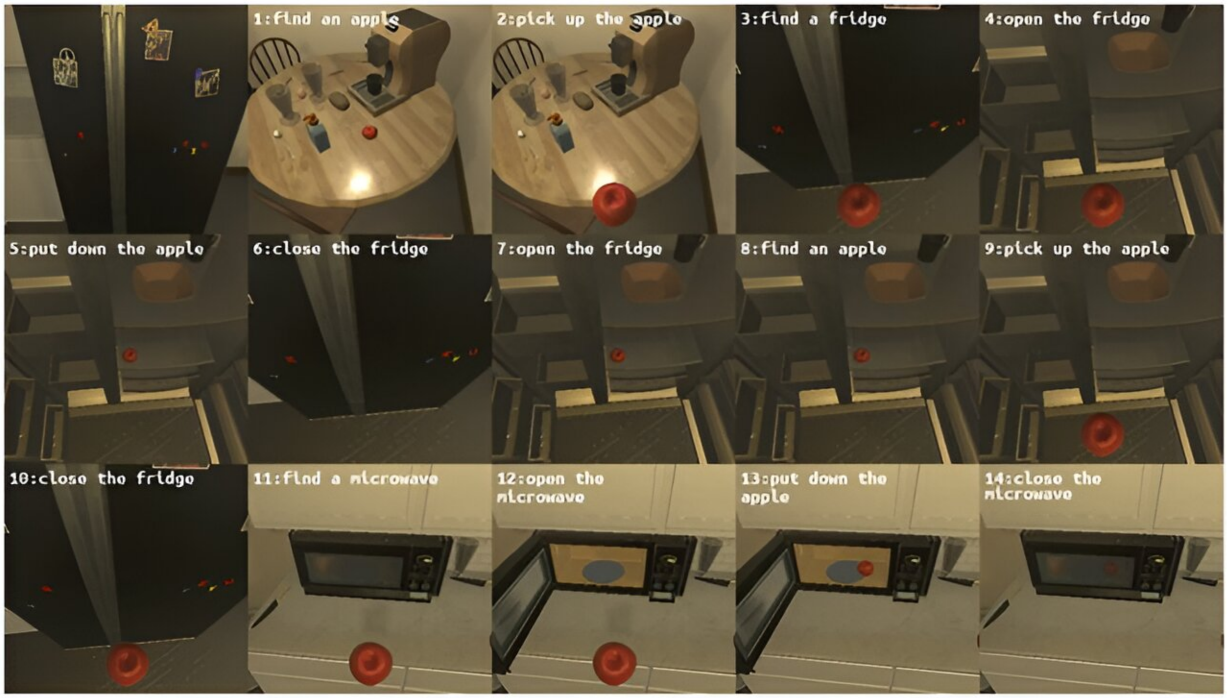
Researchers develop an automated benchmark for language-based task planners

April 26 2024



Case Study of Procedural Generation Following the Command: "Put a chilled apple in the microwave." Credit: Electronics and Telecommunications Research Institute(ETRI)

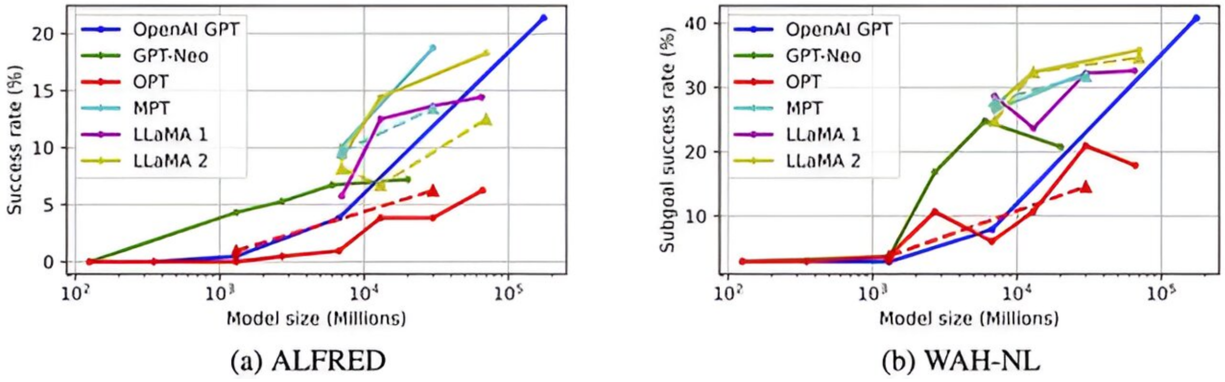
If instructed to "Place a cooled apple into the microwave," how would a robot respond? Initially, the robot would need to locate an apple, pick it up, find the refrigerator, open its door, and place the apple inside. Subsequently, it would close the refrigerator door, reopen it to retrieve the cooled apple, pick up the apple again, and close the door. Following this, the robot would need to locate the microwave, open its door, place the apple inside, and then close the microwave door.

Evaluating how well these steps are executed exemplifies the essence of benchmarking task planning AI technologies. It measures how effectively a robot can respond to commands and adhere to the specified procedures.

An Electronics and Telecommunications Research Institute (ETRI) research team has developed a technology that automatically evaluates the performance of task plans generated by Large Language Models (LLMs) and paves the way for fast and objective assessment of task planning AIs.

ETRI has announced the development of LoTa-Benchmark (LoTa-Bench), which enables the automatic evaluation of language-based task planners. A language-based task planner understands the verbal instruction from a human user, plans a sequence of operations, and autonomously executes the designated operations to fulfill the goal of the instruction.

The research team published a paper at the International Conference on Learning Representations (ICLR), and shared the evaluation results for a total of 33 large language models through GitHub.



