Voice Powered Task Automation Using AI

Smit Makodia[1], Darshana Patel[2], Aditiba Jadeja[3], Soniya Aghera[4], Darshana Bhatti[5], Vijaysinh Jadeja[6]

Department of Information Technology[1234], V.V.P. Engineering College, Rajkot, India

Department of Chemical Engineering[5], V.V.P Engineering college, Rajkot, India

Department of Computer Engineering[6], Sal college of Engineering, Ahmedabad, India

Abstract

In many developing regions, including India, the demand for scalable and intuitive automation solutions is growing rapidly. Traditional robotic systems often require complex programming knowledge, creating a barrier for non-expert users. In response to the growing need for scalable, voice-driven automation, a platform named CATALYST needs to be considered. It serves as a voice-to-action bridge, acting as a smart link and translating natural language commands into executable robotic actions. It completely gets rid of the complex programming setups by employing user spoken commands. Without any training, the system changes the human language to the sequential robotic executions. One step further in this direction is the LLM (Large Language Model) processing, speech to text services, and n8n (workflow automation) which altogether make the just mentioned system. The demonstration of the system was the use of a virtual car which was under the control of a virtual microcontroller. The research results state that CATALYST is able to perform prompt planning and task processing, having short voice processing times as well as robot execution times being of a predictable nature. This development can make the interaction between humans and robots more user-friendly and time-saving in different sectors.

Keywords: Automation, Conversational AI, Embodied AI, Human-Robot Interaction, Speech-to-Text (ASR)

Introduction

Human-computer interaction (HCI) landscape has been radically transformed by the introduction of conversational AI systems. Such systems are capable of understanding, processing, and executing natural language commands. Despite remarkable progress in AI, robotic automation is still constrained by the requirement for specially made software or complicated interfaces, which amateurs find it hard to access.[1,2]The CATALYST is a solution that takes voice commands in natural language from users, converts the words into robotics actions, and then shows them the desired output. In this way, robotics become more simple, efficient, and friendly to the user thus increasing the possibility of their application in different fields.[3,4]The production of robotic systems tasked with performing routine work remains, however, limited by the necessity of professionals setting up specific programming to do it. This in turn makes the technology inaccessible to most individuals, thus, its use would be limited to only a few

experts such as engineers. The problem is known as "last mile" in the adoption of robots.[5,6]The real problem is the humans have to change to suit the demands of machines instead of machines adapting to normal human communication. For robots to be really common and handy in our daily living, a change from the engineer-centric to the human-centric interfaces is necessary. This would grant robots the ability to become companions that talk with us in ordinary language and comprehend as well as respond.[7,8]

CATALYST is a robotic system that relies on an elaborate design to turn human-spoken instructions into the desired robot actions. Basically, it employs a Large Language Model (LLM) for semantic understanding, an automatic speech recognition system for transcription of spoken words into text, and a workflow automation platform named n8n to manage and coordinate with the robot control system.[9-11] With this system in place, it is referred to as "zero-shot" task automation, the system is able to perform new tasks without the need of pre-programming.[12-14] LLM is the most advanced cognitive planner which is able to break down complex tasks to the point where the robot only executes a series of simple commands. At the same time n8n plays the role of the connecting link, the bridge between high-level intelligence of the LLM and low-level module of robot which gives the feature of speedy prototyping and hence scalability across various hardware and software platforms.[7,8]

Objectives

To overcome the limitations of conventional robotic interfaces by designing a universal control system driven by natural, conversational language. The aim is to eliminate the need for technical or programming knowledge, thereby empowering non-programmers to intuitively command and execute complex, multi-step robotic tasks. This approach significantly broadens access to automation technologies, especially in contexts where technical expertise is a barrier, such as in education, agriculture, or home automation.

To establish that integrating large language models (LLMs) with voice interfaces enables robots to dynamically interpret and adapt to high-level user instructions. Unlike rigid, pre-programmed systems, the proposed solution supports real-time decision-making, task adaptation, and self-correction — even in the presence of ambiguous or variable input. This addresses one of the core limitations of traditional automation systems: the inability to respond flexibly to unstructured or changing environments.

To design and implement a modular system architecture that connects LLMs with a workflow automation engine (n8n), facilitating the translation of natural human intent into precise robotic actions. This architecture is demonstrated through a virtual robotic platform — specifically, a virtual car controlled via a virtual microcontroller — and is evaluated for performance metrics such as voice command latency, execution confirmation, and robustness. By releasing the system as an open-source project, this work also lays the foundation for future advancements in embodied AI, smart environments, and assistive robotics.

Applications

Voice-powered automation has been implemented in a variety of sectors to the extent that the main factor for this phenomenon is the improvement of the involved technologies in speech recognition and the rising demand for an environment where the hands-free operation is possible. In the smart home environment, the voice assistants can in fact be said to have already become the consumers' favourites as the prediction of the total market for smart home is approximately $537 billion by 2030. The systems described have been developed to carry out services such as switching the lights on and off, or controlling the security and entertainment devices through the natural language command and thus form a basis for more complex applications.[1]

It has been observed that the nature of the voice control technology embedded in smart home devices like Amazon Alexa and Google Home is such that it suffers from the lack of capability to deal with complex multi-step commands. The technology depends on strict and predetermined rules and most times struggles to recognize the subtleties of human language. Besides the highly efficient industrial automation, it is still prone to being very stiff and unchanging in nature. Such type of industry includes "dark factories"[15] and "ballet parking"[16], where the systems are solely implemented for specific task repetition and more than enough pre-programming is needed. This is totally different from that of flexible automation which is able to adapt to unplanned human commands. The LLM-robotics field is a new solution for the problem. Large language models can close the gap between fixed and flexible automations, which is the essence of CATALYST, by going from the keyword-driven approach to the intent-driven understanding and this allows complex, situational instructions to be deciphered without any previous programming.

Healthcare is a very prospective domain for the implementation of voice-controlled automation. In a surgical setting, the technology may let the surgeon handle the machinery as well as gain access to the patient data all the while maintaining sterile conditions. Besides this, it can be utilized for the monitoring of patients at a distance and for telemedicine, the development of which has been quickened by the COVID-19 pandemic. Voice-operated medical instruments can cut down on procedure time by up to 30% and their precision can be improved as well [17-20]. Robots powered by CATALYST can also come to the help of medicine delivery, supervise from a distance and provide physical, cognitive and social support to patients and people with disabilities thus facilitating their movement towards being self-reliant.[21]

