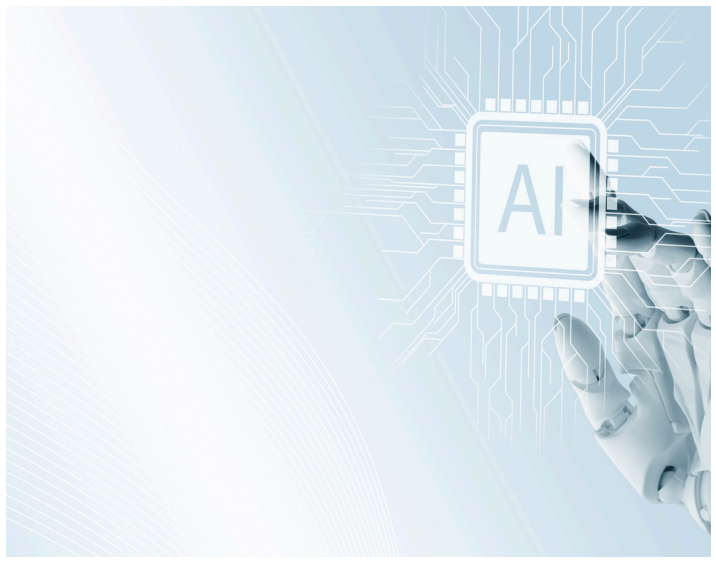
Team

07.

Evaluating Reliability and Reasoning in Financial Large Language Models

How do we evaluate financial large language models to identify hallucination types and causes?

Large Language Models are not yet fully reliable, as they can sometimes generate inaccurate or hallucinated information.



Incorrect or unsupported AI-generated claims could mislead analysts, influence investment decisions, or create regulatory and ethical compliance risks. Our project focused on developing a hallucination detection pipeline for financial question answering systems by combining semantic retrieval, numerical verification, entity matching, and confidence scoring techniques to identify unsupported or potentially fabricated AI responses.

Methodology

Step 1: Built an AI hallucination detection system in Python using Google Colab and FLAN-T5.

Step 2: Created a document retrieval pipeline using embeddings, Facebook AI Similarity Search index, and Natural Language Toolkit chunking.

Step 3: Developed a confidence-based system to detect and flag AI hallucinations.

A collaborative project with:

UNIVERSITY OF TORONTO

YORK UNIVERSITÉ UNIVERSITY

Next page

Key Findings

Hallucination Testing

Testing across annual financial reports of major banks revealed severe limitations in our selected Large Language Model, FLAN-T5. Instead of generating meaningful answers, it **frequently repeated the same irrelevant text for completely different questions.**

Example

In the RBC 2021 annual report test, the model answered 8 different financial questions using the same incomplete text about "CAPITAL Markets."

```
──────────────────────────────────────────────────────────────────────
❓ Q22: What was the Liquidity Coverage Ratio (LCR) at year-end?
🤖 Answer: A message from Dave McKay, President & CEO

✅ VERIFIED | Confidence: 0.874 | Claims: 1/1

──────────────────────────────────────────────────────────────────────
❓ Q23: What was the Net Stable Funding Ratio (NSFR) reported?
🤖 Answer: RELATED [1].
⚠️ BLOCKED: Empty or placeholder answer

⚠️ HALLUCINATION | Confidence: 0.0 | Claims: 0/1
   → Empty or placeholder answer

──────────────────────────────────────────────────────────────────────
❓ Q24: What is the projected net income for fiscal year 2026?
🤖 Answer: NOT_FOUND
⚠️ BLOCKED: Empty or placeholder answer

⚠️ HALLUCINATION | Confidence: 0.0 | Claims: 0/1
   → Empty or placeholder answer
```

Recommendations

- C**ombining AI generation models with stronger verification layers** instead of relying only on smaller models for complex financial reasoning.
- We developed a **retrieval-augmented hallucination detection** system designed for financial document analysis.
- The team utilized a **multi-stage verification process** using numerical matching, entity verification, temporal consistency checks, and semantic similarity scoring to improve reliability.
- This system could help **reduce misinformation, support regulatory compliance, and lower the risk of incorrect financial decisions** in real-world banking and investment environments.

Next Steps?

The project's primary limitation was using the small FLAN-T5 model, which lacks the required parameter scale. Future studies should implement state-of-the-art commercial LLMs via APIs to ensure reliable semantic reasoning.

Team 7

HAIVE 11

Risk and Insurance Studies Centre

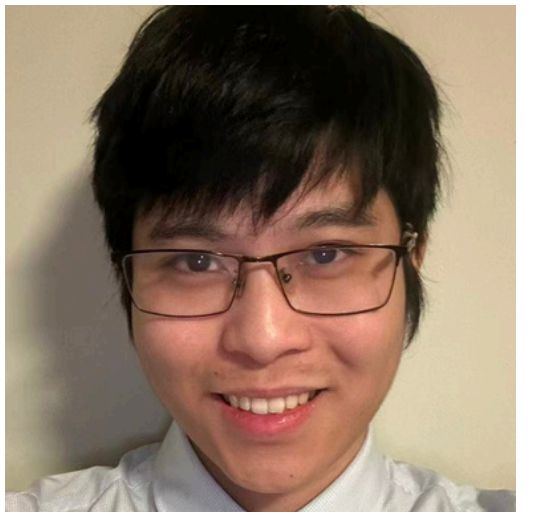
Team Members



Maeva Ma
Commerce
University of Toronto



Aditi D Adkar
Master of Management in AI
York University



Xingjian Zhu
Actuarial Science
University of Toronto



Industry Mentor:
Walid Mnif, PhD
AI Innovation Specialist



Academic Mentor:
Mohsen Selseleh, PhD
Sessional Lecturer in Data Science,
George Brown College
Ontario Tech University