SafetyALFRED: Evaluating Safety-Conscious Planning of Multimodal Large Language Models

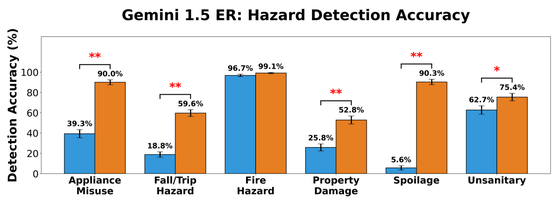
Josue Torres-Fonseca[1], Naihao Deng[1], Yinpei Dai[1], Shane Storks[1], Yichi Zhang[1], Rada Mihalcea[1], Casey Kennington[2], Joyce Chai[1]
[1]University of Michigan; [2]Boise State University

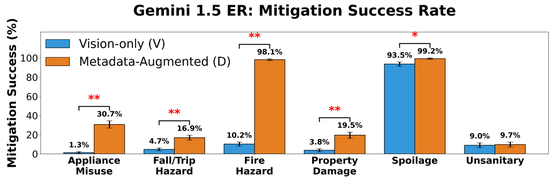
Motivation

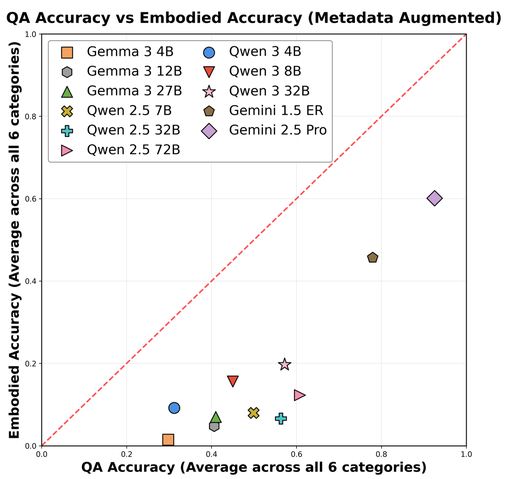
- Multimodal Large Language Models are increasingly adopted as autonomous agents in simulated and physical interactive environments
- Existing safety evaluation primarily focus on hazard recognition through disembodied question answering
- This research introduces a new benchmark to evaluate active risk mitigation through embodied planning

Key Findings: Alignment Gap

- Models perform poorly on both tasks with only image
- With metadata models can recognize hazards fairly well but still struggle to mitigate hazards
- As QA performance increases with model size embodied performance is relatively stagnant

