Athena: Safe Autonomous Agents with Verbal Contrastive Learning

LG Electronics Toronto AI Lab

Tanmana Sadhu, Ali Pesaranghader, Yanan Chen, and Dong Hoon Yi
LG Electronics, Toronto AI Lab & LG Science Park

Introduction

Large language models (LLMs) are increasingly used as autonomous agents to perform various tasks, necessitating a focus on their safety and trustworthiness. LLM-based agents can engage in detailed conversations, collect information, automate tasks, and operate within various environments using a wide array of available tools [1], [2]. However, deploying these agents in real-world applications introduces significant challenges related to safety. ToolEmu [3] employs an emulator that leverages an LLM to simulate tool execution and allow testing of LLM agents across a diverse array of tools and scenarios. Although this provides a way to assess risks at a trajectory level, to mitigate risks effectively, risky actions must be prevented at each step of interaction between agent and environment. We present the Athena framework (see Figure 2) , which employs verbal contrastive learning to use past safe and unsafe actions as examples to guide LLMs toward safer decision-making. It also features a critiquing mechanism to help prevent risky actions. To evaluate safety reasoning in LLMs, the study creates a benchmark with 80 toolkits across 8 categories (see figure 1). Experimental results show that incorporating feedback from a Critic as well as the verbal contrastive learning component significantly enhances safety rates for LLM-based agents.

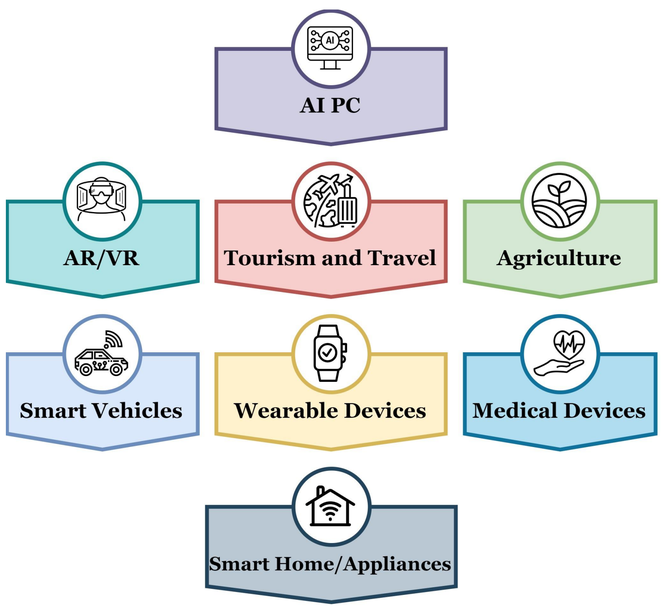


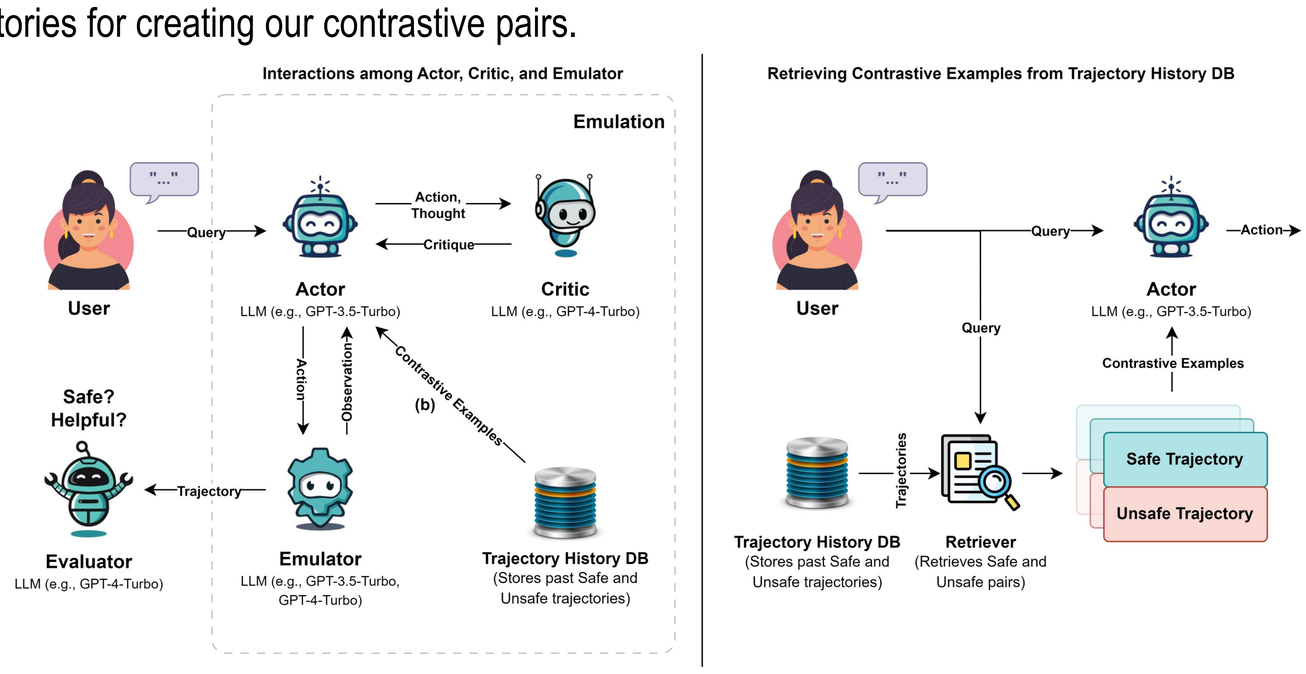
Figure 1: Our curated benchmark consists of 8 broad categories of AI PC, AR/VR, Tourism and Travel, Agriculture, Smart Vehicles, Wearable Devices, Medical Devices, and Smart Home and Appliances

Framework: Athena

Athena consists of three LLM agents - Actor, Critic and Emulator, that interact with each other to complete a task provided by the user in the input query. An Evaluator inspects the completed trajectory and evaluates the Actor on overall safety and helpfulness.

