Enhanced Hallucination Detection in Large Language Models using FEVER Dataset

This presentation examines hallucination detection in LLMs - when models confidently generate factually incorrect information.

Shubham Gaur

The Challenge of Hallucinations

LLM Confidence

Models generate incorrect information while appearing certain.

Trust Issues

Hallucinations undermine trust in AI systems.

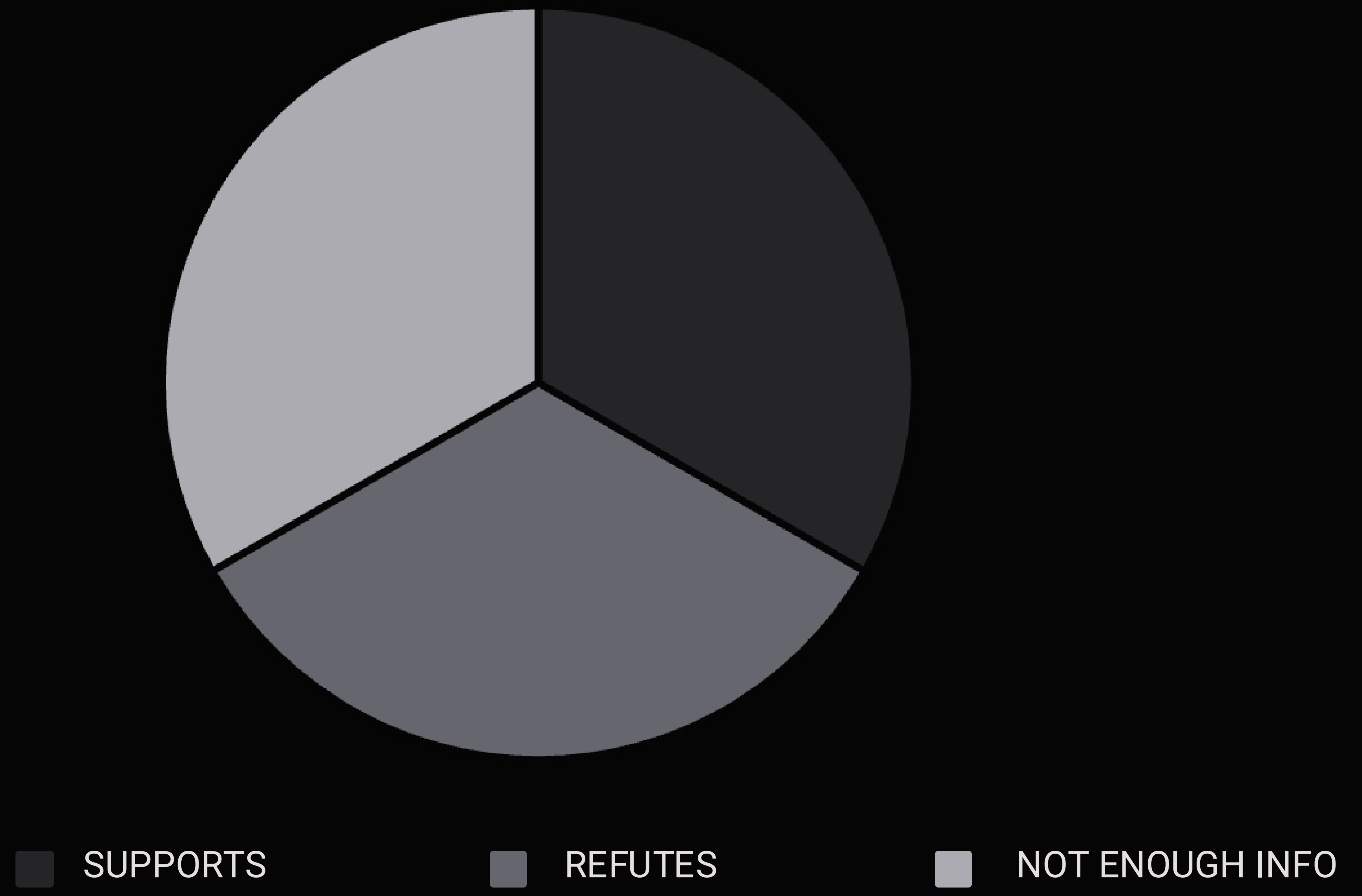
Detection Gap

Current methods struggle to identify hallucinated content.

Our project compares traditional ML (Random Forest, XGBoost) with LLM approaches (GPT-3.5, Claude) for hallucination detection.

FEVER Dataset Exploration

Class Balancing



We balanced 145,449 entries to 84,513 claims across three classes. Our dataset was split into train (59,159), validation (12,677), and test (12,677) sets.

FEVER Dataset Structure

FEVER (Fact Extraction and VERification) contains 145,449 human-generated claims for automatic fact verification.

