RADI: LLMs AS WORLD MODELS FOR ROBOTIC ACTION DECOMPOSITION AND IMAGINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robotics is irreplaceable in driving social progress, enhancing productivity and improving human life, and efficient task planning is the key to ensuring that robots accurately perform complex tasks. Traditional world models based on physical simulation or rule-based engines are limited by the high cost of environment modeling and dynamic scene generalization capabilities. Although large language models (LLMs), represented by GPT, have shown potential for generalized intelligence in natural language processing tasks and have made initial progress in robotic task planning, their generalization ability as a world model for the robotics domain has not been systematically verified. No study has yet answered the question of whether LLMs can predict physical action outcomes through task decomposition and environment imagery (rather than pure linguistic reasoning), and how to assess their world modeling capabilities. In this paper, we propose the **R**obotic **A**ction **D**ecomposition and **I**magination (RADI) framework, which combines the self-reflective capability of LLMs to improve the success rate of task planning through the two core mechanisms of action decomposition and environment imagination. Specifically, RADI first gradually decomposes a complex robot task into atomic action sequences, then imagines the execution results of each action based on the environment state, and verifies whether it meets the task expectations through the state changes. If the expectations are not met, it triggers the self-reflective mechanism to re-optimize the action decomposition. The experiments are conducted based on GPT-4 in the VirtualHome environment, and the results show that RADI significantly improves the success rate of task planning, and verifies the effectiveness of LLM as a world model in robotics.

1 INTRODUCTION

Robotics is of irreplaceable importance in promoting social progress, enhancing productivity, and improving human life (Brooks, 1986). For example, the application of robots in the manufacturing industry has greatly improved productivity and product quality (Gatla et al., 2007a;b). Robot task planning is a crucial step to ensure that robots can complete complex tasks efficiently and accurately, and it enables robots to maximize their performance in various application scenarios by analyzing, decomposing, and optimizing paths for tasks (Hanheide et al., 2017; Paxton et al., 2019; Galindo et al., 2008; Zhang et al., 2017). In the field of robot task planning, world models (Ha & Schmidhuber, 2018) play a pivotal role. It is crucial to predict the outcome of actions, reduce the cost of physical trial and error, and improve the safety of decisions. Traditional world models based on physical simulation or rule engines have significant limitations since they rely heavily on accurate environment modeling (Blumenthal et al., 2013; Roth et al., 2003; Zhang & Faugeras, 1990), a process that is costly and difficult to generalize to complex or dynamic scenarios.

Large Language Models (LLMs) (Zhao et al., 2023), represented by the GPT family of models (Floridi & Chiriatti, 2020; Lund & Wang, 2023; Achiam et al., 2023) developed by OpenAI, have achieved performance far beyond that of previous models on a wide range of tasks in natural language processing (Baktash & Dawodi, 2023), and to some extent have shown the potential for artificial general intelligence (AGI) by going beyond the language model itself and understanding the physical world (Bubeck et al., 2023), and even more recently there has been some research to show that LLMs can be effective in generating robot task plans. For example, the PROGPROMPT(Singh et al., 2023) achieves a high success rate in the VirtualHome housework task using LLMs with

program-like prompts. In addition, the RoboMatrix (Mao et al., 2024) framework provides a skill-centered hierarchical approach for scalable robot task planning and execution in the open world, demonstrating generalization performance across new objects, scenarios, tasks, and robots.

However, previous approaches have limited large language models for task planning and have not verified whether LLMs have the ability to generalize as world models in the robotics domain (Yao et al., 2023; Wu et al., 2023; Chalvatzaki et al., 2023; Wu et al., 2024). There are currently two major research gaps in LLM as a world model for robotics task planning. The first is whether LLM can predict the outcome of physical actions through environment imagination rather than purely linguistic reasoning. The other is how to evaluate the world modeling ability of LLM.

