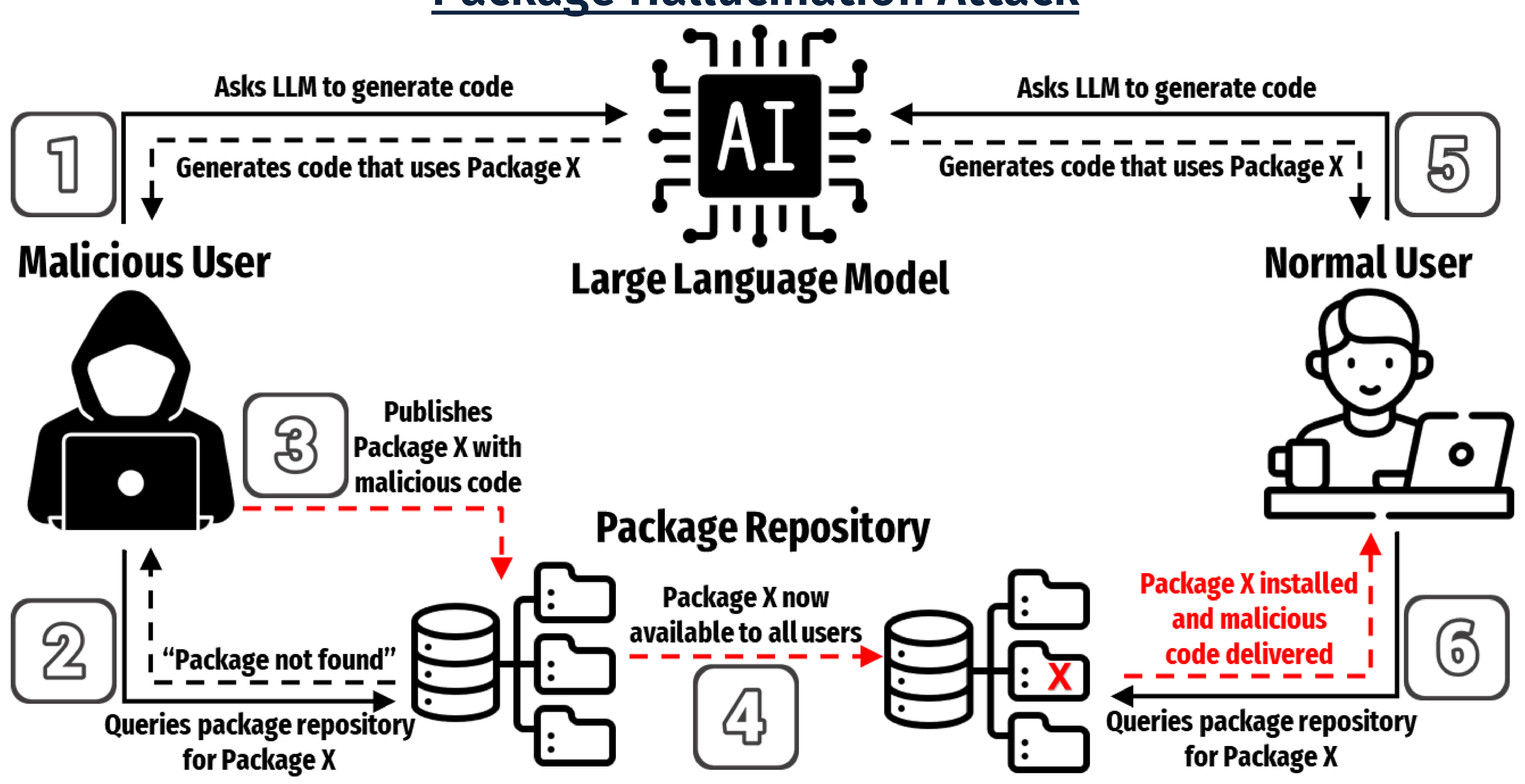
Special Delivery! Investigating the Prevalence, Causes, and Mitigation Methods of Package Hallucination in Code Generating LLMs