In the field of industrial automation and manufacturing, voice-controlled systems allow workers to have hands-free operation. They can control robotic arms or adjust parameters simply by giving voice commands. These systems have been installed in companies such as BMW and Ford, which has resulted in the production efficiency rising by 25%.[1,2,22]In the case of compact warehouses, CATALYST-powered robots are permitted to pick up, carry, and arrange products, which can greatly reduce the amount of

labor-intensive tasks. The system can also be utilized in making DIY construction easier and safer through methods such as brick laying and tool organizing. Skilled labor becomes safer and more accessible due to the help of the system.If we are talking about autonomous vehicles and transportation, voice assistants might not only be the source of entertainment but could also take control over driving modes, vehicle-to-vehicle communication, and dynamic route optimization. Voice interfaces could be implemented in advanced driver assistance systems (ADAS) for safety-critical functions, thus keeping the driver vigilant.[6,23]

The technology of speech recognition in numerous languages has played a pivotal role in making these innovations more acceptive of a wider audience. The target market for the voice-dependent technologies includes the disabled population and speech recognition potentially is the way for the differently abled to become more independent. Applications for voice-command wheelchairs, prosthetic devices, and intelligent home products can all be part of a package for voice-controlled accessibility and assistive technologies. There have been improvements made in the field of multilingual speech recognition, which have had a vast impact on the inclusiveness of the abovementioned technologies.[4,24-26] CATALYST is the hallmark of assistive living as it eases elderly and differently-abled people into the comfort of self-reliance and continuous enhancement of the quality of life.

One can only guess the future applications of such technology may span space exploration, emergency response, and education.[27,28]For example, voice-controlled drones and robots might still be used in dangerous locations but for the relief of victims of natural disasters. It can also be a living demonstration of AI for students to learn the way just by seeing the behavior of the robotics language-translated. Its voice interface can easily operate agricultural devices in remote areas, making it convenient for farmers and raising productivity.[5,14,29,30]

The merging of 5G, edge computing, and advanced AI technology are opening the door for on-the-spot voice-controlled systems to exist in very difficult areas.[6,23,31]The next stage of multimodal AI could be where it not only uses voice but also gestures and computer vision making the whole human-robot interaction more natural and straightforward. The potential of CATALYST in cultural preservation is equally promising, as verbal demonstrations of traditional skills by elderly community members can provide learning material for AI robots, thus, safeguarding non-physical heritage.[32]

Literature Review / Related Work

The table consolidates five representative methods in language-guided robotics, aligning method descriptions with outcomes, strengths, limitations, and future scope to support a structured comparison of design trade-offs in embodied instruction following. Entries emphasize how spatial mapping, modular

policies, LLM prompting, memory architectures, and re-planning influence ALFRED performance and generalization in unseen environments.

Table 1: Comparative Literature Review of Instruction-Following Methods

(Structured summary across five methods: a persistent spatial semantic mapping approach (HLSM), FILM's modular mapping with semantic search, Prompter's LLM-based planning with data efficiency, CAPEAM's context-aware planning and memory, and RoboGPT's LLM planner with re-planning and skills; columns cover Method, Result, Strength, Weakness, Future Scope, and Discussion, enabling rapid cross-paper analysis of performance and design constraints.)

Ref.	Method	Result	Strength	Weakness	Future Scope	Discussion
A[33]	Introduces an always-on 3D spatial semantic voxel map & with a hierarchical controller for instruction execution.	Set a new record, going beyond the previous results for the high-context instructions without guidance by steps.	Transparent spatial semantic maps enable complex, long-term, stored reasoning and along with efficient selecting subgoals.	However, long-horizon exploration and perception still limit system performance. The low-level controller is not designed for physical robots.	Need to extend to real robot, joint end-to-end reinforcement learning for control, and bridging the sim-to-real gap.	Acknowledges the semantic representation significance in the robot behavior conditioned on language and generalizable.
B[34]	It is a modular approach (FILM) that reconstructs a semantic top-down scene map while converting the	Set a new record on ALFRED's high-level instruction task. The semantic searching policy made exploration better.	It lowers the necessity of the training data and is resistant to the deficiency of the expert supervision. It generalizes	Nonetheless, it is still unable to detect tiny or hidden objects correctly and heavily depends on the pre-trained segmentation	Perception and locating more active objects, deploying on real robots for sim-to-real transfer, and better-designed FILM to	The success of FILM with modular memory and semantic policies validates CATALYST's strategy of decoupling language

	language to structured subtasks with the help of BERT.		well to unfamiliar scenes.	to function.	tackle these challenges.	processing from execution.
C[35]	By using prompt engineering on LLMs to come up with the most efficient plans for embodied agents from very small annotated datasets.	By utilizing pretrained LLMs for task decomposition, achieved similar or better results with drastically less training data.	By being able to adapt easily to new tasks and decrease the amount of manual labeling, this research shows the advantage of using powerful prompted LLM priors.	Does not have perceptual grounding and fully relies on the accuracy of scene models or oracles. The problem of deploying on a real robot is not handled.	First, they wish to closely couple visual perception with the help of embarking on a truly autonomous grounding mission so that annotation needs can be drastically lowered.	The study here serves as the final experimental validation of CATALYST's core claim regarding LLMs as zero-shot planners, if given ample contextual information.
D[36]	It proposes a flexible memory design (context-aware planning and environment-aware memory) to match up instructions	Better precision and transfer to new, complex, multi-step tasks as compared to transformer-only baselines.	Context is explicitly modeled for the environment, thereby allowing the understanding of ambiguous instructions. The use of dynamic	Really computationally expensive due to the dual memory updates. There is no proof of sim-to-real robustness.	Memory abstraction for different kinds of environments and policy distillation to smaller agents to deploy there.	One of the reasons why an environment-aware memory is still relevant for future versions of CATALYST is because feedback

	with scene semantics.		memory makes long-term planning possible.			from the environment can now be used to replan.
E[37]	RoboGPT is the idea that merges an LLM-based planner, navigation skill modules, and an adaptive re-planning strategy.	Compared to other LLM and template-based planners on ALFRED, it did better and was able to handle object dependencies and dynamic environment mapping with ease.	After solving the problem of missing targets by using semantic maps for re-planning, the system also adjusts to the terms of the scene to overcome this issue.	Depend heavily on very complete embodied datasets, as well as on accurate mapping. Failures to manipulation may lead to vulnerabilities.	Some possible uses of the computer: Improving real-time visual grounding, reducing errors in the generation of data, and exploring scaled robotic deployment.	RoboGPT's achievement in the area of re-planning using environmental feedback is an excellent example of the directions towards better CATALYST impacts.
F[27]	Delves into the use of large language models to plan the sequences of a robot's actions in zero-shot manipulation tasks, with emphasis on	They have proven that without any kind of specific training, LLMs designed for general purposes can come up with the right robot	Such a method is very adaptable and can change tools and tasks just by adding new function definitions to the prompt.	The system performance depends largely on the accuracy and richness of the prompt. There is no mechanism for the system to correct its errors through feedback.	One way to solve this problem is by giving the robot's eyes (LLM) access to the visual feedback so they can figure out the correct plan and also	The approach taken in this paper is very compatible with CATALYST, giving the nod to LLMs as one of the "translators" moving from human intent to machine

	the conversion of natural language instructions into scripts that can be executed by a robot.	control scripts for a wide range of manipulation tasks.			creating stronger ways to link spoken words with the real world.	code."