In this paper, we propose the **R**obotic **A**ction **D**ecomposition and **I**magination (RADI) framework. Specifically, we first utilize LLM to achieve action decomposition by planning and progressively decomposing a complex task into a series of atomic actions. Subsequently, we let the large model perform environment imagination. Based on the current state of the environment, the LLM is asked to predict the change in the state of the environment after the execution of each atomic action, and to anticipate whether such a change in the state can meet the expectations of the task. In the event that it does not meet the expectation, we can ask the LLM to re-perform the action decomposition, thus improving the success rate of task planning through the reflection of the LLM itself. The experiment conducted in VirturalHome (Puig et al., 2018) shows an improvement in task planning success rate that can be used as a measure of the LLM's ability to act as a world model. The contribution of this paper can be summarized as follows:

- We propose the RADI framework that consists of action decomposition and environmental imagination, allowing the LLMs to break down complex robot tasks and predict the outcomes of actions based on the current environmental state, as well as achieve environmental imagination-driven error correction for action decomposition.
- We provide a systematic way to explore the potential of LLMs as world models in the field of robot task planning, paving the way for more interpretable and reliable applications of LLMs in robotic systems in a variety of environments.
- We conduct experiments on four public datasets in VirtualHome using GPT-4, one of the state-of-the-art LLMs. Experimental results show the effectiveness of the RADI framework to improve robot task planning, and LLMs can serve as the world model for robotics.

2 PRELIMINARIES

Robot task planning. Robot task planning is the process of allowing a robot, upon receiving a command for a particular task, to generate a detailed executable plan in a given environment to achieve the goal of the task (Tsarouchi et al., 2016; Hanheide et al., 2017; Paxton et al., 2019). Specifically, given the task goal G, the observation O consisting of objects in the environment E and their relationships, and a set of all possible actions $A = \{a_1, a_2, \ldots, a_n\}$ executable for the robot, a task planning algorithm $\mathcal{T}$ aims to find an action sequence π to achieve the goal of the task. In other word, $\mathcal{T} : (G, O, A) \mapsto \pi$. For example, if the goal $G =$*"put one cupcake in microwave and switch on microwave"*, the observation $O =$*"one cupcake is in fridge, one cupcake is in kitchencabinet"*, the possible action set $A = \{walk, open, ..., switchon\}$, then we aim to generate an action sequence $\pi =$ *"walk to fridge, open fridge, grab cupcake,...,switchon microwave"*.

3 METHODOLOGY

We propose the RADI framework, where we first let the LLM complete the decomposition of a robot action sequence based on the task description and the observed environment state. Then, we let the LLM be a world model, imagining the change of the environment state after the execution of the action sequence to verify whether the action sequence can accomplish the corresponding task. If LLM determines that the task cannot be accomplished, we let the LLM repeat the action decomposition. The overall pipeline of the proposed framework is shown in Figure 1.

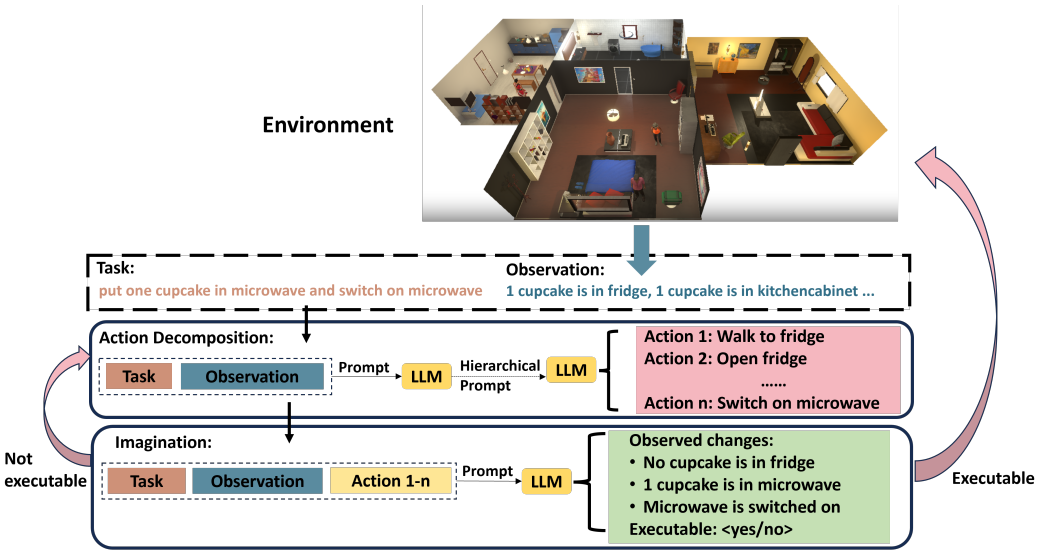

Figure 1: The overview of robot task planning pipeline of the proposed RADI framework.

3.1 Action Decomposition

The action decomposition module is independent of the followed imagination module and can be implemented by any decomposition methods. In practice, to improve the efficiency of action decomposition, we adopt a hierarchical task decomposition framework (Wu et al., 2024) that reduces the planning complexity of difficult tasks through an incremental decomposition strategy. Firstly, a goal-oriented decomposition is adopted to decompose the overall task into a number of independent sub-goals based on semantic associations; subsequently, environment observation is introduced at the task execution layer, and each sub-goal is transformed into an actionable sequential task chain through hierarchical prompt templates; and finally, at the action generation layer, each sub-goal is parsed into specific action instructions by combining with the domain knowledge base, and the standardized action sequences are extracted by a pattern-matching algorithm. Formally, given task goal G, observation O and the possible action set A, we use LLM to obtain the action sequence:

$$\pi = (a_1, a_2, ..., a_m) = LLM(G, O, A). \tag{1}$$

3.2 Imagination via LLM Simulation

The imagination module aims to let LLM predict the execution result of an action sequence based on the current environment state, focusing on the construction of a closed-loop inference mechanism. Specifically, the complete action sequence is generated by the action decomposition module. In the imagination step, the structured text description of the current environment state (including object position relationship, status, etc.) is input into the LLM together with the action sequence. The LLM is prompted to output two prediction results: (1) the natural language description of the action effect (e.g., "after the robotic arm grasps the cup, the cup will be detached from the table and move with the robotic arm"), and (2) whether the whole action sequence is feasible for task completion. We use the entire sequence of actions rather than individual atomic actions, which allows the LLM world model to imagine the state changes of the whole process rather than unilaterally considering the effects of individual actions, as Figure 2 shows. This design also led to a significant reduction in the time of inferences in the face of long action sequences.

We also design a multi-round iterative correction mechanism, i.e., if the LLM world model determines that an action sequence is infeasible, the system will automatically trigger a Self-Correction Prompt, which requires the LLM to re-generate a new action sequence by combining the conflict information. If the LLM still fails to pass the imagination after K times of action decomposition, we regard it as an abstention instead of forcing the execution of this action sequence.