(a) ALFRED

(b) WAH-NL

Results of Procedural Generation Performance Evaluation Across Various Large Language Models. Credit: Electronics and Telecommunications Research Institute(ETRI)

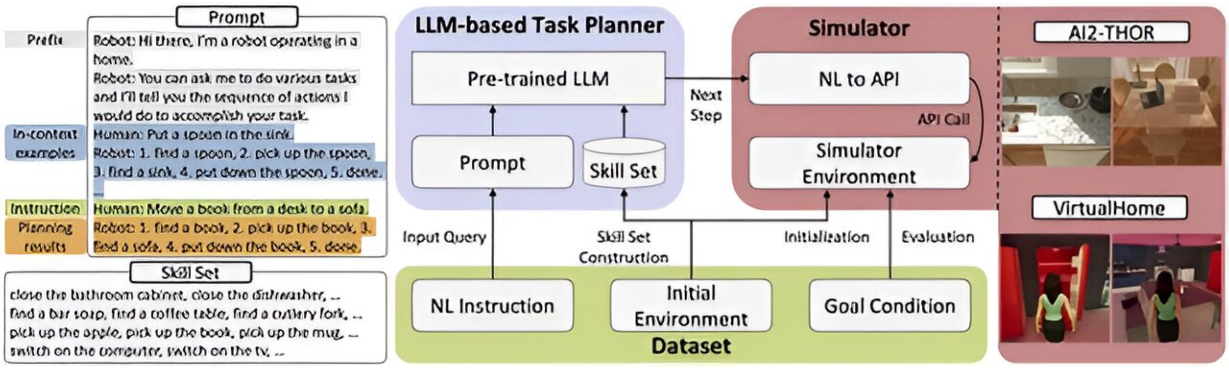
Recently, LLMs have demonstrated remarkable performance not only in language processing, conversation, solving mathematical problems, and logic proof but also in understanding human commands, autonomously selecting sub-tasks, and sequentially executing them to achieve goals. Consequently, there has been a widespread effort to apply large language models in robotics applications and service implementation.

Previously, the absence of benchmark technology capable of automatically evaluating task planning performance necessitated manual assessments, which were labor-intensive. For instance, in existing research, including Google's SayCan, the method adopted involved multiple individuals directly observing the results of tasks being executed and then voting on their success or failure.

This approach not only required a significant amount of time and effort for performance evaluation, making it cumbersome but also introduced the problem of subjective judgment influencing the results.

The LoTa-Bench technology developed by ETRI automates the evaluation process by actually executing task plans generated by large language models based on user commands and automatically compares the outcomes to the intended results of the commands to determine whether the plans were successful or not. This approach significantly reduces evaluation time and costs as well as ensures that the evaluation results are objective.

ETRI revealed benchmark results for different large language models, indicating that OpenAI's GPT-3 achieved a success rate of 21.36%, GPT-4 exhibited 40.38%, Meta's LLaMA 2-70B model showed 18.27%, and MosaicML's MPT-30B model recorded 18.75%.



Robot: Hi there, I'm a robot operating in a home.
Robot: You can ask me to do various tasks and I'll tell you the sequence of actions I would do to accomplish your task.
Human: Put a spoon in the sink.
Robot: 1. find a spoon, 2. pick up the spoon, 3. find a sink, 4. put down the spoon, 5. done.
...
Human: Move a book from a desk to a sofa.
Robot: 1. find a book, 2. pick up the book, 3. find a sofa, 4. put down the book, 5. done.

close the bathroom cabinet, close the dishwasher, ...
find a bar soap, find a coffee table, find a cutlery fork, ...
pick up the apple, pick up the book, pick up the mug, ...
switch on the computer, switch on the tv, ...

Structure of LoTa-Benchmark (LoTa-Bench). Credit: Electronics and Telecommunications Research Institute(ETRI)

It was noted that larger models tend to have superior task planning capabilities. A success rate of 20% implies that out of 100 instructions, 20 plans were successful in fulfilling the goal of the instructions.

In LoTa-Bench, performance evaluation is conducted in virtual simulation environments developed by the Allen Institute for AI (AI2-THOR) and the Massachusetts Institute of Technology (MIT's VirtualHome) aimed at research and development of robotics and embodied agent intelligence. The evaluation utilized the ALFRED dataset that included everyday household task instructions such as "Place a cooled apple in the microwave" etc.

Leveraging the benefits of the LoTa-Bench technology for easy and rapid verification of new task planning methods, the research team discovered two strategies for improving task planning performance through data-driven training: In-Context Example Selection and Feedback-Based Replanning. They also confirmed that fine-tuning effectively enhances the performance of language-based task planning.

Minsu Jang, a principal researcher at ETRI's Social Robotics Lab, stated, "LoTa-Bench marks the first step in the development of task planning AI. We plan to research and develop technologies that can predict task failures in uncertain situations or improve task generation intelligence by asking for and receiving help from humans. This technology is essential for realizing the era of one robot per household."

Jaehong Kim, the director of ETRI's Social Robotics Research Section, announced, "ETRI is dedicated to advancing robotic intelligence using foundation models to realize robots capable of generating and executing various mission plans in the real world."

By releasing the software as open source, the ETRI researchers anticipate that companies and educational institutions will be able to freely utilize this technology, thereby accelerating the advancement of related technologies.

More information: Choi et al, LoTa-Bench: Benchmarking Language-oriented Task Planners for Embodied Agents, ICLR (International Conference on Learning Representations (2024)

Provided by National Research Council of Science and Technology

Citation: Researchers develop an automated benchmark for language-based task planners (2024, April 26) retrieved 8 May 2024 from https://techxplore.com/news/2024-04-automated-benchmark-language-based-task.html