The following figure benchmarks instruction-following systems on ALFRED Test-Unseen under high-level-only instructions, providing a like-for-like view of generalization without low-level supervision bias. It highlights the progression from modular mapping-centric methods to LLM-based planning with feedback, clarifying how architectural shifts translate into success-rate gains in unseen

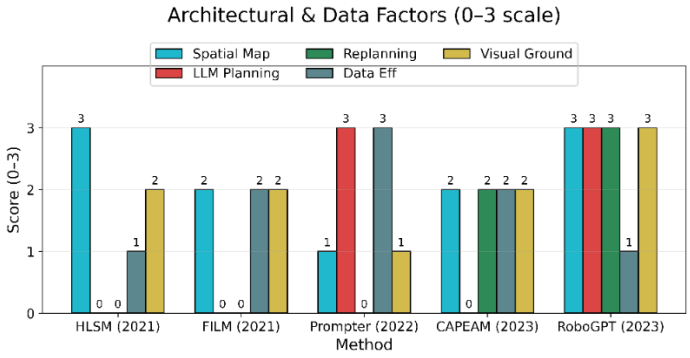

Figure 1: ***Architectural & Data Factors (0–3 scale)***

(Grouped bar chart comparing five methods—HLSM (2021), FILM (2021), Prompter (2022), CAPEAM (2023), RoboGPT (2023)—across Spatial Map, Replanning, Visual Grounding, LLM Planning, and Data Efficiency, showing RoboGPT leading with consistent 3s, Prompter excelling in LLM Planning, and earlier methods scoring lower or zero on several factors.)

scenes.

Proposed Workflow

The CATALYST system is an elaborate design that changes conversational voice instructions of human language to execute robotic actions. Its modular setup combines numerous cutting-edge technologies, such as internet-based APIs, process automation tools, and digital practice spaces. Such an architectural style endows the system with qualities of strength, expandability, and simplicity of upkeep while achieving the accurate execution of user commands by the robot.[7,9,11]

Basic Workflow

The CATALYST workflow is a control system of the closed-loop type, which keeps checking whether the commands given have been carried out successfully, thus enabling adaptive responses.

1. **User Speaks Command (A):** The process is initiated by a natural language voice instruction, which is the main interface.[9]
2. **Speech-to-Text (ASR) (B):** The spoken command is converted into a text format by an Automatic Speech Recognition (ASR) service.[10,12,38]
3. **Intent Extraction & NLP (C):** A Natural Language Processing (NLP) module looks at the text to find the user's main point and any additional data that might be related.[7,8,39]
4. **Task Decomposition & Planning (D)**: The command of the high-level is divided into a logical procession of discrete sub-tasks which results in a preliminary task plan forming.[7,40,41]
5. **Map Actions to Robot Commands (E):** The sub-tasks are converted into low-level, specific robot control commands, such as motor control parameters, that the robot's microcontroller can understand.[13,14]
6. **Arduino/Motor Control Execution (F):** The commands are carried out by a virtual microcontroller (simulating an Arduino) in a simulated environment, thus the robot's motors and actuators are activated.[13,14,29]
7. **Environment/Robot Feedback (G):** The robot is constantly receiving the data from the real-time sensors (e.g., virtual encoders, collision detection) of its virtual environment.[6,23]
8. **Task Success / Notify User (H):** When the job is done as expected, the system goes through the completion phase and it can give a visual or audible signal to the user.
9. **Error/Obstacle Detected (I):** If the feedback shows that the present state is not the one predicted, for example, an obstacle or execution error, the system will be at failure.[7,37]
10. **Replanning or User Clarification (J):** The system is thus getting ready for a recovery phase when deciding upon an error. It can either re-plan the task automatically or indicate to the user that they need to clarify by returning to the earlier stages to improve the plan. This feedback loop equips the system to be tough and to allow for smart

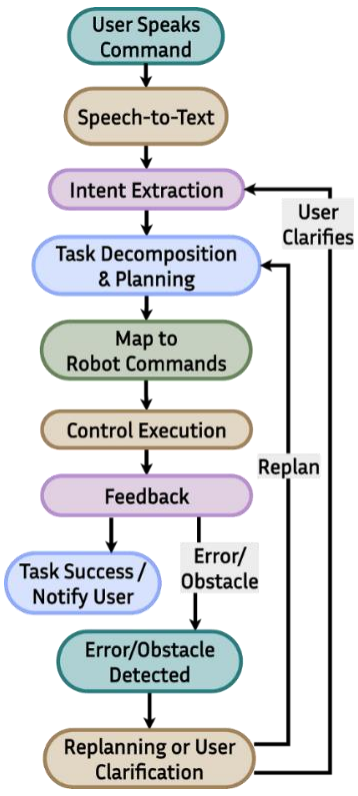

Figure 2: ***CATALYST Basic Closed-Loop Workflow***

(Block diagram from "User Speaks Command" to "Replanning or User Clarification," showing ASR, intent extraction, task decomposition and planning, mapping to robot commands, control execution, feedback, success path, and recovery paths via error/obstacle detection with user clarification or

coordination.

Advanced Workflow

While the basic workflow gives a broad picture of the operational loop, the advanced block diagram reveals the intricate cognitive processes of CATALYST, showing in particular how it relies on LLMs for

