CSCI 566: DEEP LEARNING AND ITS APPLICATIONS

ARGUS

AI REGULATORY & GOVERNANCE FOR UNBIASED AND SAFE LLM SYSTEMS

Presentation ID: 31

Team Name: Adventures of Dh2 and JV

Dhruv Shetty and Jayavibhav Kogundi

MODELS AND METHODOLOGY

What is ARGUS?

ARGUS is an Agentic Pipeline designed to enhance safety and reliability of AI generated content. It integrated advanced models to address critical challenges like prompt safety, bias and hallucinations.

Motivation:

- AI-generated content often contains biases, hallucinations, or explicit material, leading to ethical and safety concerns.
- Current governance frameworks lack comprehensive oversight of AI outputs.
- A unified, scalable solution is needed to enforce ethical guidelines and regional compliance across AI-generated text.

Agent Controller:

- Model: LLama 3.1 8B
- Method:

Prompt Injection Detection:

- Model: DistilBERT
- Custom Dataset : 50k Samples

Prompt Engineering:

- Model: NVIDIA LLaMA 3.1 Nemotron-70B
- Method: agent-based adaptive prompt improvement

G**eneration Models**: Llama 3.1 8B

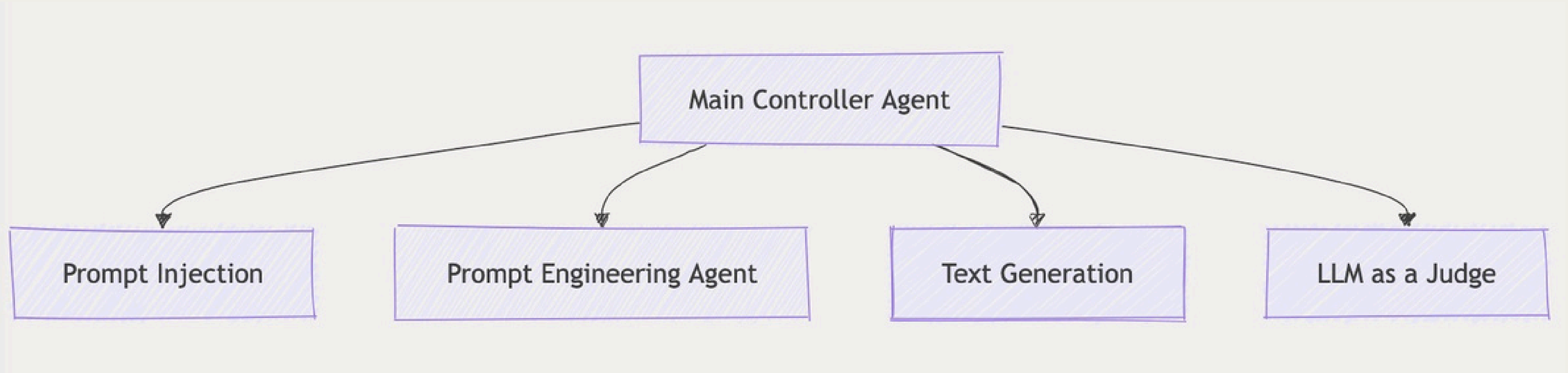
Hallucination, Bias, and Correctness Checks:

