Abstract

Large Language Models (LLMs) have become a transformative component in modern robotics, enabling robots to interpret natural language, plan sequential tasks, reason over multimodal inputs, and execute manipulation skills across dynamic environments. As robotics moves toward higher levels of autonomy and embodied intelligence, LLM-based systems increasingly form the cognitive core that integrates perception, decision-making, control, and human–robot interaction. Motivated by rapid progress in the field, this paper presents a comprehensive synthesis of recent advancements in LLM-enabled robotics, grounded in analyses from multiple foundational studies and our consolidated understanding extracted from *Summary.pdf*. We explore how LLMs enhance natural-language-based task planning, multimodal perception, manipulation control, interactive decision-making, and modular robot programming structures, referencing techniques such as grounded planning [1.Large Language Models for Robotics: Opportunities, Challenges, and Perspectives], ontology-based action parsing [2.Parsing Natural Language Sentences into Robot Actions], and systematic human–robot instruction frameworks [3.A Review of Natural-Language-Instructed Robot Execution Systems]. We further discuss domain-specific fine-tuning approaches for industrial robot programming [4.Domain-Specific Fine-Tuning…] and multimodal reasoning-tuned manipulation methods, such as RT-Grasp [5.RT-Grasp…], as well as recent advances in retrieval-augmented robotic control pipelines [6.ARRC…] and structured ROS-integrated LLM systems [7.ROS-LLM…]. By organizing insights across planning, reasoning, manipulation, perception, control architectures, and interaction strategies, we offer a unified perspective on the current state of LLM-enabled embodied intelligence. Finally, we outline challenges—including safety, grounding reliability, numerical precision, and real-world generalization—and discuss promising avenues toward scalable, human-aligned, and autonomous robotic systems capable of operating across complex unstructured environments.

Keywords

Large Language Models (LLMs); Robotics; Embodied Intelligence; Natural Language Understanding; Task Planning; Manipulation; Multimodal Reasoning; Retrieval-Augmented Generation; Human–Robot Interaction; Robot Control Systems

---

1. Introduction

1.1. Background and Motivation

The integration of Large Language Models (LLMs) into robotics marks a pivotal shift in how robots understand, reason about, and interact with the physical world. While traditional robotics has long relied on structured programming, domain-specific control policies, and carefully engineered models, recent advancements in natural language processing have introduced new frameworks where robots can interpret human intent directly through natural language instructions. This evolution reflects a broader paradigm change toward embodied intelligence,

where robots learn, act, and collaborate using intuitive and cognitively enriched representations. As highlighted in foundational surveys, LLMs introduce new ways for robots to engage in natural-language-based planning, decision-making, perception, and control [8.Large Language Models for Robotics: A Survey].

Historically, robotic systems required explicit programming and extensive domain knowledge. However, LLMs function as generalized reasoning engines capable of extracting contextual meaning, producing task plans, and synthesizing executable actions. This capability enables complex tasks—from navigation to manipulation—to be expressed conversationally rather than through specialized code or carefully structured symbolic instructions. For example, modern LLM-centric planning systems demonstrate the ability to translate linguistic instructions into action sequences even in the absence of visual context, while achieving significant performance gains when multimodal information is incorporated [1.Large Language Models for Robotics: Opportunities, Challenges, and Perspectives].

Multimodal LLMs expand these possibilities by integrating visual, linguistic, and contextual inputs. By processing both natural language and sensory data, they create grounded representations that bridge human communication with robot control. These models support complex reasoning, object affordance understanding, and spatial awareness—properties essential for reliable execution in unstructured real-world environments. Such capabilities reflect a growing consensus that LLM-based reasoning is central to enabling robots to achieve higher degrees of autonomy, adaptability, and human alignment.

1.2. From Language to Action: The Role of Natural Language Understanding

A critical function of LLMs in robotics lies in converting human language into symbolic, structured, or executable representations. Early research focused on natural language parsing pipelines, such as ontology-based systems that identified robot actions, body parts, and constraints from parsed linguistic structure [2.Parsing Natural Language Sentences into Robot Actions]. These systems demonstrated the potential for robots to respond intelligently to human commands while maintaining consistency with their internal states—for instance, rejecting infeasible actions due to balance constraints.

Modern LLMs extend this capability dramatically. With deep semantic understanding, they capture implicit intent, temporal dependencies, and task-specific logic. As noted in research on natural-language-guided execution, current systems can interpret commands such as “I am hungry” and derive the implied task “get me food” by using embedded commonsense associations [3.A Review of Natural-Language-Instructed Robot Execution Systems]. This ability presents a major departure from earlier rule-based systems, shifting toward adaptive interpretation grounded in contextual and world knowledge.

