Conference on Robot Learning*, 2023.

Siddharth Nayak, Adelmo Morrison Orozco, M. T. Have, Vittal Thirumalai, Jackson Zhang, Darren Chen, Aditya Kapoor, et al. "Long-Horizon Planning for Multi-Agent Robots in Partially Observable Environments." *Neural Information Processing Systems*, 2024.

Siwei Chen, Anxing Xiao, and David Hsu. "LLM-State: Open World State Representation for Long-Horizon Task Planning with Large Language Model," 2023.

Wen Jiang, Boshu Lei, Katrina Ashton, and Kostas Daniilidis. "Multimodal LLM Guided Exploration and Active Mapping Using Fisher Information," 2024.

Yaran Chen, Wenbo Cui, Yuanwen Chen, Mining Tan, Xinyao Zhang, Jinrui Liu, Haoran Li, Dong Zhao, and He Wang. "RoboGPT: An LLM-Based Long-Term Decision-Making Embodied Agent for Instruction Following Tasks." *IEEE Transactions on Cognitive and Developmental Systems*, 2025.

Yutao Ouyang, Jinhan Li, Yunfei Li, Zhongyu Li, Chao Yu, K. Sreenath, and Yi Wu. "Long-Horizon Locomotion and Manipulation on a Quadrupedal Robot with Large Language Models." *arXiv.org*, 2024.