SUPPORTS

Claims verified as true by evidence

REFUTES

Claims contradicted by evidence

NOT ENOUGH INFO

Claims that cannot be verified with available evidence

Each entry contains a claim, supporting evidence from Wikipedia, verification label, and unique ID.

```
{
  "id": 75397,
  "verifiable": "VERIFIABLE",
  "label": "SUPPORTS",
  "claim": "Nikolaj Coster-Waldau worked with the Fox Broadcasting Company.",
  "evidence": [[[92206, 104971, "Nikolaj_Coster-Waldau", 7],
                [92206, 104971, "Fox_Broadcasting_Company", 0]]]
}
```

Our preprocessing involved text cleaning, evidence formatting, and balanced sampling to achieve equal class distribution.

System Architecture

Data Processing

Cleaning, feature extraction, and split preparation.

Parallel Pipelines

Traditional ML, standard LLMs, and RAG-enhanced LLMs.

Evaluation System

Accuracy metrics and hallucination detection framework.

The RAG component creates a FAISS vector store for relevant evidence retrieval and augmentation.

Feature Engineering

Semantic Similarity

Used SentenceTransformer to measure content alignment.

Entity Overlap

Extracted with spaCy to detect named entity matches.

Captures factual agreement between claim and evidence.

Text Features

Length metrics correlate with verification difficulty.

Complex claims require more evidence.

Initial experiments showed LLMs struggle with REFUTES class, motivating our RAG enhancement approach.

Implementation Details

Feature Extraction

TF-IDF vectorization combined with numeric features.

Prompt Engineering

Specialized prompts for GPT and Claude models.

Metrics Framework

False positive/negative and disagreement calculations.

RAG Implementation

FAISS for efficient semantic search of evidence.

Custom transformers streamlined our pipeline for consistent evaluation across all approaches.

RAG Implementation

Retrieval-Augmented Generation combines an information retrieval system with generative AI to ground LLM responses in verified evidence.

Vector Store Creation

FAISS with HuggingFace embeddings for efficient similarity search.

Evidence Chunking

RecursiveCharacterTextSplitter (500 chars, 50 overlap).

Similarity Search

Claim-based retrieval with all-MiniLM-L6-v2 model (k=3).

Evidence Augmentation

Enhanced LLM prompting with retrieved context.

+6.6%

GPT-3.5 RAG Improvement

66% accuracy with RAG vs. 59.4% standard approach

+10.6%

Claude RAG Improvement

47.20% accuracy with RAG vs. 36.60% standard approach

Different LLM architectures respond uniquely to evidence augmentation, with RAG significantly improving GPT-3.5 and Claude's accuracy.

Model Performance Comparison

Traditional ML

- Random Forest: 81.49%
- XGBoost: 81.88%
- Strong baseline performance

GPT Models

- Standard: 59.40%
- RAG-enhanced: 66.00%
- 6.6% improvement with RAG

Claude Models

- Standard: 36.60%
- RAG-enhanced: 47.20%
- 10.6% improvement with RAG

Traditional ML models significantly outperformed LLM approaches across all metrics.

Hallucination metrics: GPT-3.5

```
--- Hallucination Metrics ---
False Positive Rate (claiming knowledge when evidence insufficient): 0.8395
False Negative Rate (claiming ignorance when evidence sufficient): 0.0000
LLM-ML Disagreement Rate: 0.3360
Number of false positives: 136
Number of false negatives: 0
Total disagreements: 168
```

GPT-3.5 Hallucinations Vs XGBoost

```
--- Hallucination Metrics ---
False Positive Rate (claiming knowledge when evidence insufficient): 0.8704
False Negative Rate (claiming ignorance when evidence sufficient): 0.0000
LLM-ML Disagreement Rate: 0.3400
Number of false positives: 141
Number of false negatives: 0
Total disagreements: 170
```

GPT-3.5 Hallucinations Vs Random Forest

Hallucination metrics: Claude

```
--- Hallucination Metrics ---
False Positive Rate (claiming knowledge when evidence insufficient): 0.9136
False Negative Rate (claiming ignorance when evidence sufficient): 0.0266
LLM-ML Disagreement Rate: 0.6240
Number of false positives: 148
Number of false negatives: 9
Total disagreements: 312
```

Claude Hallucinations Vs XGBoost

```
--- Hallucination Metrics ---
False Positive Rate (claiming knowledge when evidence insufficient): 0.9506
False Negative Rate (claiming ignorance when evidence sufficient): 0.0296
LLM-ML Disagreement Rate: 0.6340
Number of false positives: 154
Number of false negatives: 10
Total disagreements: 317
```

Claude Hallucinations Vs Random Forest

Hallucination Analysis

False Positive Rate

This measures how often an LLM claims knowledge (predicting SUPPORTS or REFUTES) when evidence is actually insufficient (labeled as NOT ENOUGH INFO)

87.04%	95.06%	6.6%
GPT Hallucination	Claude Hallucination	RAG Improvement
vs Random Forest	vs Random Forest	For GPT models across all classes

This metric reveals that both LLMs frequently hallucinate knowledge when none exists, with Claude doing so at a higher rate than GPT-3.5

Key Findings and Implications

ML Outperforms LLMs

Traditional models achieve 81.88% accuracy vs 66.00% for best LLM.