LLMs also aid in the disambiguation of commands, enabling clarification queries when information is missing. This aligns with human-robot communication frameworks where linguistic interaction serves not only as an instruction mechanism but also as a bidirectional channel for negotiation, clarification, and refinement of intended tasks [7.ROS-LLM…]. As

such, natural language serves as both the medium for commanding robots and the substrate for interactive decision-making.

1.3. Advances in LLM-Based Planning and Reasoning

Recent literature identifies task planning as one of the most impactful domains for LLM integration. Studies demonstrate that LLMs generate accurate sequential plans, incorporate commonsense reasoning, and collaborate effectively with classical planners to enhance precision and feasibility [1.Large Language Models for Robotics: Opportunities, Challenges, and Perspectives]. This synergy enables robots to:

- decompose long-horizon instructions into structured steps,
- reason about preconditions and spatial relationships,
- leverage probabilistic grounding mechanisms,
- and incorporate heuristic knowledge into search algorithms.

Furthermore, frameworks such as SayCan and LLM+P couple LLM semantic guidance with low-level reinforcement learning policies, enabling robots to map high-level language goals to executable skill policies [1…]. Other systems introduce preconditioned error cues, multimodal feedback loops, and self-reflective reasoning models such as ReAct, improving robustness in uncertain or dynamic environments.

Complex task reasoning extends beyond planning into decision-making and adaptation. For example, LM-Nav integrates language, vision, and navigation priors to allow robots to follow high-level descriptions without requiring trajectory annotations [1…]. Meanwhile, uncertainty-aware models like KnowNo help mitigate hallucinated or unsafe decisions during long-horizon tasks by incorporating statistical reliability estimates.

Together, these methods signal a shift from LLMs as passive language processors to active reasoning agents embedded in robotic control architectures.

1.4. Manipulation, Perception, and Embodied Skills Enhanced by LLMs

While language-driven task planning represents one major breakthrough, manipulation and embodied control present a distinct set of challenges. Traditional robotic manipulation relies heavily on numerical precision, geometric modeling, and carefully tuned perception pipelines. However, LLMs introduce a new dimension—semantic reasoning—which augments robots with contextual understanding of object properties, affordances, and common-sense constraints.

Systems such as LLM-GROP and VIMA demonstrate how linguistic cues can enrich spatial reasoning, enabling robots to perform tasks such as collision-free placement or multimodal imitation learning using only high-level cues [1…]. Multimodal frameworks unify vision, language, and motion, allowing robots to generalize skills across categories of tasks without explicit task-specific data.

Moreover, the introduction of reasoning-tuned multimodal manipulation—as seen in RT-Grasp—addresses a long-standing barrier: numerical precision. By combining symbolic reasoning outputs with structured numerical prediction templates, RT-Grasp leverages LLM-generated explanations to refine low-level grasp parameters, bridging the gap between semantic reasoning and geometric execution [5.RT-Grasp…]. This highlights a growing trend of using LLMs not only for planning but for guiding precise physical actions previously dominated by purely analytical or learning-based controllers.

1.5. Retrieval-Augmented and Modular Architectures

As LLMs grow in capability but remain prone to hallucinations, retrieval-augmented generation (RAG) emerges as a crucial paradigm for grounding reasoning in verified robot-specific knowledge. ARRC exemplifies this trend by indexing robot motion templates, safety heuristics, and task patterns into a vector database that supplies relevant procedural information to the LLM during planning [6.ARRC…]. This approach mitigates safety risks and enhances task validity, demonstrating how hybrid architectures combine LLM flexibility with curated operational constraints.

Similarly, modular systems such as ProgramPort or MetaMorph utilize LLMs to interpret linguistic inputs into programmatic modules or to adapt policies across varying robot morphologies [1…]. These modular strategies improve scalability and generalization, enabling robots to compose skills, adapt to new tools, and operate in diverse environments using structured semantic programs.

1.6. Human–Robot Interaction and Interactive Learning

LLMs significantly advance the domain of human–robot interaction (HRI). Systems like ROS-LLM illustrate that LLM-enabled robots can engage in iterative conversations, incorporate human feedback, execute hierarchical behaviors, and improve over time. Interactive frameworks also support multimodal input—text, images, and sensor observations—permitting robots to resolve ambiguities, learn from corrections, and adapt to environmental disturbances [7.ROS-LLM…].