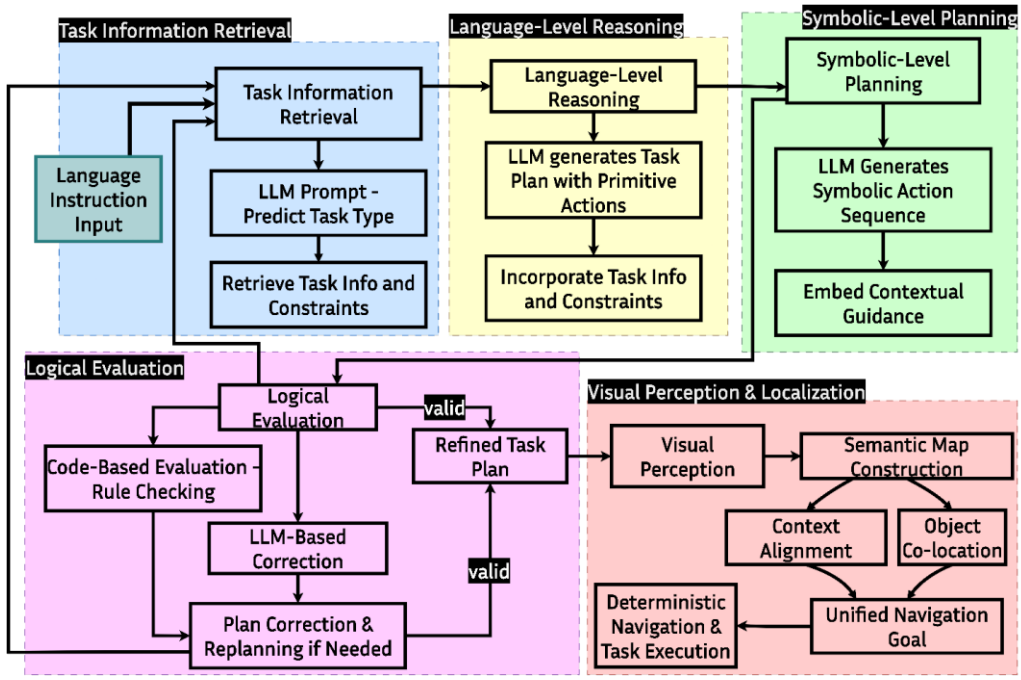

Figure 3: ***Advanced Reasoning-to-Execution Workflow***

(Block diagram with five coordinated stages: 1) Task Information Retrieval via LLM prompting and constraint fetch; 2) Language-Level Reasoning where the LLM decomposes the instruction into primitive actions; 3) Symbolic-Level Planning that converts steps into a symbolic action sequence with contextual guidance; 4) Logical Evaluation that performs rule checking and LLM-based correction, producing a refined, valid task plan; 5) Visual Perception & Localization that builds a semantic map, aligns context, co-locates objects, sets a unified navigation goal, and triggers deterministic navigation and task execution.)

multi-level planning, brings in perception, and carries out thorough logical validation for achieving dependable, grounded robotic actions.[44]

1. **Language Instruction Input (A):** This is the point where the user, through a transcription of the spoken voice into text, gives a natural language instruction, which is the norm.
2. **Task Information Retrieval (B):** After getting the text instruction, the system will activate a Large Language Model (LLM) to start the task information retrieval process. This task will consist in instructing the LLM to foresee the kind of task involved and then gather such contextual information, constraints, and operational rules on the task from the knowledge base as are relevant to that task.
3. **Language-Level Reasoning (C):** The task information fetched is then presented to the LLM for language-level reasoning. In this step, the LLM turns the high-level natural language command into a low-level one, essentially a task plan, and it also describes the task as a series of primitive

actions. The focus of this phase is on ensuring that the logical order and the purpose of the users request have been clearly implied.

4. **Symbolic-Level Planning (D):** The LLM after the language-level reasoning turns these plain-language steps into a collection of symbolic actions (e.g., "pick up," "place," "toggle") that correspond to the sequence of the first concept of one additive language. At this point, it also involves inserting any directional or partial assistance that might be necessary when localizing and manipulating the object in the physical world for the accuracy of the work.
5. **Logical Evaluation (E):** At first, the symbolic action sequences undergo strict logical evaluation. This is rule-based evaluation, code-based which besides proper rule-checking against given constraints also allows for using correction algorithms. In the event the plan is invalid or the desired changes are to be made, the system will not accomplish the task immediately. Instead, it will facilitate the re-planning procedure, which thereby is going to back down to the "Task Information Retrieval" stage (B) to recheck and prepare an updated plan.
6. **Refined Task Plan (F):** When logical evaluation accompanied by all other checkings of validity and executability of the plan takes place, what is then produced is the Refined Task Plan, entailing the first plan with robust, semantically upright, and physically realizable one.
7. **Visual Perception (G):** A computer vision unit will be launched by the system that already has a refined task plan. This step is the application of depth and vision sensors that allows the gathering of up-to-the-minute data about the robots direct surroundings.
8. **Semantic Map Construction (H**): The scene’s semantic map constructed from the raw visual data details. This map includes objects, their properties, and the relationships between the spatial locations, hence, is a memory tensor of the environment.
9. **Object Co-location (I) & Context Alignment (J):** The system identifies the objects that are nearest based on the semantic map, and it also estimates the exact locations of those objects. The former activity thus finds the precise locations of the target objects from the map, while the latter helps in the fine-grained localization aspect by using the context or the original instruction (e.g., "on the tub") of the object to be placed and the relationships with other objects to be understood.
10. **Unified Navigation Goal (K):** The data from object co-location and context alignment are melded together to form the "Unified Navigation Goal." This aim merges probabilistic maps and semantic cues to set up safe and context-aware navigation destinations for the robot.
11. **Task Execution (L):** At last the robot executes the planned sequence of actions.

Implementation

One of the most advanced implementations is the CATALYST system which combines cloud-based APIs, workflow automation platforms, and virtual simulation environments. To help with scalability, the system has been designed with a microservices architecture that not only supports the up-time or availability of services and complex error handling but also leads to reduced latency and even increased reliability.[9,11]

Technology Stack and Architecture

The CATALYST system is a complex microservices-based design that practically links cloud-based APIs, a workflow automation platform, and a virtual simulation setting to maintain scalability and reliability.[9,11]

Google Gemini 2.5 Flash serves as the system's natural language processing and code generation unit. It is an extremely compact and efficient model that makes it very suited to understanding complex instructions and providing Arduino-compatible JavaScript code on-the-fly in real time.[7,8,35,43-45]

The real-time SSH Bridge constructed on Node.js offers a real-time voice channel for the N8N workflow and the virtual Arduino environment. This mechanism provides the promptest delivery of messages, and at the same time guarantees the loose coupling between parts of the system resulting in the good responsiveness of the system.

Figure 4: Voice input

(Image that shows the front-end/webpage for the voice recording and passing via an API. Also, has additional facility to directly pass the command as text, allows for testing purposes for the workflow testing and developing.)

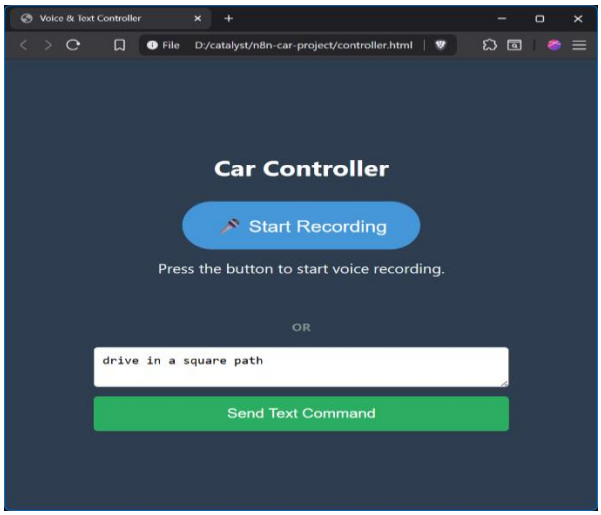

Figure 5: Simulator

(Image that shows the front-end/webpage for the virtually simulated car and the Arduino. Also, has terminal where we can see the status as well as the commands that are running. Pin state shows the live Arduino pin state working at each command run time.)