UTSA Computer Science

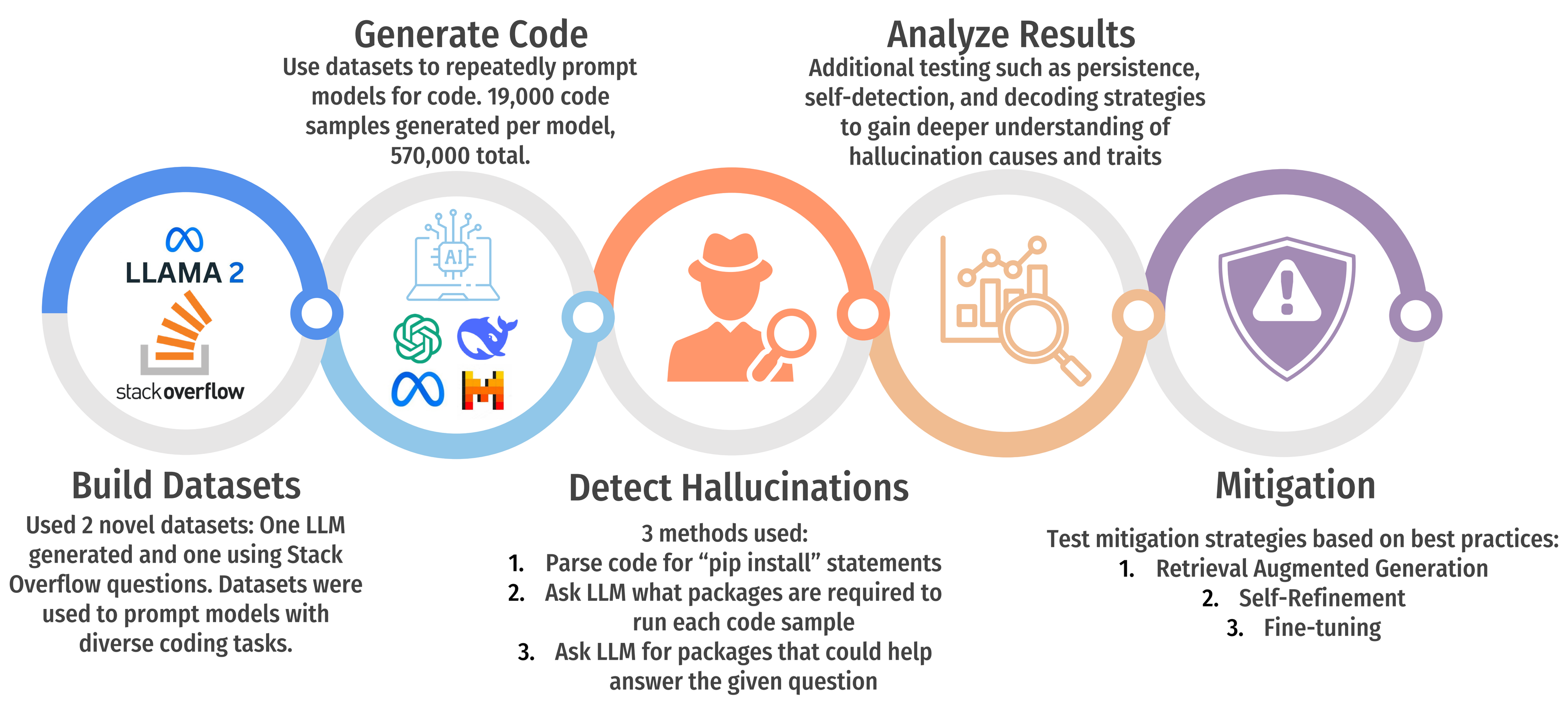
Joseph Spracklen[1], Raveen Wijewickrama[1], Nazmus Sakib[1], Anindya Maiti[2], Murtuza Jadliwala[1]
[1]University of Texas at San Antonio
[2]University of Oklahoma