Additionally, generative agents and multimodal interactive strategies enable robots to simulate human-like reasoning patterns, store memories, and act based on accumulated experience, marking a progression toward more intuitive and collaborative autonomous systems.

1.7. Contributions of This Paper

Drawing upon consolidated insights from Summary.pdf and eight foundational robotics–LLM papers, this work offers:

1. A unified academic synthesis of natural language understanding, planning, reasoning, manipulation, and interaction strategies in LLM-enabled robotics.
2. Integration of findings from industrial fine-tuning, multimodal manipulation, retrieval-augmented planning, and ROS-based interactive frameworks.

3. A cross-sectional analysis of current limitations—including grounding reliability, safety enforcement, and real-world generalization.
4. A holistic perspective on the trajectory toward embodied intelligence influenced by LLM advancements.

2. Main Body

2.1. Natural Language Understanding for Robotic Systems

Natural Language Understanding (NLU) is the foundational capability that enables LLM-based robotic systems to interpret human intentions in a form that can guide autonomous behavior. Across the literature, NLU defines how the robot parses linguistic input, extracts semantic structure, identifies actionable elements, resolves ambiguity, and transforms natural language into executable commands. Earlier approaches relied heavily on linguistic pipelines and symbolic representations, as demonstrated in ontology-based frameworks that link verbs, body parts, and actions using dependency-parsed linguistic structure [2.Parsing Natural Language Sentences into Robot Actions]. These systems leveraged ontologies to categorize robot actions, resolve missing information, and ensure consistency with the robot's physical state. Such pipelines were crucial in enabling robots to handle commands like "raise your arm" by determining whether the user meant the left or right arm and prompting clarification when necessary.

LLMs expand these capabilities beyond purely structural linguistic processing. With semantic-rich embeddings and extensive pre-training on textual corpora, modern LLMs capture nuanced relationships between instructions, environmental cues, and implied intent. For example, research on LLM-enabled planning shows that models can decompose abstract commands into coherent action sequences without explicit visual input, while additional visual context (such as multimodal prompts) greatly improves accuracy and grounding [1.Large Language Models for Robotics: Opportunities, Challenges, and Perspectives]. These results highlight the advantage of integrating linguistic reasoning with multimodal perception.

Furthermore, advancements in NL-instructed execution systems demonstrate the importance of commonsense grounding. Robots increasingly interpret indirect commands by identifying latent intent (e.g., interpreting "I am cold" as a request to turn off the fan or adjust the heater) [3.A Review of Natural-Language-Instructed Robot Execution Systems]. Such capabilities move beyond strict verb–object parsing toward a deeper model of human intention. Because LLMs incorporate vast world knowledge, they allow robots to infer meaning in ways previously limited to human reasoning.

Another critical advancement is the ability to incorporate logic-based semantics into execution. NL-based execution control systems integrate linguistic temporal connectors such as "then," "after," or "before," enabling structured task decomposition [3…]. These logic relations are key

for hierarchical task plans—for example, distinguishing between “fill the cup, then deliver it” versus “deliver the cup, then fill it,” which produces drastically different outcomes.

Taken together, these developments reveal how LLM-driven NLU forms the cognitive foundation for robotics systems that can interpret, reason, and act in response to complex human instructions. As LLMs continue to evolve, their NLU capabilities increasingly resemble an intuitive grasp of human intent, broadening the potential for natural, efficient, and collaborative human–robot interaction.

2.2. Task Planning With Large Language Models

Task planning represents one of the domains where LLMs have delivered the most impressive performance improvements. Traditional robotic planning frameworks depend heavily on symbolic planners, search algorithms, or reward-driven reinforcement learning policies. While effective in well-defined environments, these methods often struggle in dynamic, ambiguous, or conceptually complex tasks requiring contextual reasoning. LLMs address this limitation by infusing high-level reasoning and semantic understanding into robotic planning pipelines.

Recent research shows that LLMs can produce long-horizon action sequences by interpreting abstract language and inferring essential task steps, even when the environment is only partially observed [1…]. One key advantage is their ability to integrate world knowledge and common sense. For example, an LLM-based planner understands that retrieving a drink typically involves locating and grasping a cup before pouring liquid into it. This built-in knowledge allows robots to generate valid steps even when certain specifics are omitted in the instruction.