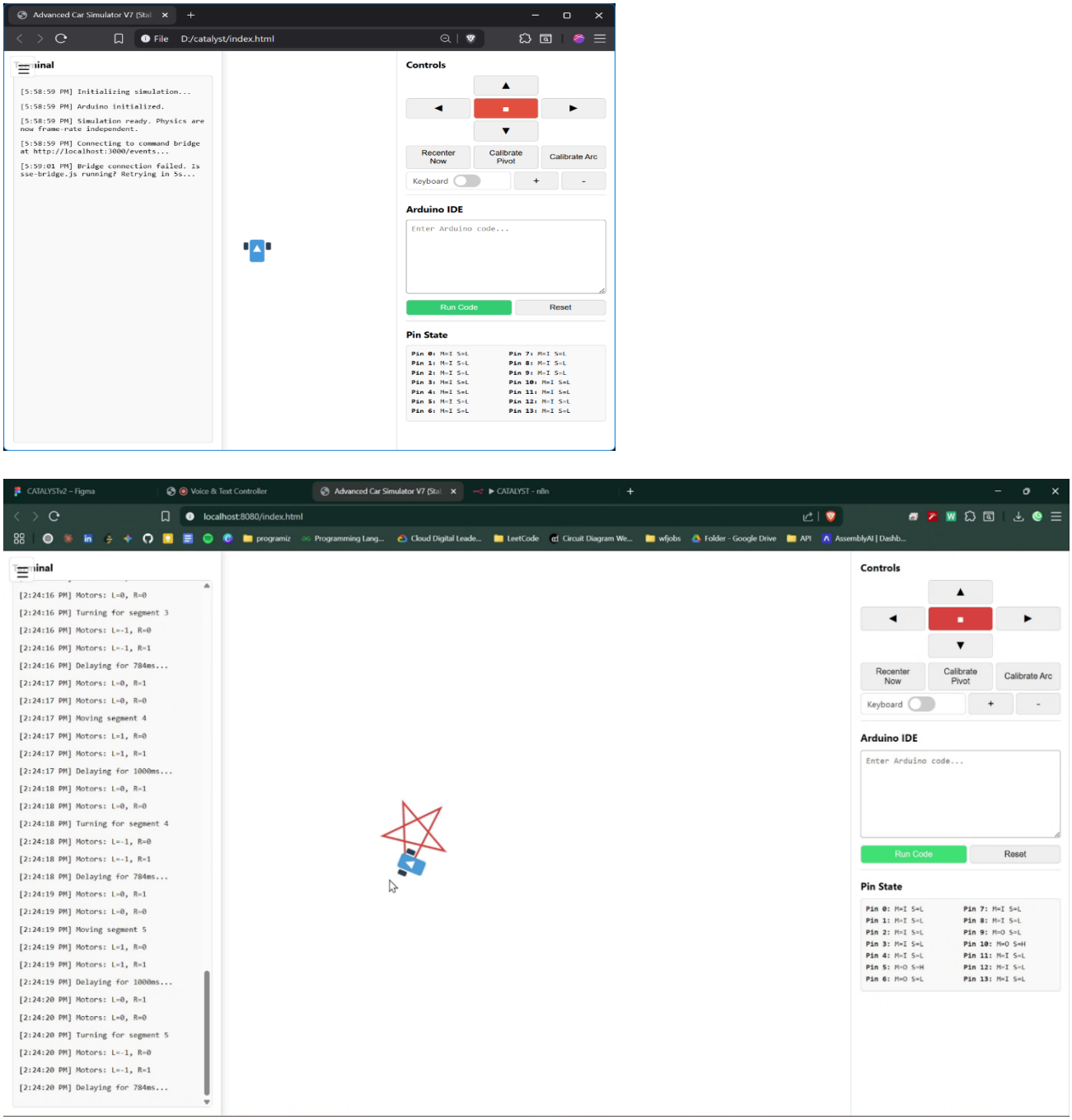

Figure 6: End result or the output for CATALYST in action

(Image that shows the output of the complete execution. All commands generated using AI can be seen in action inside the terminal on the left.)

Back-End (SSE Bridge) Stack and Techniques:

The back-end is a minimalistic Node.js/Express server with a CORS middleware. The main idea behind the back-end command distribution is Server-Sent Events (SSE) via a /events endpoint which is a very productive approach in a unidirectional streaming scenario. The /command endpoint runs separate process that takes JSON code snippets from n8n, validates, and then sends them to all connected simulators.

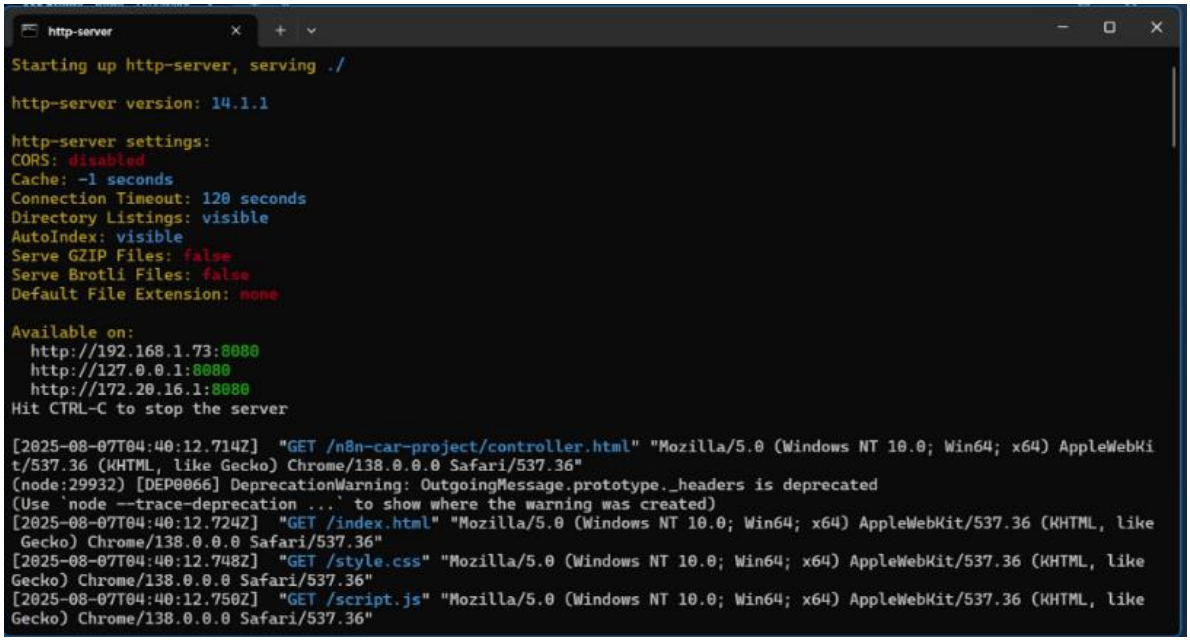

Figure 7: ***Local bridge connection***

(Image that shows the locally hosted webpages on node.js server via SSE bridge.)

n8n workflow is pretty much the producer of the whole sequence. It first sets up a Webhook node that collects binary audio data. Then AssemblyAI API is used for transcription from the audio. Finally, the transcribed text is passed to a "Basic LLM Chain" node having Google Gemini 2.5 Flash as the LLM engine. The LLM's prompt that is used here, strictly instructs the AI to output only one line of executable JS for the virtual Arduino. Included in the prompt are very concrete timing constants (e.g., pivot turn 90°=490ms) and a small fixed set of APIs and motor control pins that come from the training data thus, the model can provide the exact and instantly executable output.

