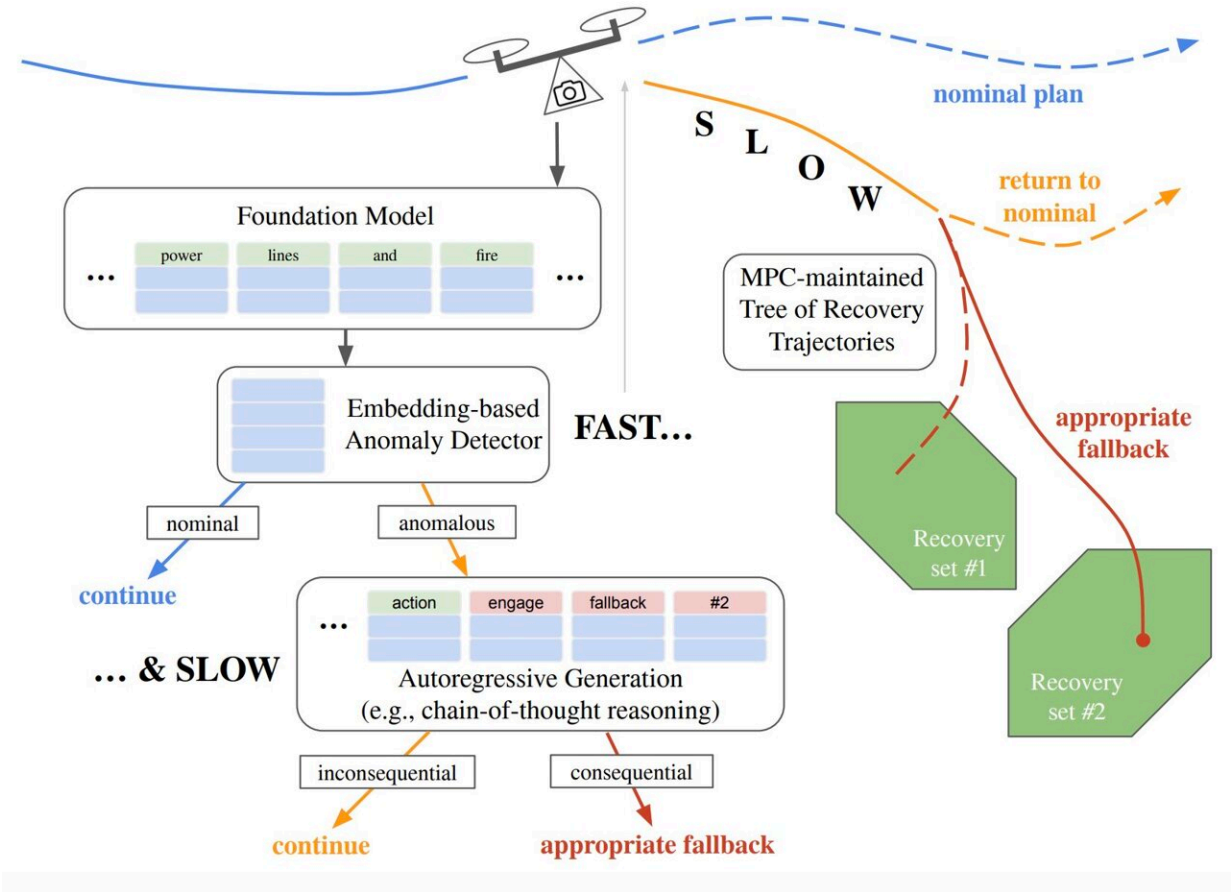
A two-stage framework to improve LLM-based anomaly detection and reactive planning

August 15 2024, by Ingrid Fadelli



An embedding-based runtime monitoring scheme using fast and slow language model reasoners in concert. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2407.08735

Large language models (LLMs), such as OpenAI's ChatGPT, are known to be highly effective in answering a wide range of user queries, generalizing well across many natural language processing (NLP) tasks. Recently, some studies have also been exploring the potential of these models for detecting and mitigating robotic system failures.

Researchers at Stanford University and NVIDIA recently introduced a new two-stage framework that could facilitate the use of LLMs for detecting system anomalies and planning robotic actions in real-time.

This approach, introduced in a paper that won the Outstanding Paper Award at the Robotics: Science and Systems conference (RSS 2024), could significantly enhance the trustworthiness of various robotic systems, including self-driving vehicles. The work is available on the *arXiv* preprint server.

"This line of work started when we came across examples of real-world failure modes of self-driving vehicles, such as the case of a self-driving car that gets confused by a set of traffic lights being transported by a truck or a case where a self-driving car stopped on the freeway because it drove past a billboard with a picture of a stop sign on it," Rohan Sinha, co-author of the paper, told Tech Xplore.

"Such examples are often called out-of-distribution (OOD) inputs, rare corner cases that differ significantly from the data on which AVs are trained."

As part of their previous studies, Sinha and his collaborators identified OOD failures that still hinder the performance of autonomous vehicles. They then set out to investigate the extent to which existing OOD detection methods could uncover these failures.

"For example, existing methods that track visual novelty did poorly at

detecting these particular cases, as seeing stop signs, billboards or similar objects is not visually novel compared to training data, it is only once such objects appear on billboards that they become anomalous," Sinha said.