LLMs also serve as heuristic modules in hybrid classical-planning systems. Approaches like LLM+P and SayPlan combine LLM-based semantic reasoning with symbolic planners described in formal languages such as PDDL, enabling robots to ground task descriptions into domain models [1…]. This integration supports robust, long-horizon planning where symbolic planners ensure feasibility, while LLMs provide semantic structure and contextual knowledge.

Furthermore, techniques such as grounded decoding enhance planning reliability by aligning LLM-generated sequences with physical constraints. By incorporating probabilistic models of the physical world, robots can generate behavior sequences consistent with environmental affordances, improving feasibility and consistency in real-world deployment [1…].

Multimodal planners further extend these capabilities by integrating vision-language models, enabling grounded task sequencing based on visual cues. Systems like LM-Nav demonstrate how LLMs can coordinate vision-based navigation, image-language correlations, and high-level reasoning to follow natural-language navigation commands [1…].

Collectively, these innovations demonstrate that LLMs provide a flexible, powerful planning framework capable of tackling complex tasks, integrating human intent, and supporting robust execution in dynamic environments.

2.3. Complex Reasoning and Decision-Making

Beyond planning, LLMs excel at high-level reasoning, making decisions, selecting between options, and anticipating the consequences of actions. This capability distinguishes LLM-enabled robots from traditional systems that rely solely on predefined policies or rule-based algorithms.

Several studies explore how LLMs incorporate semantic memory, generalize across tasks, and utilize structured reasoning templates. For instance, preconditioned error cues allow robots to refine action plans by extracting procedural insights from natural language [1…]. Meanwhile, advanced frameworks such as ReAct integrate reasoning with acting, enabling robots to maintain internal state, improve decision-making, and avoid hallucinated or contradictory actions by cross-validating outputs through intermediate reasoning steps.

KnowNo introduces an uncertainty-aware mechanism for decision-making, helping models avoid overconfident erroneous actions—a critical capability in safety-sensitive contexts [1…]. This framework minimizes human intervention by providing statistical guarantees for multi-step decision-making, improving reliability in complex tasks.

In multimodal decision-making, models like LM-Nav merge sensory input with linguistic reasoning. By combining pre-trained vision navigation models with LLM-based goal interpretation, robots can determine paths to user-specified landmarks using only high-level verbal descriptions [1…].

Other works emphasize collaborative or multi-agent decision-making, where LLMs coordinate between multiple agents or human partners. These systems highlight the importance of LLMs' ability to reason over distributed knowledge, negotiate multi-step strategies, and resolve conflicting objectives.

Collectively, complex reasoning frameworks reveal that LLMs serve as generalist decision-makers capable of integrating linguistic, perceptual, and contextual cues. This expands the scope of robotic applications beyond simple command execution toward autonomous, context-aware, adaptive behavior.

2.4. Manipulation and Physical Interaction Enhanced by LLMs

Manipulation tasks—such as grasping, placing, and object rearrangement—pose unique challenges because they require precise geometric reasoning and fine-grained physical control. Traditional methods rely heavily on numerical optimization, pose estimation, and data-driven control policies. LLM-based approaches introduce new opportunities by infusing manipulation systems with semantic insight and commonsense reasoning.

Research such as LLM-GROP illustrates how LLMs provide semantic cues to guide manipulation strategies, enabling robots to choose sensible object placements or infer high-level constraints (e.g., avoiding placing fragile items under heavy objects) [1…]. Similarly, frameworks like VIMA and TIP adopt multimodal cueing strategies where textual descriptions and visual cues jointly define manipulation tasks, enabling generalization across unseen combinations of objects and motions.

Fine-tuning multimodal LLMs for specific manipulation tasks demonstrates additional progress. RT-Grasp introduces "reasoning tuning," combining structured explanation templates with numerical prediction tasks. By leveraging the LLM's reasoning phase, RT-Grasp transforms abstract object categories (such as "cup" or "sunglasses") into concrete grasp strategies before predicting the exact numerical grasp parameters [5.RT-Grasp…]. This hybrid approach bridges high-level reasoning and low-level motor control—a long-standing gap in robotics.

Other works highlight sample-efficient fine-tuning of visual-language models (VLMs) to improve downstream manipulation performance. For example, R3M and LIV demonstrate that pre-trained human video datasets can significantly improve generalization and efficiency when fine-tuning robotic manipulation strategies [1…]. These techniques leverage the representational power of large-scale pretraining, enabling robots to understand object semantics, affordances, and manipulation constraints without extensive robot-specific training data.

Interactive manipulation strategies also gain from natural-language-guided adjustment. Systems like VoxPoser extract manipulability constraints directly from language, enabling robots to adapt actions dynamically based on textual instructions [1…].