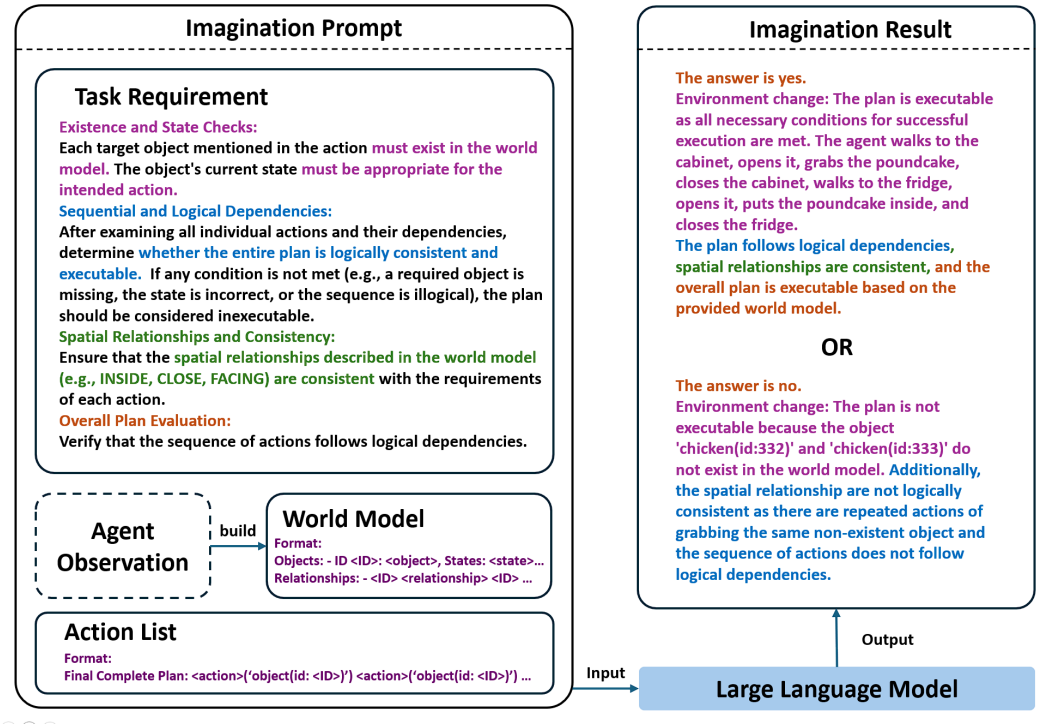

Figure 2: An example with prompt and result of imagination in our RADI framework.

4 Experiments

4.1 Experimental Settings

Environment. All experiments are performed in VirtualHome (Puig et al., 2018), a three-dimensional household simulation platform that includes various rooms (e.g., kitchens, living rooms, and bedrooms) and a wide array of interactive objects. VirtualHome offers a realistic yet controlled setting, enabling systematic evaluation of GPT-4's capacity for predicting physical outcomes. Through successive iterations of action decomposition and outcome verification, our method assesses whether GPT-4 can consistently produce valid, goal-aligned action sequences in an environment that closely mirrors everyday scenarios.

Datasets We exclusively utilize the In-Distribution dataset from LID Li et al. (2022), which contains tasks conforming to the original training distribution. By restricting our experiments to these tasks, we establish a controlled testbed to evaluate how effectively GPT-4, under purely prompt-based guidance, can decompose actions, anticipate their consequences, and validate state changes in a household environment. Crucially, we do not employ additional in-context examples for GPT-4; all performance results stem directly from the model's inherent reasoning abilities and its capacity to internalize and respond to carefully crafted prompts.

Large Language Models We adopt GPT-4 as the primary large language model for this study. Unlike methods that rely on extensive fine-tuning, our framework remains purely prompt-based, leveraging GPT-4's capacity for iterative task decomposition and environment "imagination." Specifically, GPT-4 breaks down complex robotic tasks into atomic actions and predicts the resulting state transitions without any parameter updates. This design rigorously tests GPT-4's ability to infer physical dynamics and assess the feasibility of each action in a simulated environment.

Evaluation Metrics We employ three complementary metrics as our exclusive means of evaluation: Success Rate After Abstention (SRA), Abstention Rate (AR), and Overall Success Rate (OSR). SRA is defined as the percentage of tasks successfully executed among those the system does not

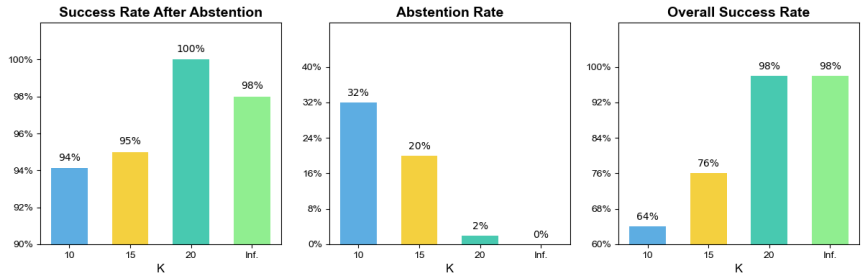

Figure 3: The performance of our RADI framework with varying repeat generation limit K.