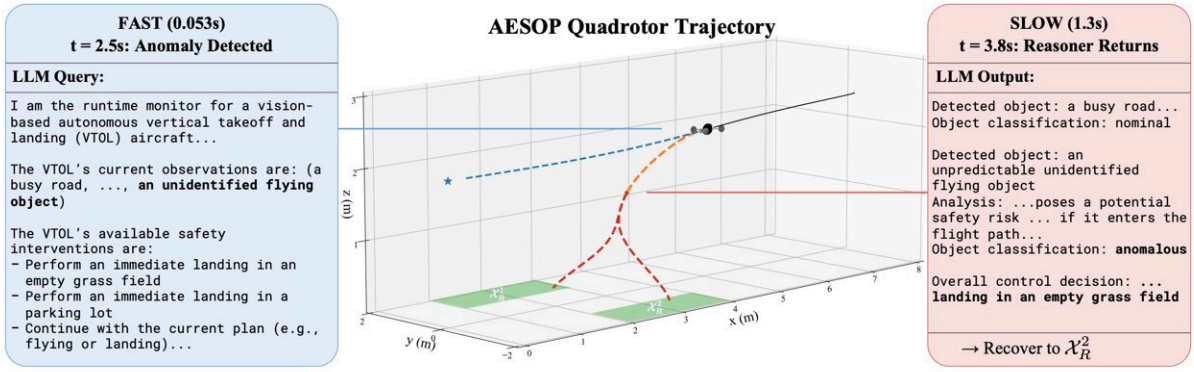
"In addition, we found that these kinds of failure modes are not easy to attribute to a specific component (e.g., a perception system) failure, rather they reflect system-level deficiencies in contextual reasoning. This makes them difficult to catch with existing component-level monitoring techniques."

In a 2023 paper, the researchers demonstrated the potential of LLMs for detecting and understanding these "semantic anomalies." Yet to effectively use these models to avert OOD failures affecting autonomous robots, they first had to overcome two key research challenges.

"First, we had to mitigate the computational costs of LLMs to enable real-time reactivity—the best LLMs are very large, and this makes them very slow, which is not very practical for a fast-moving robot," Sinha said.

"Second, we need to integrate LLM-based reasoners into the control of dynamic and agile robots. The goal of our recent paper was to address these two key challenges and thereby demonstrate that LLMs can significantly increase the safety of autonomous robots."

Compared to other computational models, LLMs can be slow in processing information. The main reason for this is that to create a new text, they autoregressively and individually generate tokens. To generate a chain-of-thought-like text that reasons what a robot should do (i.e., planning a robot's actions), the transformer models underpinning the LLM thus need to predict hundreds or even thousands of tokens one-by-one.



Closed-loop trajectory of a quadrotor using the AESOP algorithm. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2407.08735

"To overcome this limitation, we propose a 2-stage reasoning pipeline, where the first (fast) stage leverages intermediate outputs, a single embedding resulting from a single forward pass through a transformer model, to enable low-latency reactivity," Sinha explained.

"In the second (slow) stage, we still rely on the full generative chain of thought capabilities of the largest models to make zero-shot decisions on OOD scenarios that have never been recorded in data before."

Sinha and his colleagues first created a database of semantic embedding vectors using a foundation LLM model offline and an existing dataset of nominal experiences. At runtime, the team's framework embeds what a robot is currently observing and computes the similarity of the observation's embedding to those included in the embedding dataset. This is their model's first stage (i.e., the fast stage).

"If the observation is similar to prior observations, we continue with the

decisions made by the base autonomy stack," Sinha said. "If the observation is anomalous, we query a large model to reason about what safety-preserving intervention to take (stage 2: slow). We paired this 2-stage reasoning framework with a model predictive control (MPC) framework that plans multiple fallbacks and takes the latency of the slow reasoner into account."

With these two steps, the framework allows a robot to rapidly detect an anomaly and slow down its actions, so that an LLM model can reason about what can be done to mitigate failures. The adaptive plan proposed by the LLM is then executed by the robot.

Sinha and his colleagues evaluated their proposed framework in a series of tests and found that it could enhance anomaly detection and reactive planning in autonomous robotic systems. Notably, their approach was found to outperform other methods that solely rely on the generative reasoning of LLMs.

"Interestingly, we found that smaller models (e.g., MPNet with 110M params) can do just as well as larger models (e.g., Mistral 7B) at anomaly detection," Sinha said. "Embedding-based anomaly detectors are really good at detecting when observations are different from prior experiences, whereas zero-shot chain-of-thought reasoning with large models is really necessary to determine the safety criticality of an OOD scenario and the appropriate fallback."

Overall, the recent work by this team of researchers suggests that the deployment of both fast and slow reasoning can improve the performance and practicality of using LLMs for anomaly detection and robotic planning tasks. In the future, their framework could facilitate the use of LLMs to enhance the robustness of robots, potentially contributing to the improvement of various autonomous robotic systems.

"Our fast reasoners run in real-time, approximately 360X faster than querying GPT-4, while slow reasoning with GPT-4 achieved the highest accuracy at determining the safety risks of nuanced anomalies in our evaluations," Sinha added

"We now plan to continue building upon this framework. For example, we plan to use continual learning based on the delayed anomaly assessment of the generative reasoner to avoid triggering the slow reasoner on non-safety-critical anomalies a second time."

More information: Rohan Sinha et al, Real-Time Anomaly Detection and Reactive Planning with Large Language Models, *arXiv* (2024). DOI: 10.48550/arxiv.2407.08735

Citation: A two-stage framework to improve LLM-based anomaly detection and reactive planning (2024, August 15) retrieved 15 August 2024 from https://techxplore.com/news/2024-08-stage-framework-llm-based-anomaly.html