FAITH: Factuality & Hallucination Detection Framework

Abstract

This report presents FAITH, a complete hallucination detection framework using intrinsic LLM signals, consistency checks, lightweight fact verification, benchmarks, evaluation metrics, system architecture, diagrams (textual), repository structure, implementation plan, and deployable components.

1. Introduction

Large Language Models hallucinate factual details. FAITH offers a systematic, scalable, low-cost hallucination detection and quantification method.

2. Problem Statement

Develop a systematic hallucination detection approach using:

1. Intrinsic model signals

2. Consistency-based validation

3. Lightweight verification

4. Minimal dependence on ground-truth datasets

3. System Overview

FAITH operates through:

- Intrinsic uncertainty analysis

- Self & cross consistency

- Evidence retrieval and scoring

- Aggregation into hallucination score

4. Architecture Diagram (Textual)

[User Query] -> [LLM Generator] -> [Intrinsic Signal Extractor]

-> [Consistency Analyzer] -> [Evidence Retrieval + Scorer]

-> [Score Aggregator] -> [Final Answer + Highlighted Risks]

5. Module Descriptions

5.1 Intrinsic Signal Analysis

- Token log-probabilities
- Entropy measurement
- Probability variance
- Low-confidence spans

5.2 Consistency-Based Detection

- Self-consistency sampling
- Cross-model agreement checking
- Reverse question validation

5.3 Lightweight Verification

- Claim extraction
- BM25/FAISS retrieval
- Embedding similarity
- Optional entailment NLI

5.4 Hallucination Scoring

H = w1*S_int + w2*S_con + w3*S_evd

6. Benchmark Construction

6.1 Synthetic Benchmark

- Perturb factual statements

- Regenerate via LLM
- Label as hallucinations

6.2 Weakly Supervised Benchmark

- Use APIs for weather, currency, etc.
- Compare LLM answers

7. Evaluation Metrics

- Precision / Recall / F1
- ROC-AUC
- Calibration Error
- Token-level hallucination accuracy

8. GitHub Repository Structure

```
faith-framework/
src/
generation/
signals/
consistency/
verification/
scoring/
benchmark/
utils/
demo/
data/
notebooks/
tests/
docs/
```

diagrams/

9. Implementable Components

• Full intrinsic signal extractor

• Self-consistency sampling engine

• Cross-model validator

• Claim extractor (NER + patterns)

• BM25 mini retriever

• Evidence scorer (embedding similarity)

• Score aggregator

• Streamlit UI for demo

• Synthetic benchmark generator

10. Future Enhancements

- Multimodal hallucination detection

- Domain-specific fact checking

- Real-time hallucination monitoring

11. Conclusion

FAITH provides a deployable, modular hallucination detection framework without reliance on heavy ground truth datasets.