Together, these advances demonstrate an emerging paradigm where manipulation is guided not solely by geometry or reinforcement learning, but by a fusion of semantics, multimodality, and structured reasoning. This marks a significant evolution toward embodied intelligence that mirrors human-like understanding of physical tasks.

2.5. Interactive Strategies and Human Feedback Integration

Robots must often adjust behavior based on human feedback, environmental changes, or task failures. LLM-centric frameworks provide powerful tools for this through dialogue-based correction, iterative refinement, and feedback-driven policy improvement.

Systems such as TEXT2REWARD convert natural language feedback into reward codes, enabling reinforcement learning agents to interpret human corrections as optimization objectives [1…]. InstructRL similarly uses LLMs to generate initial policy structures based on language descriptions, helping agents align with human preferences through iterative refinement.

Human-robot interaction frameworks like LILAC enable users to adjust trajectories or modify robot actions using natural language, supported by multimodal inputs such as images and scene

context [1…]. These systems demonstrate how LLMs facilitate rapid adaptation by translating human feedback into actionable control modifications.

The ROS-LLM framework extends this idea by integrating LLMs with ROS action libraries, enabling robots to reflect on task outcomes, refine their internal plans, and learn from user feedback during interactive sessions [7.ROS-LLM…]. This includes handling failure states, recovering from disturbances, and adjusting behavior in real time.

Generative agents further expand interactive capabilities by providing synthetic memory systems that allow robots to store and integrate past experiences, improving behavioral consistency and long-term adaptation.

In summary, LLM-enabled interactive strategies allow robots to learn, adapt, and improve in real time—essential characteristics for flexible deployment in homes, factories, and unstructured environments.

2.6. Modular Approaches for Scalable Robotic Intelligence

Modularity is a key architectural principle in robotics, and LLM-enabled systems have expanded the scope and flexibility of modular design. A modular robot system decomposes complex tasks into reusable components—skills, policies, planners, perception modules, behavior trees—and recomposes them dynamically. LLMs amplify this by mapping natural language into modular structures and enabling rapid reconfiguration of behaviors based on linguistic cues.

One example of modularity enhanced by LLM reasoning is PROGRAMPORT, which transforms natural language descriptions into programmatic modules representing robotic manipulation behaviors [1.Large Language Models for Robotics: Opportunities, Challenges, and Perspectives]. Instead of generating raw code directly, the system converts semantic language structures into neural modules, enabling robust generalization across unseen objects and hybrid environments. This modular structure supports the creation of complex behaviors without requiring developers to manually specify procedural logic.

Other work focuses on adapting robot strategies to new physical tools or morphologies using LLM-guided meta-learning. By generating geometric abstractions and vector representations of tools using linguistic descriptions, robots can rapidly adjust their strategies to accommodate unfamiliar tools [1…]. This demonstrates how LLMs serve as semantic translators between high-level concepts (e.g., “this tool is like a hook”) and low-level controllers.

Another major modular framework is NLMap, which combines vision-language models like CLIP and ViLD with planning frameworks such as SayCan to create flexible scene-level semantic maps [1…]. This integration enables robots to understand open-world environments, handle ambiguous instructions, and plan over long horizons with minimal additional data.

MetaMorph provides a complementary perspective by learning Transformer-based policies that generalize across robot morphologies. Instead of tying policies to specific joints or geometries, the system outputs morphology-aware strategies, enabling broad generalization in manipulation and locomotion tasks [1…].

Together, these modular approaches show that LLMs are not merely language interpreters—they are organizational frameworks that structure perception, control, and reasoning into scalable, reusable robotic intelligence. This opens the door to robots that can dynamically assemble and refine behaviors based on user intent, environmental context, and system capabilities.

2.7. Retrieval-Augmented Generation (RAG) for Robust Robotic Planning

As LLMs grow more powerful, one persistent challenge remains: hallucinations and incorrect reasoning. In robotics, such errors can lead to dangerous or physically impossible actions. Retrieval-Augmented Generation (RAG) provides a practical solution by grounding LLM outputs in curated external knowledge databases.

ARRC (Advanced Reasoning Robot Control) demonstrates how RAG can be used to generate safe, structured, and context-aware robotic manipulation plans [6.ARRC…]. In ARRC:

- A vector-indexed knowledge base stores movement primitives, templates, task patterns, and safety heuristics.
- The retrieval module extracts relevant information based on the user prompt and the robot's current environment.
- The LLM conditions its planning on this grounded information, producing JSON action plans.
- A safety gate validates the plan before execution.