SPriTE LAB

Package Hallucination Attack



Experiment Design



Models Tested

Number of Parameters	1B	7B	14B	34B	70B+
OpenAI					● ● ●
Meta / code llama		●	●	●	●
deepseek	●	●		●	
MISTRAL AI_		●		●	
ISE Magicoder		●			
WizardCoder	●	●		●	

Key Findings

01 Prevalence
- 19.4% of all packages generated were hallucinated (i.e. non-existent, fictitious)
- 205,474 unique hallucinated package names were generated

02 Persistence
- 48% of the time a hallucinated package will be repeated when given the same prompt
- A hallucination will repeat within 10 iterations 60% of the time

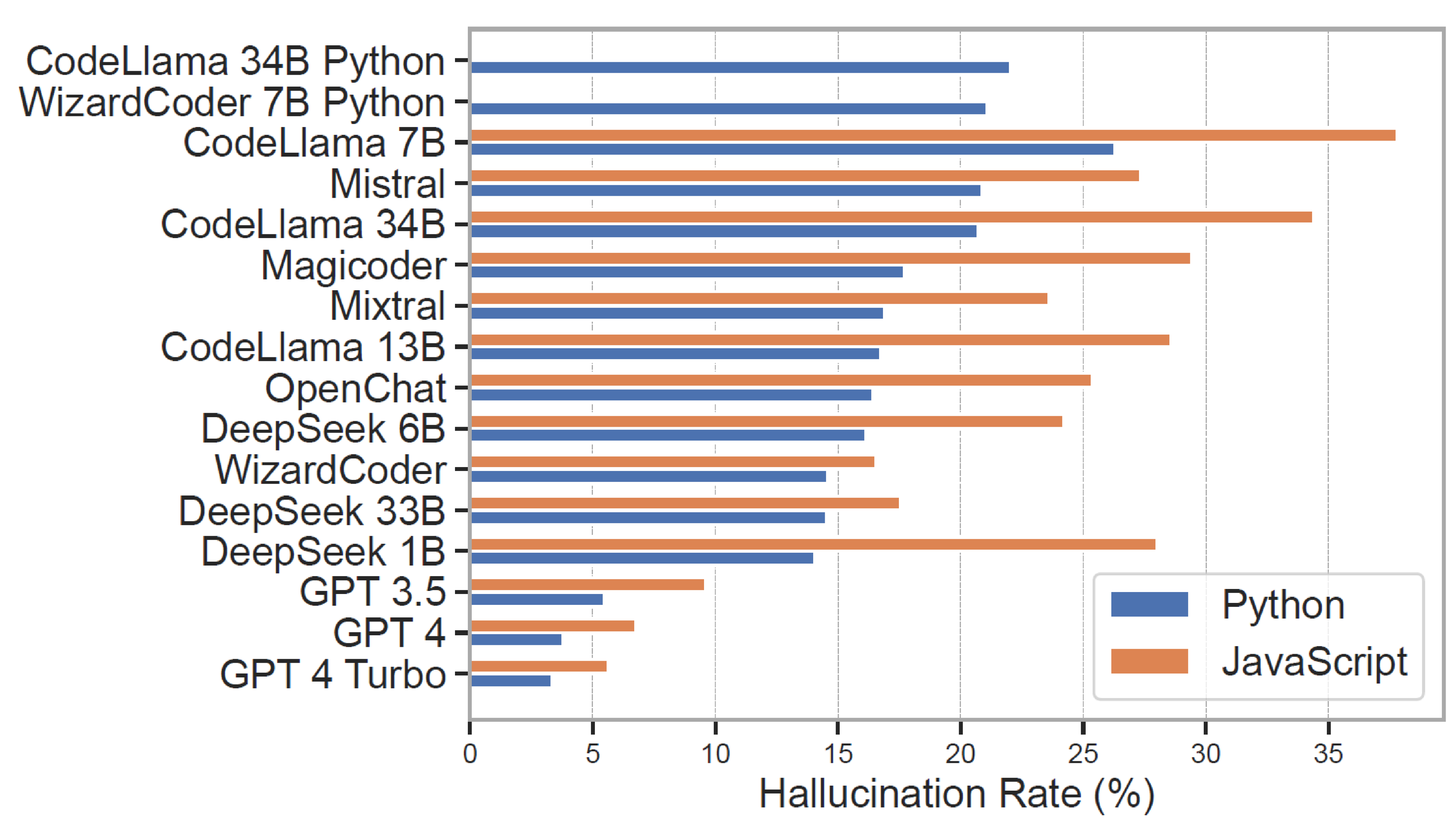
03 Self-Detection
- 3 out of 4 models were able to correctly identify their own hallucinations more than 75% of the time