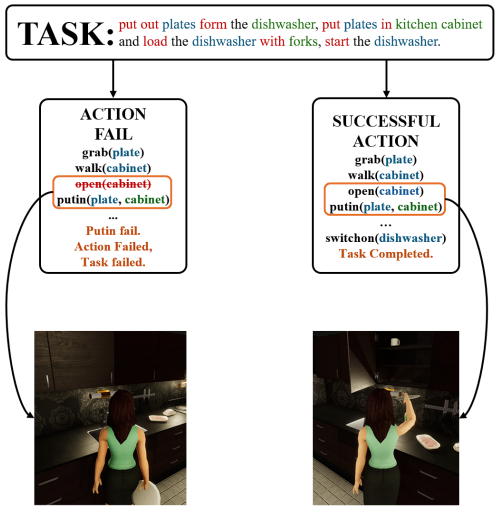

Figure 4: An example of failure without imagination, but success after imagination and correction.

abstain from, AR measures the proportion of tasks the system opts to skip, and OSR indicates the fraction of all tasks that are ultimately completed. In our framework, a plan is deemed successful only if it satisfies two critical criteria: (1) all actions can be executed in a logically consistent manner, and (2) the resultant state transitions precisely align with the intended outcomes.

4.2 Experimental Results

The results of our quantitative experiments are shown in Figure 3. An important finding is that the tasks that pass the verification in the imagination module by the LLM world model have very high success rates, and, unsurprisingly, as the repeat generation limit K increases, the percentage of abstentions gradually decreases to 0, and the overall success rate shows an increasing trend. A qualitative example is shown in Figure 4. Without imagination and correction, the generated action sequence is missing the step of opening the cupboard and cannot complete the task, whereas after imagination and correction, the correct and complete action sequence is generated.

5 Conclusion

In this paper, we explore the key challenge of utilizing LLM as a world model for robot task planning. We propose the RADI framework, which integrates action decomposition and imagination to improve the success of robotic task planning. By progressively decomposing complex tasks into atomic actions and modeling their outcomes through environment state change prediction, RADI enables LLMs to self-reflect and iteratively improve action sequences. Experiments in the VirtualHome show the effectiveness of our framework.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jawid Ahmad Baktash and Mursal Dawodi. Gpt-4: A review on advancements and opportunities in natural language processing. *arXiv preprint arXiv:2305.03195*, 2023.

Sebastian Blumenthal, Herman Bruyninckx, Walter Nowak, and Erwin Prassler. A scene graph based shared 3d world model for robotic applications. In *2013 IEEE International Conference on Robotics and Automation*, pp. 453–460, 2013. doi: 10.1109/ICRA.2013.6630614.

Rodney Brooks. A robust layered control system for a mobile robot. *IEEE journal on robotics and automation*, 2(1):14–23, 1986.

Sébastien Bubeck, Varun Chadrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Georgia Chalvatzaki, Ali Younes, Daljeet Nandha, An Thai Le, Leonardo FR Ribeiro, and Iryna Gurevych. Learning to reason over scene graphs: a case study of finetuning gpt-2 into a robot language model for grounded task planning. *Frontiers in Robotics and AI*, 10:1221739, 2023.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

Cipriano Galindo, Juan-Antonio Fernández-Madrigal, Javier González, and Alessandro Saffiotti. Robot task planning using semantic maps. *Robotics and Autonomous Systems*, 56(11):955–966, 2008. ISSN 0921-8890. doi: https://doi.org/10.1016/j.robot.2008.08.007. Semantic Knowledge in Robotics.