This pipeline significantly reduces the likelihood of hallucinated or infeasible actions because the LLM is guided by real, robot-specific prior knowledge. Moreover, RAG systems reduce the need for expensive retraining—new skills can simply be added to the database.

The value of retrieval also extends to multimodal reasoning. By providing the LLM with environment-specific information (e.g., object poses from AprilTags), the robot can generate precise action sequences grounded in the actual scene configuration [6…]. This tight integration of perception and retrieval facilitates higher reliability in real-world settings.

RAG-based planning represents a convergence of conventional robotics (which values predictability, safety, and verifiable logic) and modern machine learning (which contributes adaptability, flexibility, and linguistic intelligence). The result is a hybrid architecture ideal for robust, safe, and scalable autonomous manipulation.

2.8. Perception, Scene Understanding, and Multimodal Integration

Perception is central to embodied intelligence. Robots interacting with physical spaces must interpret visual, spatial, and semantic information reliably. LLMs and VLMs contribute significantly to this domain by integrating linguistic context with visual cues to guide perception, reduce ambiguity, and enrich scene understanding.

Systems such as PaLM-E incorporate continuous sensory inputs—including images, state vectors, and text—into a unified representation, demonstrating strong generalization across visual question answering and manipulation tasks [8.Large Language Models for Robotics: A Survey]. This integration enables robots to respond to questions like “What is inside the drawer?” or “Where is the green cup?” with understanding tied to real-world objects and spatial scenes.

Similarly, LM-Nav uses pre-trained CLIP embeddings to link linguistic landmark descriptions to visual scenes and navigation maps. The integration of ViNG, CLIP, and GPT-based processing allows robots to navigate complex environments using high-level verbal commands like “go to the cafeteria and then turn left at the couch” [8…].

Perception modules also support manipulation tasks by providing object-centric information. For example, ARRC relies on AprilTags fused with depth data to estimate object poses accurately, enabling precise pick-and-place actions [6.ARRC…]. VIMA and TIP demonstrate how text and images jointly encode manipulation primitives, allowing multimodal models to understand task constraints such as “place the block to the left of the cone.”

Visual-language pretraining (e.g., R3M, LIV) is another major trend. By training on large-scale human interaction videos, these representations capture affordances, surface textures, and functional relationships, improving sample efficiency and zero-shot generalization in downstream tasks [1…].

In summary, LLM-enabled perception systems unify symbolic reasoning with sensor data, enabling robots to interpret the world not just geometrically but conceptually. This integration is critical for navigating unstructured, dynamically changing environments.

2.9. Control and Execution in LLM-Enabled Robotics

Control—the translation of intent into precise motor actions—is where the abstract reasoning of LLMs meets the physical constraints of robotic hardware. Because language models generate high-level sequences, they must integrate closely with classical controllers, low-level policies, and motion primitives to ensure safe, predictable behavior.

In many systems, LLMs output structured representations such as:

- JSON action sequences (ARRC),
- Python scripts (ROS-LLM),
- XML behavior trees (ROS-LLM),
- code-like plans (Cap),
- or subgoal specifications for reinforcement-learning agents (SayCan).

These structured outputs can then be parsed and executed by downstream controllers. For example, ROS-LLM maps LLM outputs to ROS actions and services, enabling seamless integration with navigation stacks, kinematic solvers, and motion planning libraries like MoveIt [7.ROS-LLM…].

Other systems leverage learning-based policies. SayCan uses reinforcement learning to evaluate the executability of LLM-suggested skills based on learned value functions [1…]. In this setup, the LLM proposes a sequence of high-level tasks, while RL ensures that the proposed actions align with the robot's actual capabilities and current environmental conditions.

VIMA, TIP, and MetaMorph highlight how Transformer-based architectures encode manipulation trajectories or multi-morphology motor patterns. These models bridge the gap between language cues and continuous control by representing trajectories as multimodal sequences.

Collectively, these approaches underscore a fundamental shift: instead of hand-coded logic, robotic control increasingly relies on interpretable language-based structures that interface with existing controllers. LLMs provide structure, while classical robotics ensures precision and feasibility.

2.10. Safety-Constrained Execution

Safety is paramount in robotics, especially when operating near humans or handling delicate objects. LLMs introduce new opportunities—and risks—because they rely on statistical

reasoning rather than deterministic logic. Several systems address this through explicit safeguards.