Detailed User Interaction Flow

The user interaction with CATALYST is methodical and unambiguous, performed through clear, step-by-step procedures:

1. **User describes a desired action through speech**: The user by way of talking in natural language gives the instruction (e.g., "Move forward 20 centimeters" or "Draw a square").
2. **Voice is gotten, and it is delivered to the speech-to-text service:** The voice command is passed on to AssemblyAI, the Automatic Speech Recognition (ASR) service for transcription.
3. **The text command is handled by n8n:** Here the transcription is received by n8n workflow, which is the central orchestrator.
4. **n8n gets this command and shares it with the LLM**: The n8n workflow, through an API call, transfers the text command to the Google Gemini LLM.

5. **The LLM takes the natural language input and returns a structured command:** The LLM figures out the user's need and by a strict prompt, it prepares a single-line JavaScript program as the output command which is the executable for the virtual microcontroller.

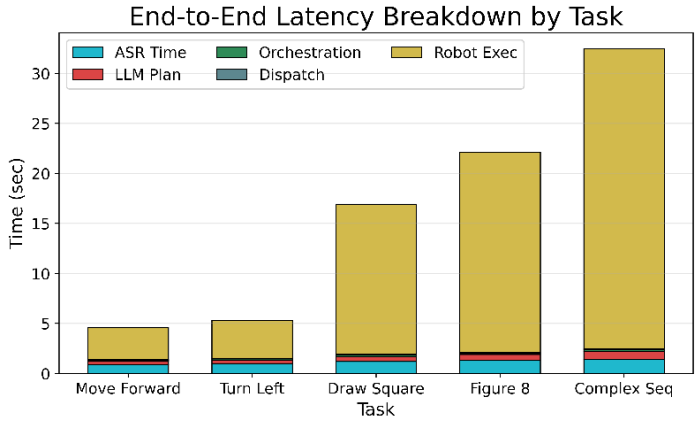

Figure 6: ***End-to-End Latency Breakdown by Task***

(Stacked bar chart comparing total latency for Move Forward, Turn Left, Draw Square, Figure 8, and Complex Sequence, decomposed into ASR Time, LLM Plan, Orchestration, Dispatch, and Robot Execution; cognitive/coordination overheads remain small and flat, while Robot Execution increases with task complexity.)

6. **n8n converts this one for the virtual microcontroller:** The n8n workflow makes sure that the output from the LLM is in a directly executable form and sends it to the simulator.
7. **The virtual car moves:** The JavaScript command is fired through the SSE bridge to the simulator where the virtual Arduino executes the code that makes the virtual car move as per the original voice command and thus visualizes the path.

Output of the Working Model

The CATALYST model visually and numerically reveals its effectiveness.

End-to-end latency

This stacked chart breaks total delay from microphone to motion into distinct stages so the composition of overall delay is visible at a glance.End-to-end latency is a delay within which a user trigger is followed by a completed response across all network and processing steps, not just one subsystem.

Stage definitions

- ASR: Automatic Speech Recognition takes speech as input and produces a text file as output which is the next step in the pipeline.
- LLM planning: Large Language Models turn the given text command into a stepwise task plan or action sequence for the robot.

- Orchestration: Managing multiple automated tasks and services into one combined workflow, thus getting the right step executed at the right time.[48]
- Dispatch: The process of giving an agent or a device that is poised and ready to start working on a prepared job/command.
- On-robot execution: Time the robot spends doing the commanded motions and actions by its controller until the time that it is over.

LLM latency by complexity

With this grouped chart, the comparison of Gemini 2.5 Flash[45], Gemini 2.0, and ChatGPT response times for simple, moderate, and complex prompts is depicted. Latency is used in the comparison as the delay between request and model reply.“Complexity” here indicates that harder prompts are the ones which usually require more reasoning or planning, thus extending model processing time and as a result the latency.[9,50-53]

Metric definitions:

- Latency: The time delay from the initiation of a request to the receipt of a response, which in interactive systems is often represented in milliseconds.
- Median: The center value when data is sorted, which is a summary of a typical value less affected by extreme values than the average.
- Instruction translation accuracy
- The information presented in the grouped chart conveys the frequency of each model in converting a user’;s instruction into a comprehendible, executable plan of dealing with simple, multi-step, and ambiguous categories.
- Accuracy means the rate of correctly predicted cases out of all the predictions made, which is a standard classification metric summarizing the degree of success in instruction-following.

Command categories

- Simple: Goals are stated in a one, clear and simple manner, easy to be decomposed in actions.
- Multi-step: The instructions that need to be carried out only after the sequence of several sub-tasks is completed in the correct order.

- Ambiguous: Commands that are not provided with key details or context, which require disambiguation or clarification to be able to recognize the correct way of functioning.

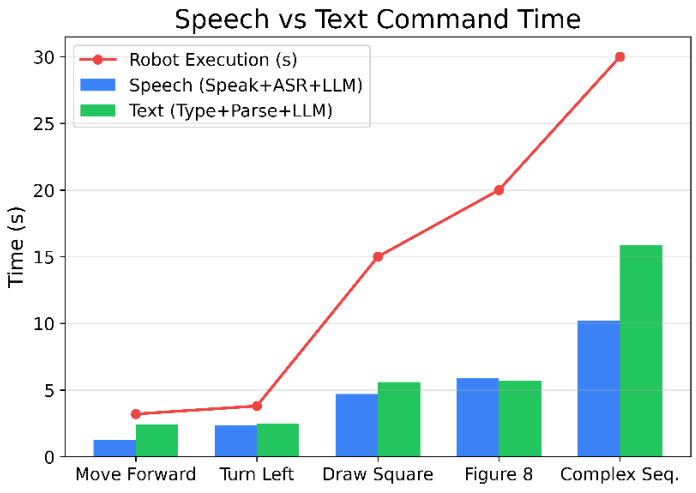

Figure 7: ***Speech vs Text Command Time***

(Line-and-bar composite comparing robot execution time (red curve) with Speech pipeline time (Speak+ASR+LLM, blue bars) and Text pipeline time (Type+Parse+LLM, green bars) for Move Forward, Turn Left, Draw Square, Figure 8, and Complex Sequence; interaction overheads are low and relatively flat, while execution grows with task complexity.)