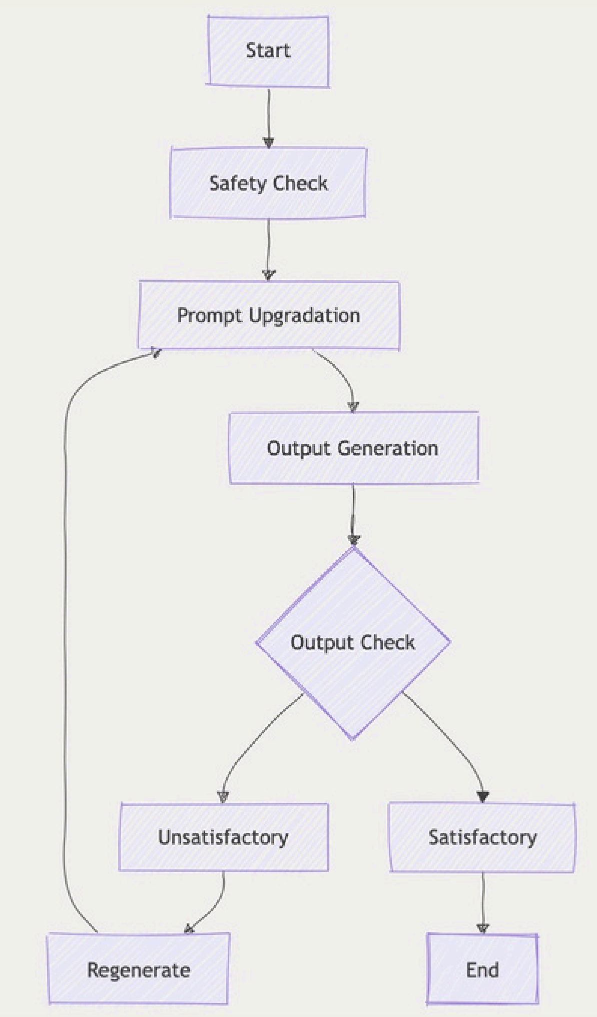
- Model: Nemotron
- Method: LLM as a judge
  - Improvisation: regeneration with annotations

RAG Pipeline:

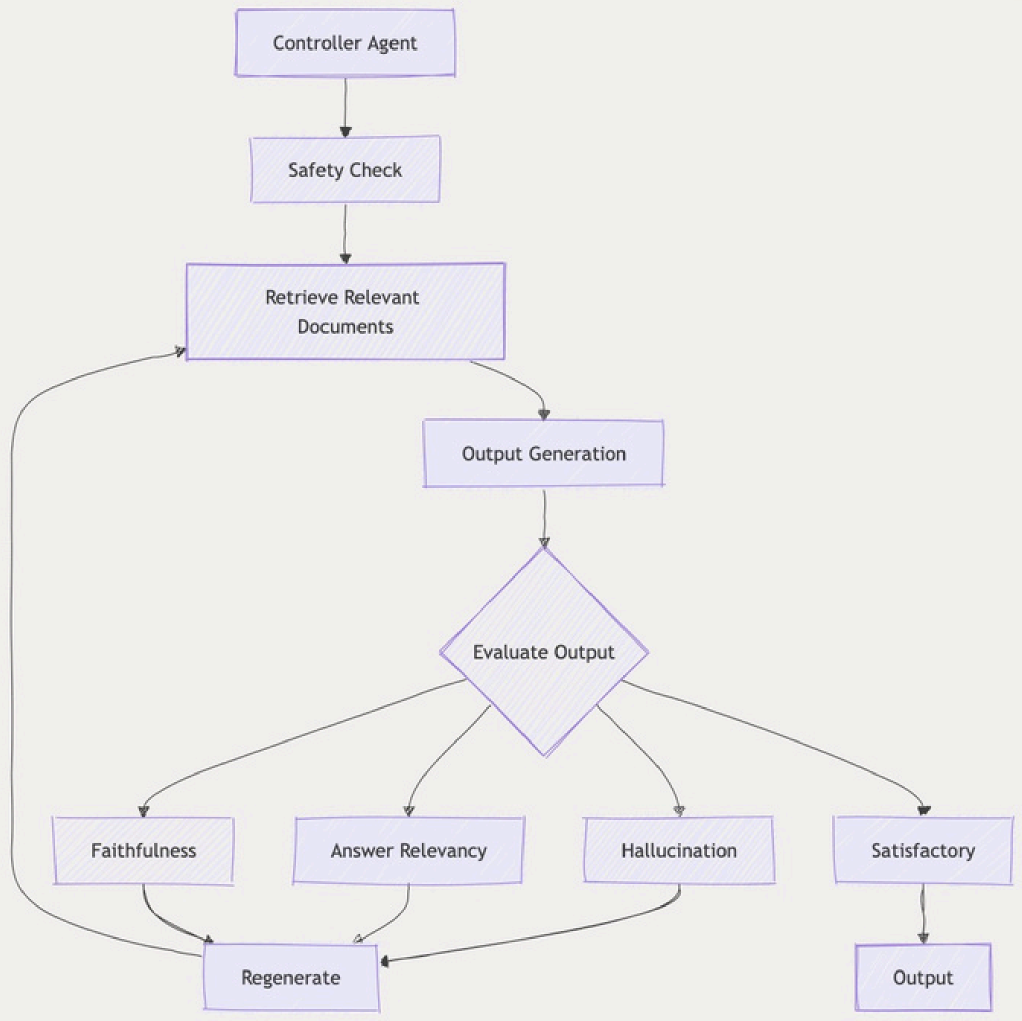
- Model: Llama 3.1 8B
- Vector Database: FAISS
- Evaluation: RAGAS