In ARRC, every action passes through a multi-layer safety gating mechanism [6.ARRC…]:

- Workspace validation
- Speed and acceleration limits
- Gripper force and torque gating
- Per-step timeouts
- Emergency retreat modes
- Bounded retries for unstable tasks

These constraints act as physical filters ensuring that even if the LLM generates an incorrect or unsafe plan, the system will catch and reject it before execution.

KnowNo provides complementary theoretical safety by applying uncertainty estimation to LLM-generated plans. It prevents the robot from executing actions with insufficient confidence, adding a statistical safety layer [1…].

ROS-LLM integrates failure flags into its MDP formalism, enabling the system to recognize execution failures and respond with corrective actions or human queries [7.ROS-LLM…]. This improves resilience and adaptability in dynamic environments.

Together, these systems demonstrate that LLM-enabled robotics requires hybrid architectures: LLMs provide reasoning, while symbolic validators ensure safety. This dual approach will remain central as robots increasingly operate in unstructured human-centric environments.

2.11. Human–Robot Interaction and Collaborative Intelligence

Human–Robot Interaction (HRI) is one of the areas where LLMs have produced the most transformative impact. By converting complex human instructions into robot behavior, LLMs eliminate the need for expert programming and enable natural, conversational interfaces.

Several research threads highlight this transformation:

Conversational Task Specification

Systems like ChatGPT-based robotic interfaces allow users to specify high-level tasks in free-form language—for example:
“Clean the table and put the dishes into the sink.”

LLMs interpret such tasks into sequences involving navigation, manipulation, and object sorting [1…].

Interactive Feedback and Correction

ROS-LLM enables iterative human feedback loops:

- The user observes the robot's behavior.
- Provides corrective natural language feedback.
- The system updates the plan and learns from the interaction [7…].

This shift toward continual learning mirrors human teaching paradigms.

Commonsense and Emotional Understanding

NL-based execution systems increasingly extract implicit intentions from statements such as "I'm thirsty," demonstrating emotional and contextual awareness [3…].

Multimodal Interaction

Some frameworks integrate text with images or scene observations, enabling users to guide robots through instructions like "move the cup next to this" while pointing at an object.

Collaborative Agents

Research on generative agents enables synthetic memory and long-term planning, allowing robots to develop richer models of user preferences [1…].

LLMs enable HRI that is intuitive, adaptive, and aligned with human expectations—critical for widespread deployment in homes, hospitals, and collaborative industrial settings.

2.12. Limitations, Open Challenges, and Future Directions

Despite substantial progress, LLM-enabled robotics faces several limitations.

1. Grounding and Symbolic Alignment

LLMs often hallucinate unrealistic actions, especially when lacking grounded sensory input. Systems like ARRC and SayCan mitigate this, but full grounding across all modalities remains an open challenge.

2. Numerical Precision and Physical Constraints

LLMs excel in semantics but struggle with numerical predictions, spatial geometry, and continuous control—hence the need for frameworks like RT-Grasp [5.RT-Grasp…].

3. Data Requirements and Domain Transfer

Industrial fine-tuning, as shown in domain-specific robot programming studies, remains resource-intensive despite innovations like QLoRA [4.Domain-Specific Fine-Tuning…].

4. Safety and Verification

Ensuring physical safety requires strict validation pipelines. LLMs alone cannot guarantee safety due to their probabilistic nature.

5. Real-World Generalization

Robots often perform worse in real-world conditions than in simulation-driven benchmarks. Multimodal fine-tuning helps but does not fully close the gap.

6. Latency and Real-Time Constraints

LLM inference may introduce delays incompatible with time-sensitive tasks. On-device acceleration and smaller specialized models may help.

7. Long-Horizon Task Performance

Plans involving dozens of steps remain difficult due to compounding reasoning errors and context window limitations.

8. Ethical and Societal Concerns

As robots gain autonomy, questions about accountability, privacy, employment, and human safety become increasingly relevant.

Despite these challenges, the trajectory is clear: LLM-enabled embodied intelligence is rapidly becoming the new paradigm in robotics. The fusion of language, perception, reasoning, and control promises unprecedented levels of adaptability, usability, and intelligence.

3. Conclusion

The rapid emergence of Large Language Models has reshaped the foundations of modern robotics, enabling unprecedented levels of abstraction, reasoning, and naturalistic interaction between humans and autonomous systems. The literature reviewed—spanning multimodal task planning, reasoning-based manipulation, industrial fine-tuning, retrieval-augmented control, and ROS-integrated frameworks—demonstrates that LLMs are no longer auxiliary components but have become central cognitive engines that unify perception, decision-making, control, and interaction. Across all the domains discussed, a consistent trend emerges: LLMs bridge the gap between symbolic reasoning and embodied action, allowing robots to translate human intent directly into physical behavior.