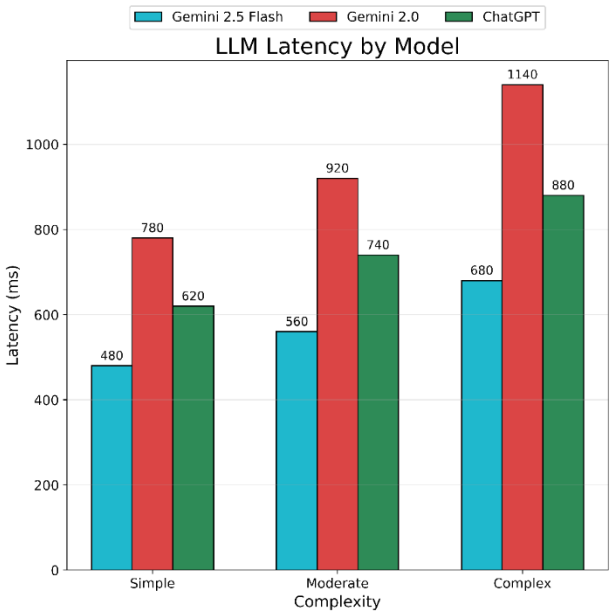

Figure 8: ***LLM Latency by Model***

(Grouped bars comparing median planning latencies for Gemini 2.5 Flash, Gemini 2.0, and ChatGPT across Simple, Moderate, and Complex tasks; the lighter-weight model exhibits the lowest latency across categories, with all models remaining in the sub- to low-single-second range for plan generation.)

Conclusion

CATALYST provides an accessible, voice first automation stack that can perform long horizon tasks efficiently through closed loop error handling and automatic re planning. By facilitating the removal of programming barriers, it aims at the domains of assistive care, logistics, healthcare, cultural preservation, and home environments while maintaining a low cognitive level; the end to end delay is primarily due to physical execution. There are still some key limitations: ASR induced latency, sim to real transfer, English centric prompts, network dependence, and reduced reliability for very complex, multi step instructions. The near term priorities are: (i) on device/edge ASR and planning to reduce latency and outage areas; (ii) integration with physical robots with advanced sensing, actuation, and safety; (iii) multilingual prompts and multimodal inputs (speech, vision, gesture, haptics); (iv) memory and reasoning augmented re planning for long horizons; and (v) cooperative multi robot task sharing with role coordination. These actions move CATALYST closer to the deployment of time sensitive, safety critical, and resource constrained scenarios while still fulfilling its primary objective democratizing advanced robotics for non experts.

References

[1] IEEE BRACU Student Branch, Voice-controlled robots: Bridging the gap between fundamentals and industrial integration, (2021). Available: https://ieeebracu.com/voice-controlled-robots-bridgingthe-gap-between-fundamentals-and-industrial-integration/ (14 Apr 2025).

[2] Purdue University, Human robot interaction with cloud assisted voice control and vision systems, (2021). Available: https://hammer.purdue.edu/mechaerospace_theses/863/ (10 Apr 2025).

[3] Techpacs, Voice-activated robotics: Revolutionizing control with speech recognition-based robotic car movement system, (2022). Available: https://techpacs.ca/voice-activated-robotics-revolutionizingcontrol-with-speech-recognition-based-robotic-car-movement-system-1828 (06 Apr 2025).

[4] The role of natural language processing in robotic systems: Enhancing human-robot interaction, (2020). Available: https://www.atlantispress.com/article/126003454.pdf (02 Apr 2025).

[5] LA-RCS: LLM-agent based robot control system for autonomous planning and environmental analysis, arXiv preprint arXiv:2505.18214, (2025). Available: https://arxiv.org/abs/2505.18214 (30 Sep 2025).

[6] Milvus, How multimodal AI is used in robotics: Vision, sound, touch, and language integration, (2023). Available: https://milvus.io/ai-quick-reference/how-is-multimodal-ai-used-in-robotics (26 Sep 2025).

[7] Gemini robotics: Bringing AI into the physical world with zero-shot robot control, arXiv preprint arXiv:2503.20020, (2025). Available: https://arxiv.org/abs/2503.20020 (22 Sep 2025).

[8] Google Cloud, Gemini 2.0 Flash | Generative AI on Vertex AI with enhanced agentic capabilities, (2024). Available: https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash (18 Sep 2025).

[9] Hostinger, What is n8n? Introduction to workflow automation tool architecture and implementation, (2023). Available: https://www.hostinger.in/tutorials/what-is-n8n (14 Sep 2025).

[10] Futurepedia, AssemblyAI reviews: Use cases, pricing & alternatives for speech-to-text API integration, (2024). Available: https://www.futurepedia.io/tool/assemblyai (10 Sep 2025).

[11] n8n.io, AI workflow automation platform & tools — n8n architecture and implementation framework, (2024). Available: https://n8n.io (06 Sep 2025)

[12] AssemblyAI, AssemblyAI | AI models to transcribe and understand speech with industry-leading accuracy, (2024). Available: https://www.assemblyai.com (02 Sep 2025).

[13] Duino4Projects, Top 5 best online Arduino simulators for virtual prototyping and development, (2023). Available: https://duino4projects.com/en/online-arduino-simulator/ (29 Aug 2025).

[14] Programming Electronics Academy, The Arduino simulator you've been looking for: TinkerCAD review and tutorial, (2024). Available: https://www.programmingelectronics.com/arduino-simulator-tinkercad/ (25 Aug 2025).

[15] Wikipedia, Lights-out (manufacturing), (2023). Available: https://en.wikipedia.org/wiki/Lights_out_(manufacturing) (21 Aug 2025).

[16] Hyundai Motor Group TV, Ballet parking: When robots dance to park your car | #Shorts, (2023). Available: https://www.hyundaimotorgroup.com/tv/CONT0000000000161850 (17 Aug 2025).

[17] EAI Endorsed Transactions on Creative Technologies, Automatic speech recognition for human-robot interaction on humanoid robot Barelang 7, 10(37) (2023) e3. Available: https://eudl.eu/doi/10.4108/eai.7-11-2023.2342940 (25 May 2025).

[18] Jorie, What is workflow automation in healthcare today? Applications and benefits, (2023). Available: https://www.jorie.ai/post/what-is-workflow-automation-in-healthcare-today (13 Aug 2025).

[19] Voice control interface for surgical robot assistants with Whisper integration and ROS framework, arXiv preprint arXiv:2409.10225, (2024). Available: https://arxiv.org/abs/2409.10225 (09 Aug 2025).

[20] Xerox, Workflow automation solutions for healthcare: Providers and payers implementation, (2024). Available: https://www.xerox.com/en-us/services/workflow-automation/healthcare (05 Aug 2025).

[21] Boomi, 8 ways healthcare workflow automation improves patient care and system efficiency, (2024). Available: https://boomi.com/blog/healthcare-workflow-automation/ (01 Aug 2025).