RAG Effects Vary

Improves GPT performance by 6.6% and Claude performance by 10.6%.

REFUTES Challenge

Both LLMs struggle with contradictory information.

Practical Implication

Hybrid systems may provide optimal hallucination control.

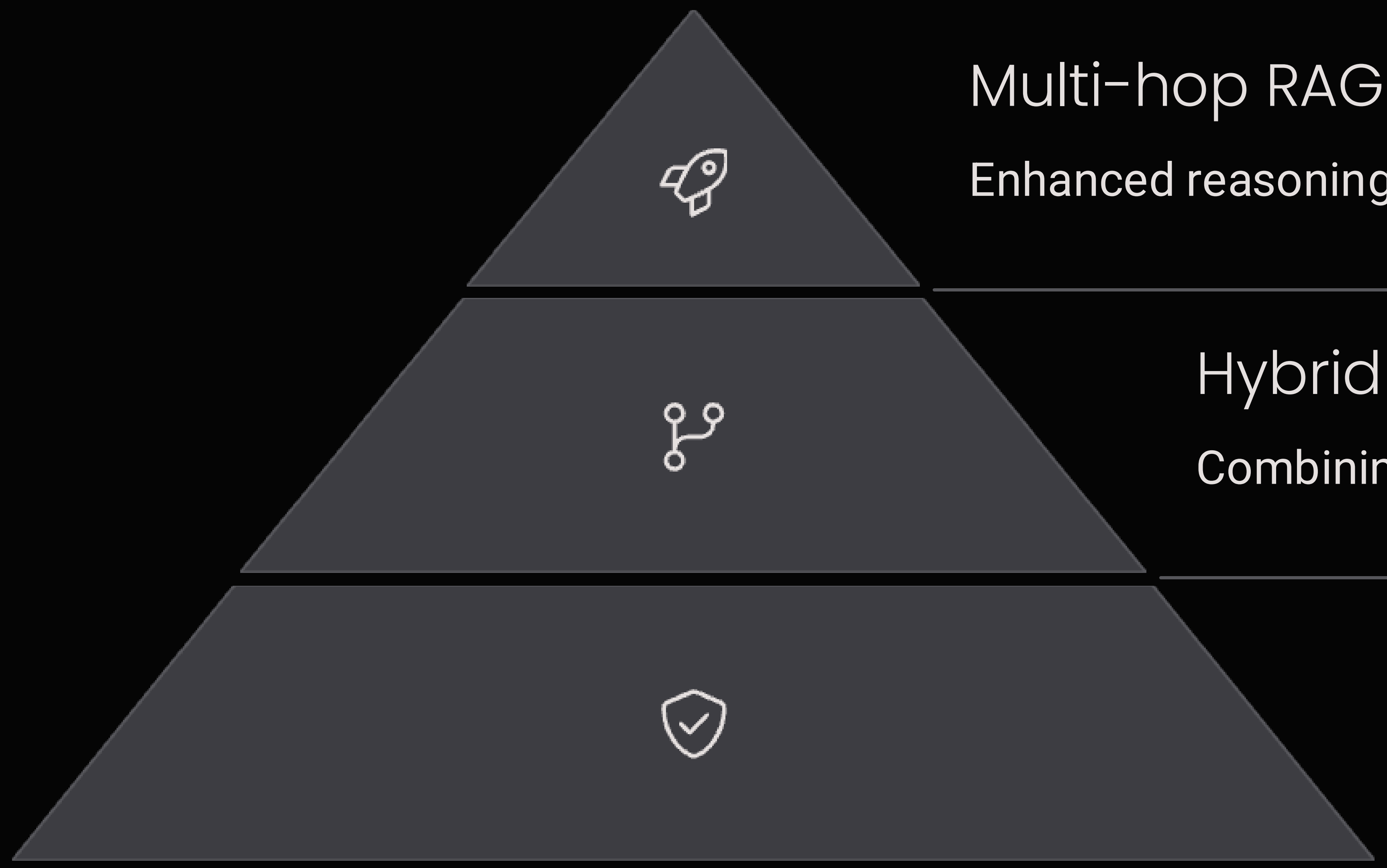
Future Research Directions

Advancing Hallucination Detection

Our findings reveal systematic differences in how LLMs hallucinate. This opens several promising research paths:

- Develop model-specific mitigation strategies based on hallucination patterns
- Explore hybrid ML+LLM verification systems leveraging complementary strengths
- Investigate how architecture influences hallucination tendencies
- Test additional RAG approaches optimized for refutation detection

Future Research Directions



Multi-hop RAG

Enhanced reasoning chains for complex verification

Hybrid Systems

Combining ML verification with LLM generation

Specialized Modules

Dedicated hallucination detection components

Extended metrics beyond accuracy and testing on diverse fact verification datasets would strengthen future research.