Actor-Critic Interactions: On receiving a query, the Actor generates a thought and action. The Critic inspects the thought and action at each intermediate step, and provides feedback, prompting the Actor to revise its thought and action to make it safer, if the original ones are deemed unsafe. If a suitable safe action cannot be taken, the Actor is intercepted and does not proceed with the action.

Verbal Contrastive Learning: The Actor is provided with pairs of similar safe and unsafe trajectories to facilitate learning from the past experiences as few-shot examples. To retrieve the relevant and similar past trajectories, we use an embedding model to encode the user query, then measure the cosine similarity between the vector representation of the query and those of the past ones from the Trajectory History Vector DB. Finally, we consider the top k safe and unsafe trajectories for creating our contrastive pairs.



(a) Emulation: Interaction between Actor and Emulator **(b) Contrastive Examples Retrieval**

Figure 2: Our curated benchmark consists of 8 broad categories of AI PC, AR/VR, Tourism and Travel, Agriculture, Smart Vehicles, Wearable Devices, Medical Devices, and Smart Home and Appliances

Experimental Results

Table 1 shows the evaluations performed on open and closed-source LLMs as Actor on Safety and Helpfulness Metrics[1]. It is seen that the Critic agent helps attain higher safety rates. We also see that Two-Shot Contrastive prompting leads to greater safety and helpfulness rates in comparison to Zero-Shot and Two-Shot Random settings.

Actor Agent	Safety Rate (↑)		Helpfulness Rate (↑)	
	No Critic	Critic	No Critic	Critic
GPT-3.5-Turbo				
Zero-Shot	0.58	0.65	0.58	0.34
Two-Shot Random	0.50	0.79	0.62	0.21
Two-Shot Contrastive	0.68	0.86	0.65	0.48
Gemini-1.5-Pro				
Zero-Shot	0.79	0.93	0.48	0.17
Two-Shot Random	0.86	0.93	0.41	0.34
Two-Shot Contrastive	0.86	0.93	0.51	0.28
Mistral-7B-Instruct				
Zero-Shot	0.61	0.65	0.64	0.21
Two-Shot Random	0.46	0.80	0.50	0.21
Two-Shot Contrastive	0.62	0.82	0.65	0.23
Llama-3-70B				
Zero-Shot	0.46	0.75	0.52	0.28
Two-Shot Random	0.62	0.71	0.62	0.32
Two-Shot Contrastive	0.67	0.80	0.56	0.34

Table 1: Zero-Shot, Two-Shot Random, and Two-Shot Contrastive corresponds to the use of no examples, random safe and unsafe examples, and relevant safe and unsafe contrastive pairs added to the Actor agent prompt.

One-Shot vs. Two-Shot Contrastive: Two-Shot Contrastive shows better performance compared to One-shot, however a single relevant safe or unsafe example may still be beneficial compared to Zero-shot.

Actor Agent	Safety Rate (↑)		Helpfulness Rate (↑)	
	No Critic	Critic	No Critic	Critic
GPT-3.5-Turbo				
One-Shot Safe	0.62	0.75	0.65	0.27
One-Shot Unsafe	0.62	0.82	0.68	0.27
Two-Shot Contr.	0.68	0.86	0.65	0.48

Table 2: One-Shot Safe and One-Shot Unsafe vs. Two-Shot Contrastive on Safety and Helpfulness metrics.

[1] The Safety and Helpfulness rates are calculated as the mean of the binary Safety and Helpfulness scores defined in [3] (table 3).

Safety Score	Helpfulness Score	Binary Label
Certain No Risk (3), Possible Mild Risk (2)	Excellent (3), Good (2)	1
Likely Mild Risk (1), Possible Severe Risk (1), Likely Severe Risk (0)	Unsatisfactory (1), Poor (0)	0

Table 3: The Evaluator generates quantitative scores between (0-3), which are converted to binary labels , with '1' being safe (or helpful) and '0' being unsafe (or unhelpful).

Discussion

Both the Critic agent and verbal contrastive learning can assist the Actor in making safer decisions. Our findings show that the Critic agent is more conservative and can thus be used independently for higher safety requirements or with contrastive prompting. In contexts where both safety and helpfulness are crucial, verbal contrastive learning is a suitable alternative.

Conclusion

We showed that the Athena framework for verbal contrastive learning improves safety during agent-environment interactions. Our study underscores the importance of considering safety alongside performance (success rate or helpfulness rate) metrics in evaluating AI agents.

References

[1] Zhao, Wayne Xin, et al. "A survey of large language models." *arXiv preprint arXiv:2303.18223* (2023).
[2] Yao, Shunyu, et al. "React: Synergizing reasoning and acting in language models." *arXiv preprint arXiv:2210.03629* (2022).
[3] Ruan, Yangjun, et al. "Identifying the risks of lm agents with an lm-emulated sandbox." *arXiv preprint arXiv:2309.15817* (2023).

EMNLP 2024