Chandra Sekhar Gatla, Ron Lumia, John Wood, and Greg Starr. An automated method to calibrate industrial robots using a virtual closed kinematic chain. *IEEE Transactions on Robotics*, 23(6): 1105–1116, 2007a. doi: 10.1109/TRO.2007.909765.

Chandra Sekhar Gatla, Ron Lumia, John Wood, and Greg Starr. Calibration of industrial robots by magnifying errors on a distant plane. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3834–3841, 2007b. doi: 10.1109/IROS.2007.4398969.

David R Ha and Jürgen Schmidhuber. World models. *ArXiv*, abs/1803.10122, 2018. URL `https://api.semanticscholar.org/CorpusID:4807711`.

Marc Hanheide, Moritz Göbelbecker, Graham S. Horn, Andrzej Pronobis, Kristoffer Sjöö, Alper Aydemir, Patric Jensfelt, Charles Gretton, Richard Dearden, Miroslav Janicek, Hendrik Zender, Geert-Jan Kruijff, Nick Hawes, and Jeremy L. Wyatt. Robot task planning and explanation in open and uncertain worlds. *Artificial Intelligence*, 247:119–150, 2017. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2015.08.008. Special Issue on AI and Robotics.

Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, Jacob Andreas, Igor Mordatch, Antonio Torralba, and Yuke Zhu. Pre-trained language models for interactive decision-making. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 31199–31212. Curran Associates, Inc., 2022.

Brady D Lund and Ting Wang. Chatting about chatgpt: how may ai and gpt impact academia and libraries? *Library hi tech news*, 40(3):26–29, 2023.

Weixin Mao, Weiheng Zhong, Zhou Jiang, Dong Fang, Zhongyue Zhang, Zihan Lan, Fan Jia, Tiancai Wang, Haoqiang Fan, and Osamu Yoshie. Robomatrix: A skill-centric hierarchical framework for scalable robot task planning and execution in open-world. *arXiv preprint arXiv:2412.00171*, 2024.

Chris Paxton, Yotam Barnoy, Kapil Katyal, Raman Arora, and Gregory D. Hager. Visual robot task planning. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8832–8838, 2019. doi: 10.1109/ICRA.2019.8793736.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8494–8502, 2018.

M. Roth, D. Vail, and M. Veloso. A real-time world model for multi-robot teams with high-latency communication. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453)*, volume 3, pp. 2494–2499 vol.3, 2003. doi: 10.1109/IROS.2003.1249244.

Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11523–11530. IEEE, 2023.

Panagiota Tsarouchi, Sotiris Makris, and George Chryssolouris. Human–robot interaction review and challenges on task planning and programming. *International Journal of Computer Integrated Manufacturing*, 29(8):916–931, 2016.

Yike Wu, Jiatao Zhang, Nan Hu, Lanling Tang, Guilin Qi, Jun Shao, Jie Ren, and Wei Song. Mldt: Multi-level decomposition for complex long-horizon robotic task planning with open-source large language model. In *International Conference on Database Systems for Advanced Applications*, pp. 251–267. Springer, 2024.

Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan. Embodied task planning with large language models. *arXiv preprint arXiv:2307.01848*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Yu Zhang, Sarath Sreedharan, Anagha Kulkarni, Tathagata Chakraborti, Hankz Hankui Zhuo, and Subbarao Kambhampati. Plan explicability and predictability for robot task planning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1313–1320, 2017. doi: 10.1109/ICRA.2017.7989155.

Z. Zhang and O. Faugeras. Building a 3d world model with a mobile robot: 3d line segment representation and integration. In *[1990] Proceedings. 10th International Conference on Pattern Recognition*, volume i, pp. 38–42 vol.1, 1990. doi: 10.1109/ICPR.1990.118061.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.