04 Decoding Strategies
- Higher temperatures dramatically increase hallucinations
- Package hallucinations are usually generated by the most probable tokens

05 Mitigation
- Fine-tuning is an extremely effective mitigation strategy for package hallucination
- Combining mitigation methods brings hallucination rate below ChatGPT

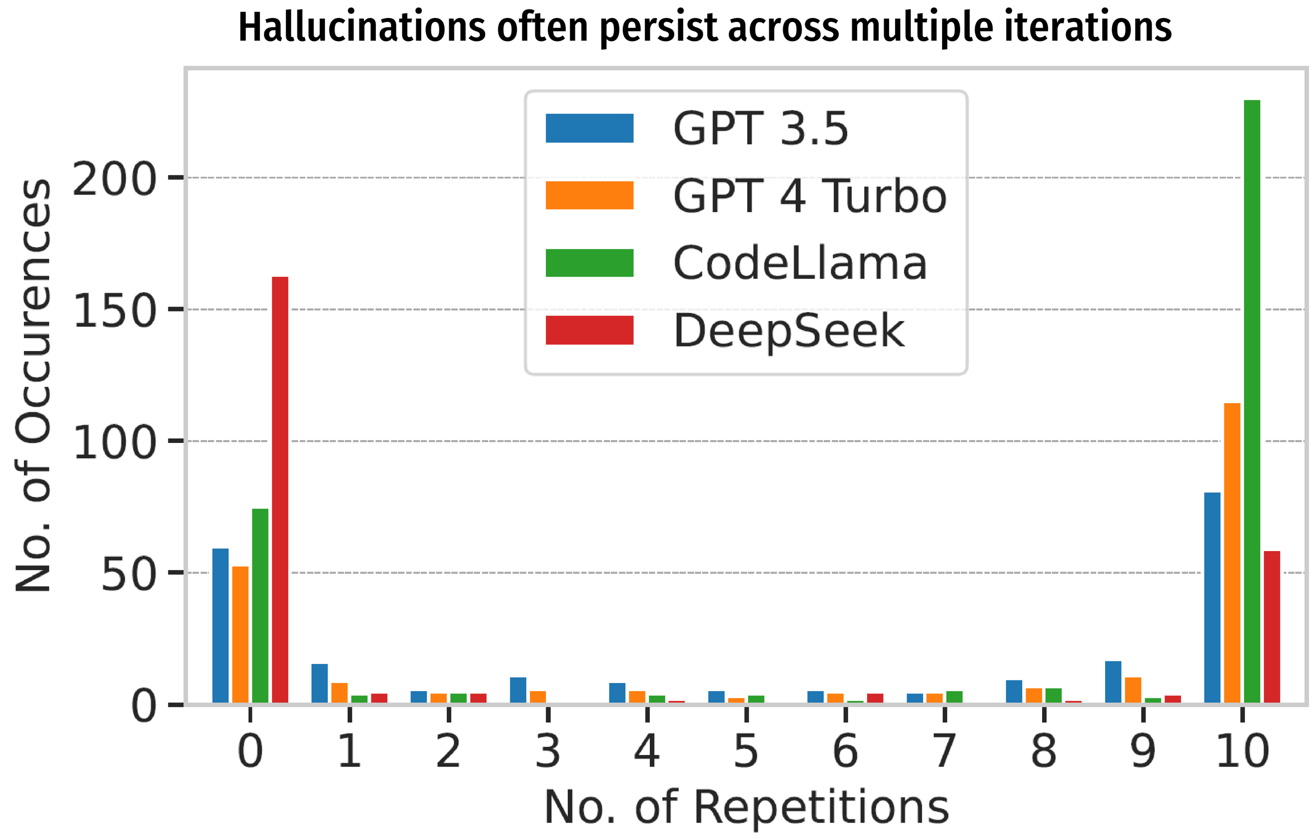
01 PREVALENCE

Package hallucinations occur at a significant rate across all models tested