Agent Controller Working



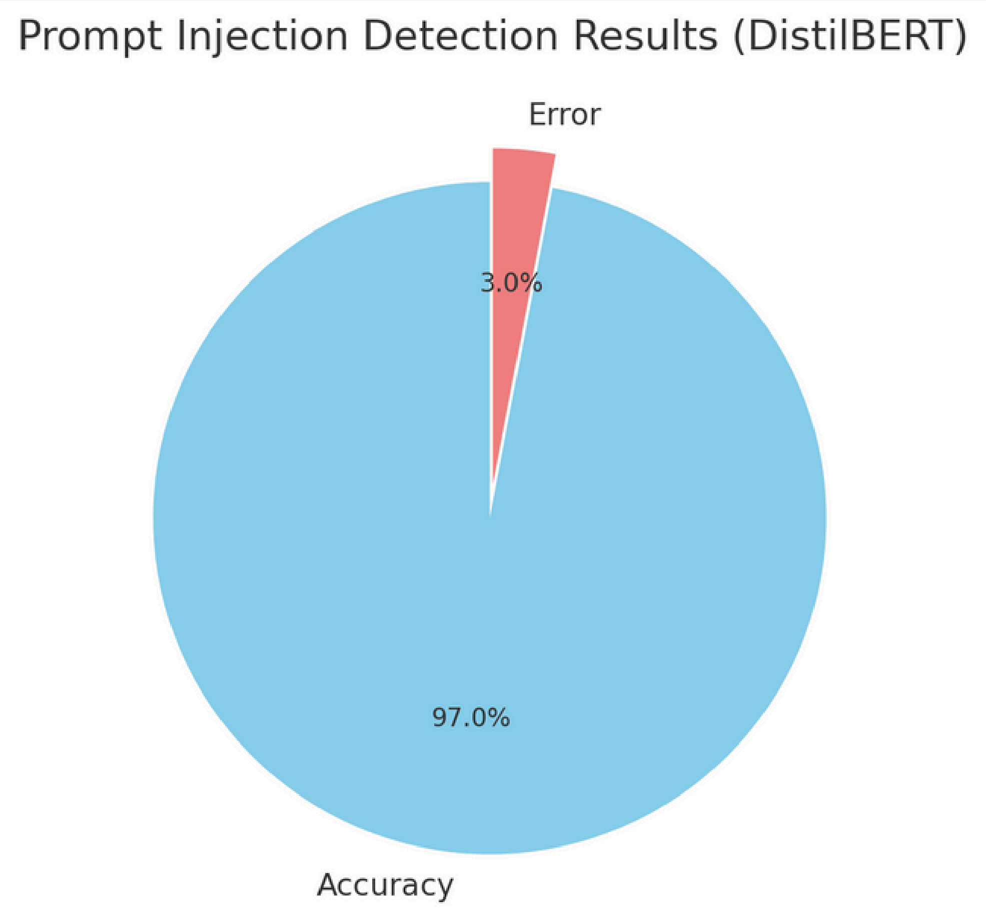


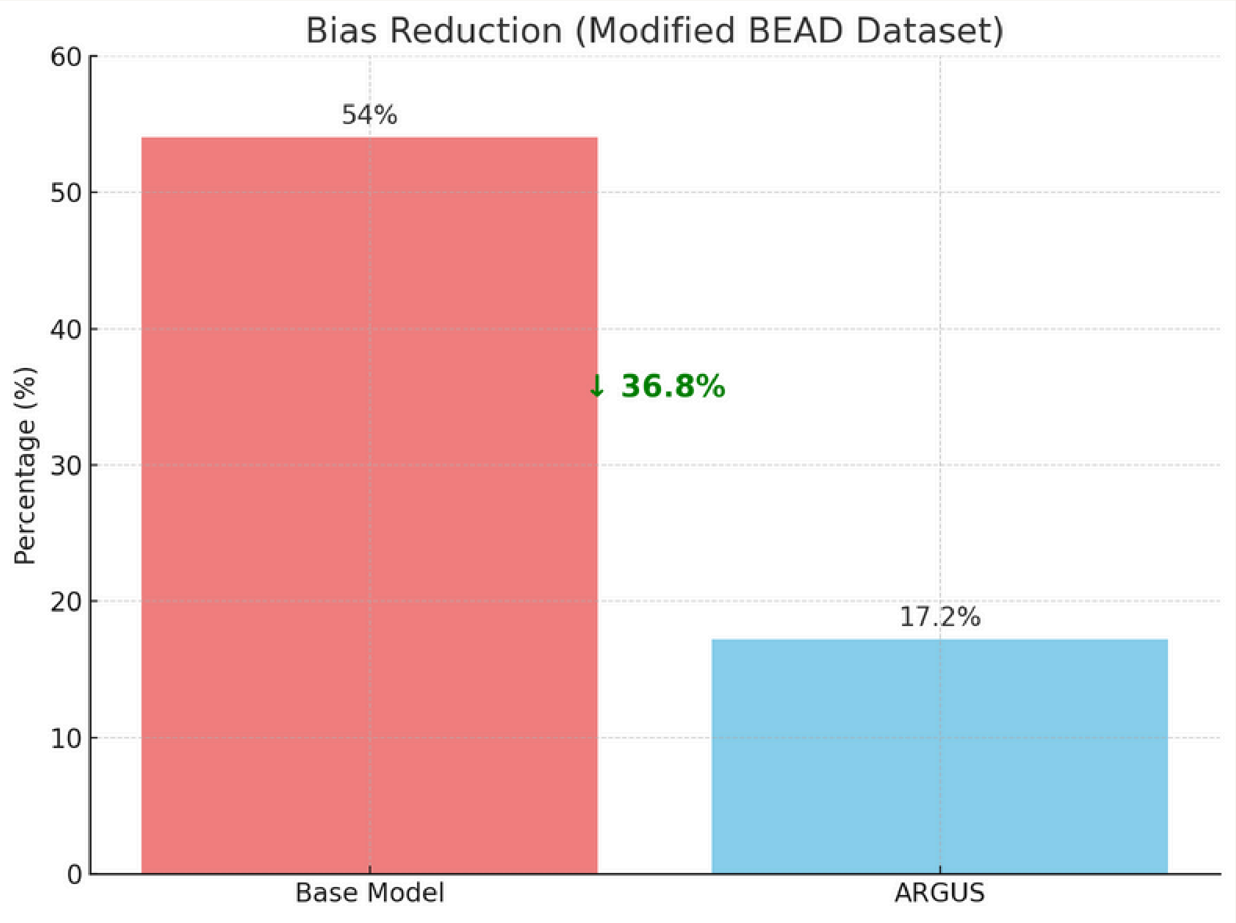
General Pipeline of ARGUS

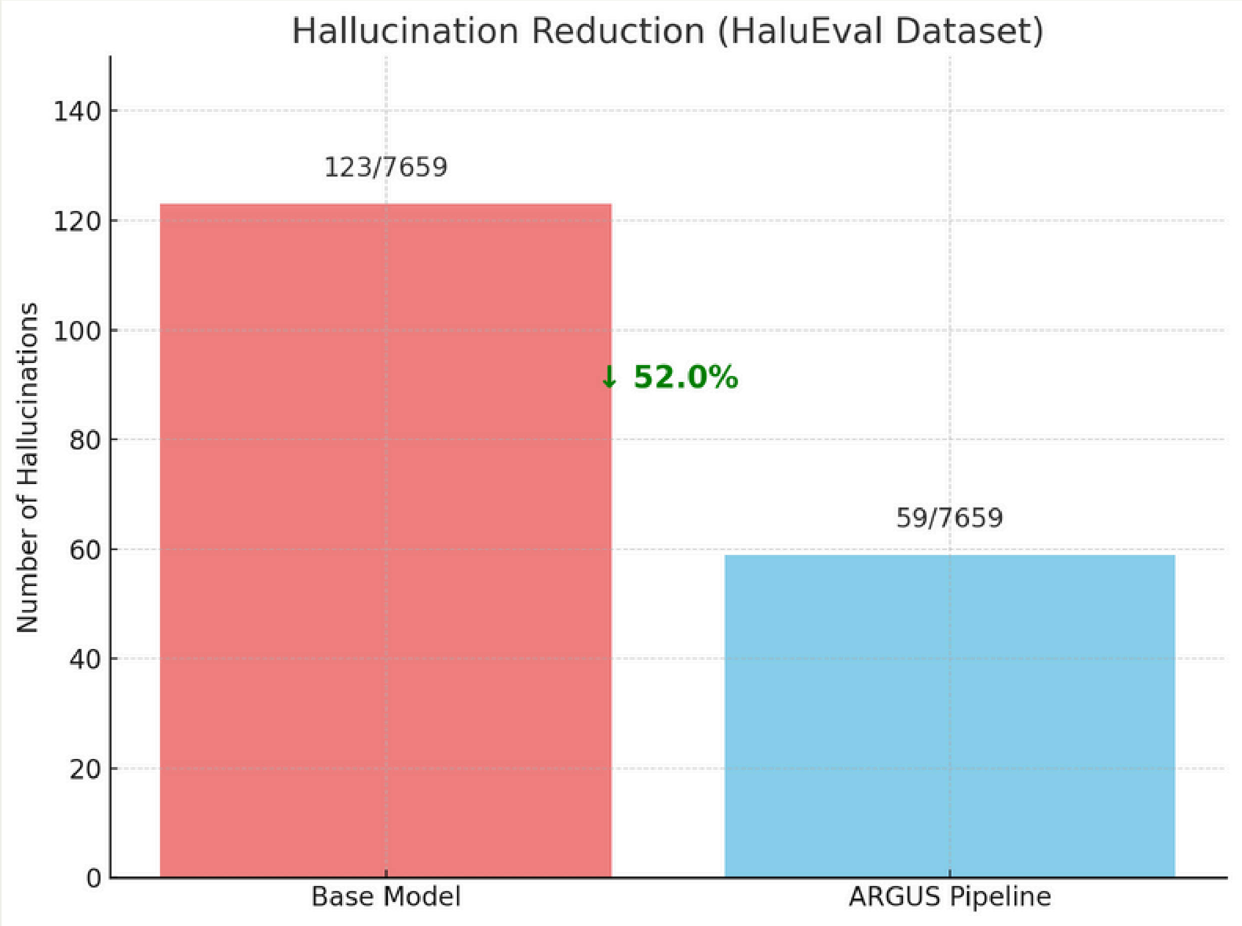


ARGUS with RAG

RESULTS







Prompt Injection Detection

- Model Used: DistilBERT
- Dataset:
  - Custom dataset with 50k training samples.
  - Tested on 10k samples.
- Purpose: Robust safety mechanism to classify:
  - Safe, Unsafe, and Prompt Injections.
- Performance:
  - Achieved 97% accuracy on the test set.

Bias Reduction

- Dataset Used: BEAD Dataset (Modified)
  - Synthetic dataset generated with 3500 rows using BEAD’s text generation subset.
  - Evaluation conducted on 1,000 samples
- Comparison:
  - Base Model Bias: 54%.
  - ARGUS Model Bias: 17%.
- Improvement: Reduced bias by 37%, ensuring more equitable model outputs.

Hallucination Evaluation

- Dataset Used: HaluEval Dataset
  - used the QA Subset of 7569 samples
  - Evaluates hallucinations using knowledge, question, ground truth, and judgment.
- Results:
  - Base Model: Significant hallucinations
  - ARGUS Model: 52% reduction in hallucinations.
- Improvement: Demonstrates enhanced reliability in handling knowledge-grounded questions.

Thank you!

QUESTIONS?

Email us

Dhruv Shetty - ddshetty@usc.edu

Jayavibhav Kogundi - jniranja@usc.edu