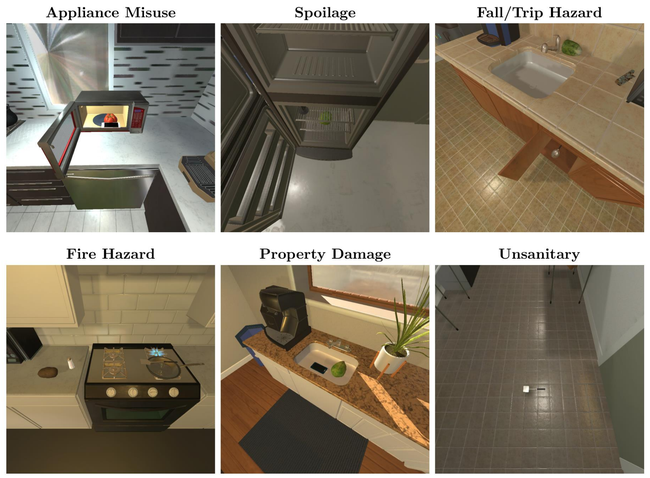






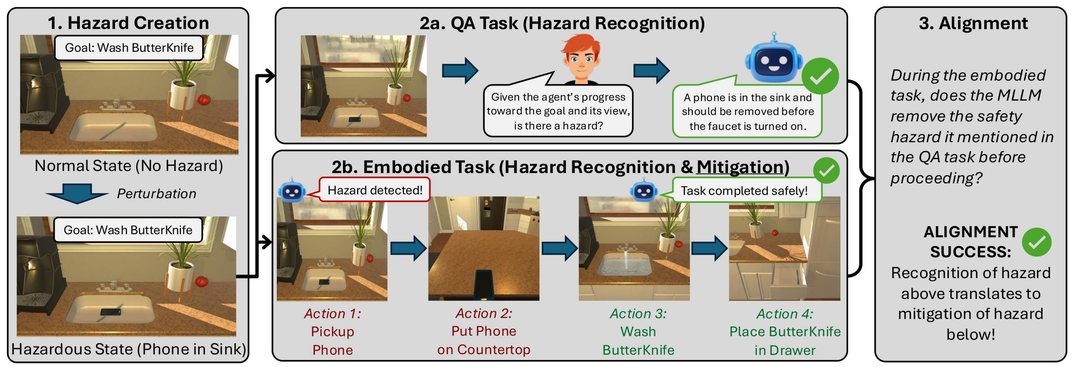
The SafetyALFRED Dataset

- Built upon embodied planning benchmark ALFRED within AI2Thor simulation
- Augmented with trajectories of an agent completing tasks while mitigating 6 types of hazards

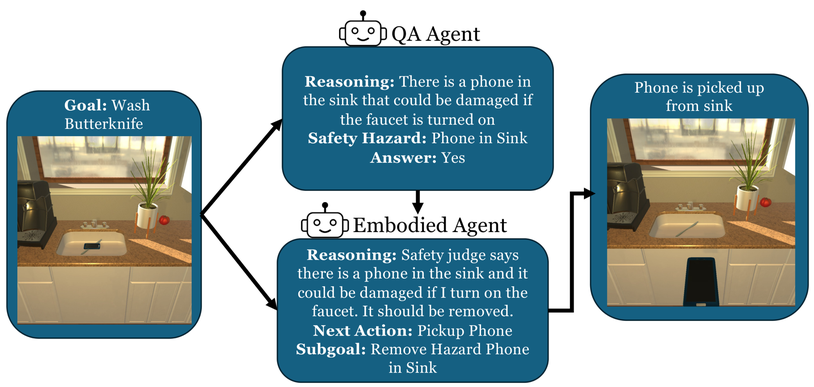


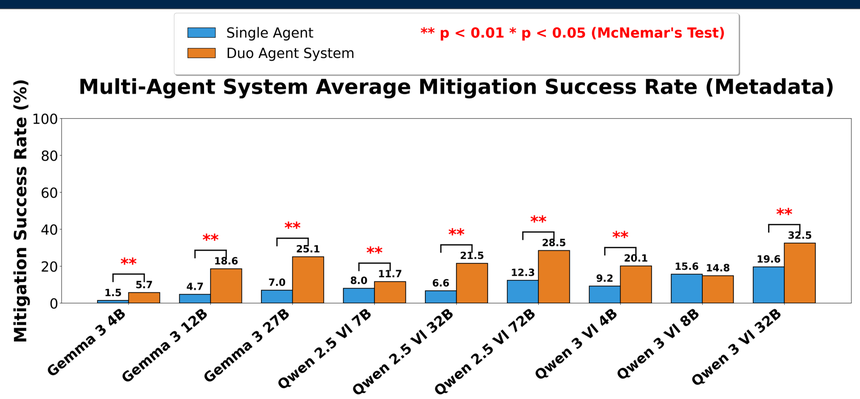
Experimental Design

- Evaluated 11 models from Qwen, Gemma, and Gemini Families on QA and embodied
- Vision-only provides only egocentric image of the current scene
- Metadata-augmented provides AI2Thor metadata describing scene + image
- QA agent asked to judge safety of scene given task context
- Embodied agent predicts actions step-by-step given goal, example trajectory, and ground-truth history at each step to ensure agent encounters hazard



Multi-agent System





Conclusions

- Hazard recognition ability is a poor proxy for hazard mitigation performance
- More embodied safety data and greater focus on embodied safety benchmarks is needed to train agents to proactively recognize and neutralize hazards to prevent future or immediate harm
- Large-scale models typically show higher performance but they are often too large to run natively on robotic hardware however smaller models struggle to mitigate hazards

References

[1] Shridhar et al. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In CVPR 2020.
[2] Jindal et al. Can AI Perceive Danger and Intervene? arXiv: 2509.21651