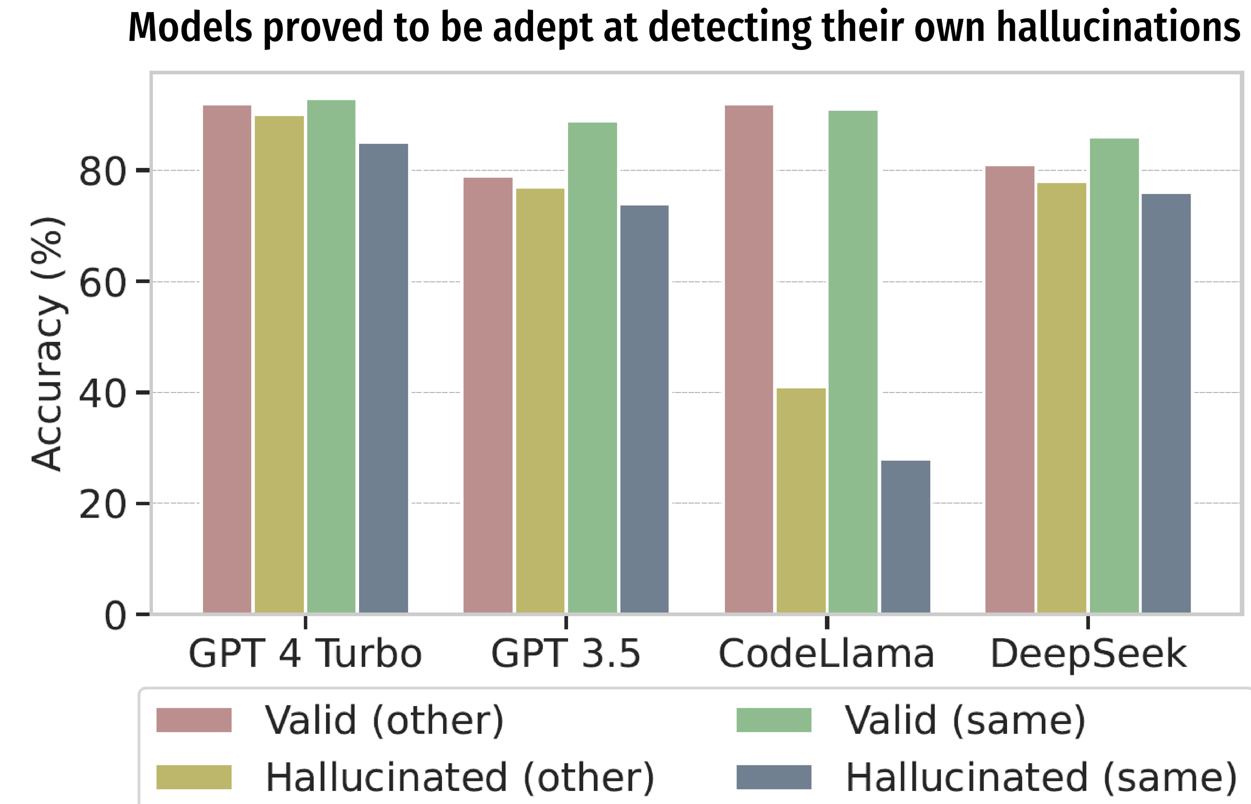
02 PERSISTENCE

Hallucinations often persist across multiple iterations



03 SELF-DETECTION

Models proved to be adept at detecting their own hallucinations



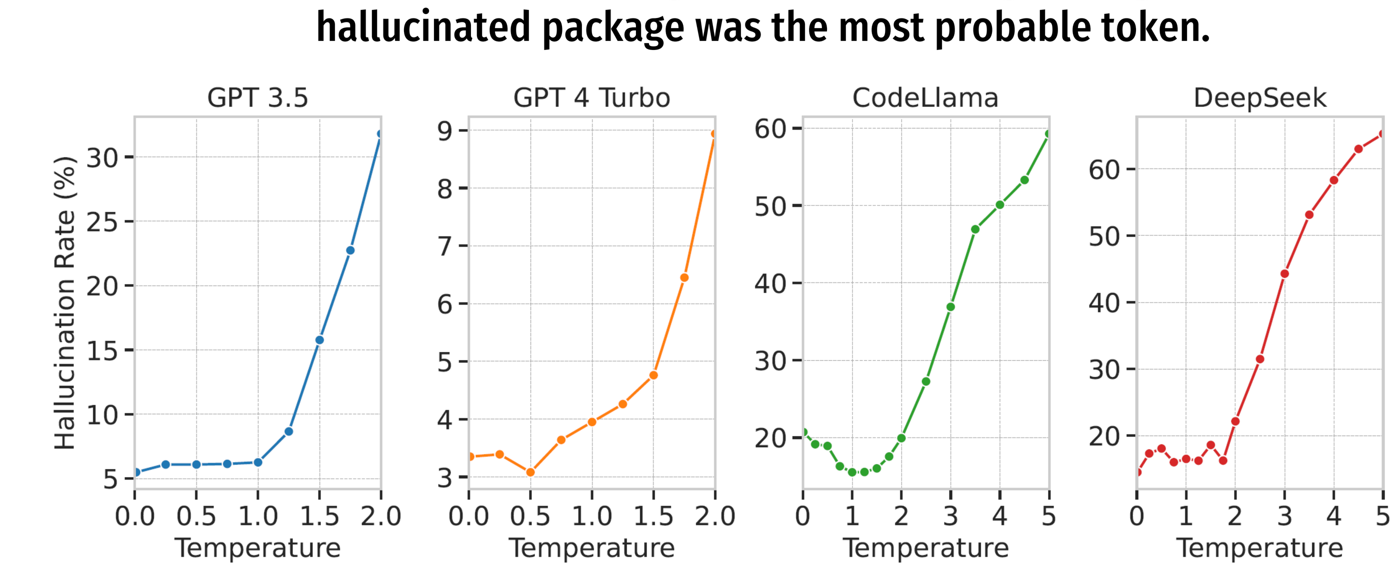
05 MITIGATION

Hallucinations were reduced using best practices but not eliminated

	DeepSeek	CodeLlama
Baseline (No Mitigations)	16.14%	26.28%
Retrieval Augmented Generation (RAG)	12.24%	13.40%
Self-Detected Feedback	13.04%	25.51%
Fine-tuning	**2.66%**	**10.27%**
Ensemble	**2.40%**	**9.32%**

04 DECODING STRATEGIES

Hallucination rate was unchanged across decoding strategies. Often the hallucinated package was the most probable token.



Altering decoding strategies produced higher hallucination rates in all cases

	DeepSeek	CodeLlama
Baseline (Default Decoding)	**16.14%**	**26.28%**
Top-k Lower	17.1%	27.8%
Top-k Higher	18.1%	28.3%
Top-p Lower	17.5%	28.0%
Top-p Higher	18.4%	28.3%
Min-p Lower	17.8%	27.9%
Min-p Higher	19.2%	28.6%