A major theme throughout the reviewed works is the evolution from rule-based natural language processing toward rich semantic understanding. Early ontology-driven systems [2.Parsing Natural Language Sentences into Robot Actions] constrained robots to limited sets of actions and grammar, restricting generality. Contemporary LLMs, with their embedded world knowledge and commonsense reasoning, enable robots to interpret abstract, indirect, or underspecified commands, infer user intentions, and adapt to environmental changes dynamically. This marks a decisive shift from rigid instruction-following toward cooperative, cognitively aligned assistants.

Task planning—one of the earliest areas of LLM adoption—has matured into sophisticated pipelines combining classical planning methods, probabilistic grounding, and LLM-based semantic reasoning [1.Large Language Models for Robotics: Opportunities, Challenges, and Perspectives]. Robots now generate long-horizon, context-aware sequences that account for preconditions, constraints, and spatial relationships. The integration of multimodal cues further enhances this, proving that vision-language models and LLMs provide complementary strengths for grounded planning.

Manipulation, previously dominated by geometric and numerical models, is now augmented by semantic reasoning through frameworks like VIMA, LLM-GROP, and RT-Grasp [5.RT-Grasp…]. These systems demonstrate that language-guided understanding of affordances, safety constraints, and object properties can dramatically improve the adaptability and generality of robotic manipulation skills. Reasoning templates and multimodal fine-tuning also reveal a promising direction: combining structured language-derived reasoning with precise numerical prediction for safe, robust physical execution.

Modular approaches, including PROGRAMPORT and MetaMorph, highlight how LLMs support scalable, compositional robotics. Their ability to translate linguistic descriptions into reusable program blocks or morphology-conditioned policies accelerates the development of general-purpose robotic systems. Retrieval-Augmented Generation (RAG), exemplified by ARRC [6.ARRC…], further shows that grounding LLMs in domain-specific knowledge bases significantly improves reliability, safety, and contextual relevance—addressing one of the most persistent concerns in LLM-based robotics: hallucination.

Human–robot interaction stands out as one of the most transformative domains affected by LLMs. Through systems like ROS-LLM [7.ROS-LLM…], natural language becomes a first-class interface that allows non-experts to program, control, correct, and collaborate with robots intuitively. Interactive strategies, multimodal feedback mechanisms, and memory-augmented agents push robotic systems toward increasingly human-centered design philosophies.

Despite these advancements, substantial challenges remain. LLMs struggle with grounding precision, continuous control, real-world generalization, and long-horizon reliability. Their probabilistic nature introduces safety concerns that necessitate layered safeguards, symbolic validators, and conservative control policies. Domain-specific fine-tuning is still expensive for many industrial applications, and the integration of multimodal sensory input with language-based reasoning remains an open problem. Finally, ethical considerations—privacy, accountability, agency, and labor impacts—must be addressed as robots become increasingly autonomous and cognitively capable.

Looking forward, the convergence of multimodal LLMs, real-time perception, retrieval-enhanced reasoning, and physically aware controllers promises a new era of embodied intelligence. As LLMs evolve, we anticipate robots that not only execute tasks but understand goals, anticipate needs, and collaborate seamlessly with humans across diverse environments. The trajectory described in this paper suggests that LLM-enabled robotics is transitioning from research novelty to an emerging backbone of real-world autonomous systems. Continued progress in grounding, safety, and multimodal integration will be crucial for unlocking the full potential of language-driven embodied intelligence.

---

4. References

Below are references **following your required format**, pointing to the specific paper identifiers you provided:

1. **Large Language Models for Robotics: Opportunities, Challenges, and Perspectives**
2. **Parsing Natural Language Sentences into Robot Actions**
3. **A Review of Natural-Language-Instructed Robot Execution Systems**
4. **Domain-Specific Fine-Tuning of Large Language Models for Interactive Robot Programming**
5. **RT-Grasp: Reasoning Tuning Robotic Grasping via Multi-modal Large Language Model**
6. **ARRC: Advanced Reasoning Robot Control—Knowledge-Driven Autonomous Manipulation Using Retrieval-Augmented Generation**
7. **ROS-LLM: A ROS Framework for Embodied AI with Task Feedback and Structured Reasoning**
8. **Large Language Models for Robotics: A Survey**