[22] Botasys, Human-robot interaction: Key concepts & what lies ahead in collaborative robotics, (2024). Available: https://www.botasys.com/post/human-robot-interaction (28 Jul 2025).

[23] Zilliz, How multimodal AI is used in robotics: Text, images, audio, and video processing, (2024). Available: https://zilliz.com/ai-faq/how-is-multimodal-ai-used-in-robotics (24 Jul 2025).

[24] University of Texas at Arlington, Human robot interaction with cloud assisted voice control and vision implementation, (2021). Available: https://mavmatrix.uta.edu/mechaerospace_theses/863/ (20 Jul 2025).

[25] Natural language processing in robotics: Leveraging Python for human-robot interaction, Int J Sci Res, 13(11) (2024). Available: https://www.ijsr.net/getabstract.php?paperid=SR241021044213 (25 May 2025).

[26] Improving robotic arms through natural language processing, computer vision, and edge computing, arXiv preprint arXiv:2405.17665, (2024). Available: https://arxiv.org/abs/2405.17665 (16 Jul 2025).

[27] Meta AI, PARTNR: A benchmark for planning and reasoning tasks in human-robot collaboration, (2024). Available: https://ai.meta.com/research/publications/partnr-a-benchmark-for-planning-and-reasoning-in-embodied-multi-agent-tasks/ (12 Jul 2025).

[28] Lamarr Institute, Embodied AI explained: Principles, applications, and future perspectives in physical intelligence, (2023). Available: https://lamarrinstitute.org/blog/embodied-ai-explained/ (08 Jul 2025).

[29] YouTube, Virtual Arduino Playground: Simulate, experiment, create with Tinkercad Arduino simulation tutorial, (2023). Available: https://www.youtube.com/watch?v=HinjXIqXFww (04 Jul 2025).

[30] 10xDS, Multimodal AI in robotics: Enabling smarter machines through unified perception and reasoning, (2024). Available: https://10xds.com/blog/how-multimodal-ai-is-reshaping-industries-through-deeper-intelligence/ (20 Jun 2025).

[31] Arm Newsroom, Transforming the future of AI and robotics with multimodal LLMs and autonomous systems, (2024). Available: https://newsroom.arm.com/blog/llms-and-autonomous-robots (26 Jun 2025).

[32] IBM, What is multimodal AI? Integration of text, images, audio, and video processing, (2023). Available: https://www.ibm.com/think/topics/multimodal-ai (22 Jun 2025).

[33] Blukis V, Paxton C, Fox D, Garg A &Artzi Y, A persistent spatial semantic representation for high-level natural language instruction execution, Conf Robot Learn (PMLR), (2022). Available: https://arxiv.org/abs/2107.05612 (25 May 2025).

[34] Min K, Chaplot D S, Davidson J, Savva M & Batra D, FILM: Following instructions in language with modular methods, arXiv preprint arXiv:2110.07342, (2021). Available: https://arxiv.org/abs/2110.07342 (18 Jun 2025).

[35] Inoue K & Ohashi T, Prompter: Utilizing large language model prompting for data efficient embodied instruction following, arXiv preprint arXiv:2211.03267, (2022). Available: https://arxiv.org/abs/2211.03267 (14 Jun 2025).

[36] Context-aware planning and environment-aware memory for instruction following embodied agents, arXiv preprint arXiv:2308.07241, (2023). Available: https://arxiv.org/abs/2308.07241 (10 Jun 2025).

[37] Chen Y, Cui W, Zhang Z, Liu C & Yang Z, RoboGPT: An LLM-based embodied long-term decision making agent for instruction following tasks, arXiv preprint arXiv:2311.15649, (2023). Available: https://arxiv.org/abs/2311.15649 (06 Jun 2025).

[38] AssemblyAI Blog, The top free speech-to-text APIs, AI models, and open source engines comparison, (2023). Available: https://www.assemblyai.com/blog/the-top-free-speech-to-text-apis-and-open-source-engines/ (02 Jun 2025).

[39] Google DeepMind, Gemini robotics brings AI into the physical world: Interactive, generalizable, and dexterous, (2023). Available: https://deepmind.google/discover/blog/gemini-robotics-brings-ai-into-the-physical-world/ (29 May 2025).

[40] AWS Marketplace, Start building voice intelligence with AssemblyAI's speech-to-text model from AWS Marketplace, (2024). Available: https://aws.amazon.com/blogs/awsmarketplace/start-building-voice-intelligence-with-assemblyais-speech-to-text-model-from-aws-marketplace/ (25 May 2025).

[41] Int J Recent Adv Technol, Enhancing human-robot interaction through voice-driven natural language processing, 12(4) (2024). Available: https://www.ijraset.com/research-paper/enhancing-human-robot-interaction (25 May 2025).

[42] GitHub, n8n-io/n8n: Fair-code workflow automation platform with native AI capabilities and 400+ integrations, (2024). Available: https://github.com/n8n-io/n8n (21 May 2025).

[43] Planning as in-painting: A diffusion-based embodied task planning framework for environments under uncertainty, arXiv preprint arXiv:2312.01097, (2023). Available: https://arxiv.org/abs/2312.01097 (17 May 2025).

[44] LLM-based symbolic planner for complex task reasoning, arXiv preprint arXiv:2410.24164, (2024). Available: https://arxiv.org/abs/2410.24164 (13 May 2025).

[45] LLM-based robot task planning with exceptional handling for general purpose service robots, arXiv preprint arXiv:2405.15646, (2024). Available: https://arxiv.org/abs/2405.15646 (05 May 2025).

[46] YouTube, How to use AssemblyAI with Python, (2024). Available: https://www.youtube.com/watch?v=5LJFK7eOC20 (05 May 2025).

[47] All3DP, Best Arduino simulators of 2024 (online & offline), (2024). Available: https://all3dp.com/2/best-arduino-simulators-online-offline/ (01 May 2025).

[48] Embodied task planning with large language models for grounded planning with physical scene constraints, arXiv preprint arXiv:2307.0184, (2023). Available: https://arxiv.org/abs/2307.0184 (23 Apr 2025).

[49] Google Cloud, Google Models | Generative AI on Vertex AI: Comprehensive model documentation and specifications, (2024). Available: https://cloud.google.com/vertex-ai/generative-ai/docs/models (23 Apr 2025).

[50] NMM-HRI: Natural multimodal human-robot interaction with voice and deictic posture via large language model, arXiv preprint arXiv:2501.00785, (2025). Available: https://arxiv.org/abs/2501.00785 (15 Apr 2025).

[51] LLM-empowered embodied agent for memory-augmented task planning in household robotics, arXiv preprint arXiv:2504.21716, (2025). Available: https://arxiv.org/abs/2504.21716 (15 Apr 2025).

[52] Large language models as generalizable policies for embodied tasks, In: Int Conf on Learning Representations, (2023). Available: https://arxiv.org/abs/2310.17722 (15 Apr 2025).