Evaluating and Enhancing Trustworthiness of LLMs in Perception Tasks

Malsha Mahawatta, University of Gothenburg, Sweden
Department of Computer Science and Engineering

Can an LLM support perception related tasks within a vehicle?

Yes, It can! Read paper [1]

How well can we trust an LLM within a vehicle?

Research Goals

RQ1 : What are potential hallucination detection strategies suggested in the literature?

RQ-2: How can hallucinations be characterized when applying LLMs to pedestrian detection and localization for ADAS/AD?

RQ-3: How can hallucination detection strategies be enhanced for use in ADAS/AD perception and monitoring systems?

Hallucination detection strategies

- SelfCheckGPT [2]
- BO3 [3]
- DreamCatcher [4]

Types of hallucinations

	GT	LLM
FN		
FP		

LLM rejecting the processing of the image

LLM performance on unmodified images and ROIs

- GPT4V - Higher recall and F1 Score values
- LLaVA - Lower recall and F1 score values

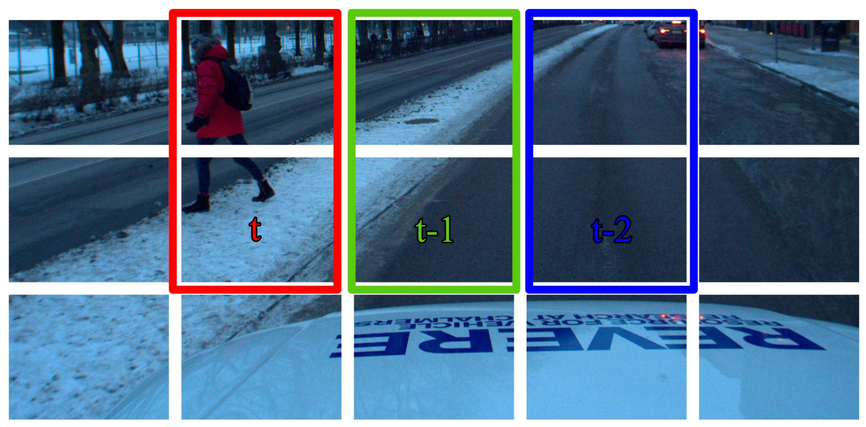
Results

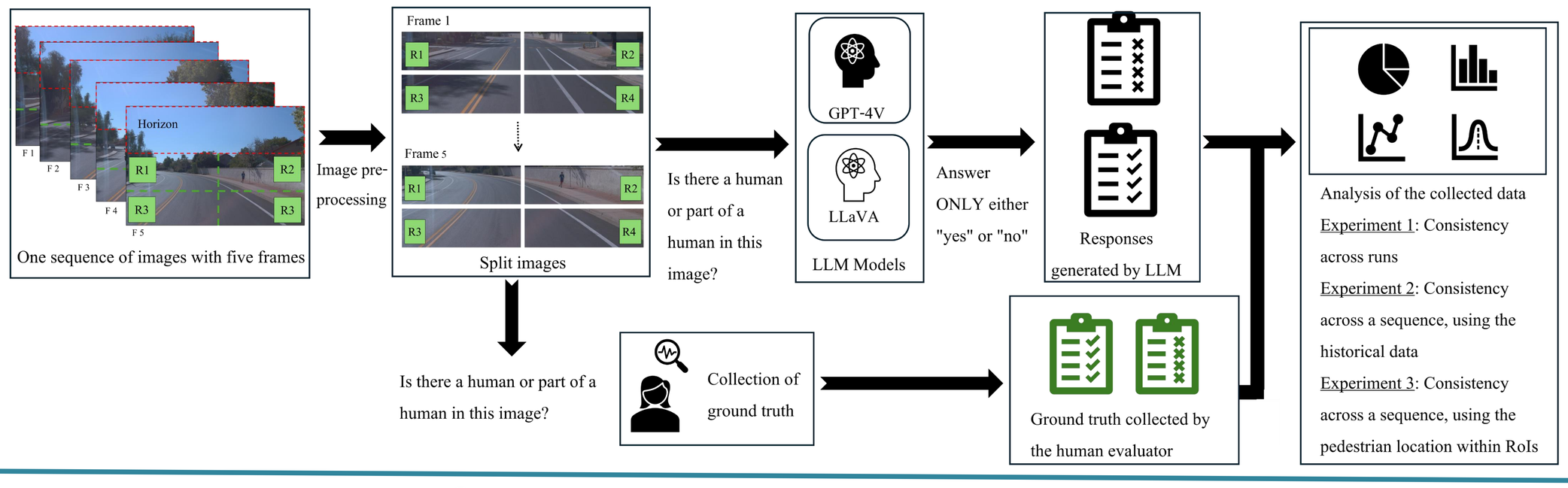
Hallucination detection using historical frames in an automotive context

THV – Three consecutive frames (f,f-1,f-2)
If one or both (f-1,f-2) has a pedestrian, t is identified as a frame with pedestrian.

THV2 – THV was applied to correct the LLMs response for frame f.

Physical Plausibility Check





References

1. Dona, M.A.M., Cabrero-Daniel, B., Yu, Y. and Berger, C., 2024. Tapping in a Remote Vehicle's onboard LLM to Complement the Ego Vehicle's Field-of-View. *arXiv preprint arXiv:2408.10794*.
2. Manakul, P., Liusie, A. and Gales, M.J., 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.
3. Ronanki, K., Cabrero-Daniel, B. and Berger, C., 2022, June. ChatGPT as a Tool for User Story Quality Evaluation: Trustworthy Out of the Box?. In *International Conference on Agile Software Development* (pp. 173-181). Cham: Springer Nature Switzerland.
4. Liang, Y., Song, Z., Wang, H. and Zhang, J., 2024. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation. *arXiv preprint arXiv:2401.15449*.

Contact

https://malshamahawatta.github.io/
malsha.mahawatta@gu.se

WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM