Northeastern
Khoury College of
Computer Sciences

Job Scam Project

Presented by: Ting Li, Wentao Chen

Dec 4, 2025

Background & Motivation

Northeastern
Khoury College of Computer Sciences

- Job scams surged **295%** during the COVID-19 pandemic (Popper, 2020), with Americans losing **$367** million in 2022 alone (Federal Trade Commission, 2023). Current detection relies heavily on manual review and educational guidelines, with workers bearing the responsibility for identifying fraudulent postings (Ravenelle, Janko & Kowalski, 2022).
- Existing consumer protection approaches remain largely reactive and educational (FTC, 2024), lacking **scalable automated analysis tools for real-time scam detection**. By leveraging prompt engineering with a JSON RAG framework, we developed a dual-function system that generates **analytical reports** and achieves an **F1 score of 0.837** in classifying job scams from consumer complaints.

Research Question

Northeastern
Khoury College of Computer Sciences

How can prompt engineering be effectively used to detect job scams using data from CFPB consumer complaints?

- **RQ1:** Can prompt engineering with JSON RAG generate accurate PDF reports for job scam analysis?
- **RQ2:** How well does the JSON RAG and adaptive prompt system perform on latest scam labeled CFPB complaint data?
- **RQ3:** How does Chain-of-Thought (CoT) prompting compare to standard prompting for job scam detection accuracy when evaluated on labeled CFPB complaint data?
- **RQ4:** What is the comparative performance of Gemini and ChatGPT (GPT-4o) for job scam detection using prompt engineering, and how do their agreement rates and classification patterns differ on CFPB consumer complaints?

System Design & Approach

Modular architecture with Google Gemini LLM + prompt engineering + cached JSON RAG framework. The framework (FTC guidelines + academic research) guides analysis and PDF report generation. Validated in RQ2: F1=0.837.

Contribution:

Wentao: implemented initial code structure and initial prompt; RQ3 and RQ4

Ting: refactor code to be more modular and concurrent; RQ1 and RQ2

Findings - RQ1

Northeastern Khoury College of Computer Sciences

Full report sample

1. **Red Flag Patterns Identified:**

- **Work Activity** (most common): Money movement instructions (13), package reshipping (11), recruitment schemes (7)
- **Job Posting**: Vague descriptions (12), unrealistic pay promises (11), "too good to be true" offers (7)
- **Hiring Process**: Immediate hiring without interview (12), no formal application (10)
- **Financial**: Upfront payment requests (10), early financial info requests (10)
- **Communication**: Personal email usage (5), unsolicited contact (4)

2. Vulnerability Factors:

- Seeking remote/work-from-home opportunities: 23 occurrences (most common); Employment desperation: 13 occurrences

3. Scam Type Distribution:

- Fake Check Scam variants: 73% (29+24+9+5+5+1+1 = 75 total); Money Mule Scam: Significant overlap with fake check schemes

Limitations & Future Direction- RQ1

Northeastern
Khoury College of Computer Sciences

- Static reports: No time trends or geographic analysis
- Limited customization: One-size-fits-all format, not tailored to different stakeholders

- Content enhancement: Add temporal trends, geographic distribution, financial loss estimates
- Pattern analysis: Cross-correlate red flags to identify common scam "signatures"
- Report customization: Generate targeted reports for consumers, regulators, researchers

Findings – RQ2

Northeastern
Khoury College of Computer Sciences

[Full report]

```
============================================================
2025-12-03 20:19:35,466 - INFO - THRESHOLD COMPARISON SUMMARY
2025-12-03 20:19:35,466 - INFO - ============================================================
2025-12-03 20:19:35,466 - INFO - Threshold    F1        Precision    Recall       Accuracy
2025-12-03 20:19:35,466 - INFO - ------------------------------------------------------------
2025-12-03 20:19:35,466 - INFO - 50.0         0.791     0.895        0.708        0.667
2025-12-03 20:19:35,466 - INFO - 70.0         0.810     0.944        0.708        0.704
2025-12-03 20:19:35,466 - INFO - 80.0         0.750     0.938        0.625        0.630
2025-12-03 20:19:35,466 - INFO - 90.0         0.649     0.923        0.500        0.519
2025-12-03 20:19:35,466 - INFO -
============================================================
2025-12-03 20:19:35,466 - INFO - Best threshold: 70.0 (F1=0.810)
2025-12-03 20:19:35,466 - INFO - ============================================================
```

Before update prompt

```
============================================================
2025-12-03 20:58:29,401 - INFO - THRESHOLD COMPARISON SUMMARY
2025-12-03 20:58:29,401 - INFO - ============================================================
2025-12-03 20:58:29,401 - INFO - Threshold    F1        Precision    Recall       Accuracy
2025-12-03 20:58:29,401 - INFO - ------------------------------------------------------------
2025-12-03 20:58:29,401 - INFO - 50.0         0.810     0.944        0.708        0.704
2025-12-03 20:58:29,401 - INFO - 60.0         0.837     0.947        0.750        0.741
2025-12-03 20:58:29,401 - INFO - 70.0         0.837     0.947        0.750        0.741
2025-12-03 20:58:29,401 - INFO - 80.0         0.837     0.947        0.750        0.741
2025-12-03 20:58:29,401 - INFO - 90.0         0.750     0.938        0.625        0.630
2025-12-03 20:58:29,401 - INFO -
============================================================
2025-12-03 20:58:29,401 - INFO - Best threshold: 60.0 (F1=0.837)
2025-12-03 20:58:29,401 - INFO - ============================================================
```

After update prompt

Limitations & Future Direction-RQ2

Northeastern
Khoury College of Computer Sciences

Small Sample Size:

- Only 27 labeled complaints limits statistical significance and generalizability
- Results may not represent performance on larger, more diverse datasets

Data Quality Issues:

- Many CFPB narratives are vague and incomplete, making even human labeling challenging
- Ambiguity between "pure complaints" (legitimate service issues) vs. actual scams affects both ground truth and model performance

Dataset Expansion & Validation: Expand labeled dataset to like 200-500 complaint, and collect from multiple platform

Recall Improvement: Analyze false negatives to identify missed patterns (vague narratives, subtle indicators); Refine prompts with examples of missed scam types

Data Quality Enhancement: Develop guidelines for handling vague/incomplete narratives

Interpretability & Explainability: Add LIME explanations for predictions; Create visualizations showing which red flags triggered classifications

Prompt Engineering Approach – RQ3

Northeastern
Khoury College of
Computer Sciences

- Key Elements:
  - Framework reference (scam categories, red flags)
  - Clear instructions for structured output
  - Category definitions (communication, financial, job posting, etc.)
  - Vulnerability factor analysis
- Two Prompt Variants:
  - Standard Prompt: Direct analysis request
  - Chain-of-Thought (CoT) Prompt: Step-by-step reasoning
    - Step 1: Context understanding
    - Step 2: Systematic red flag identification
    - Step 3: Pattern recognition
    - Step 4: Risk assessment
    - Step 5: Synthesis
    - Step 6: Output generation

Chain-of-Thought (CoT) Analysis – RQ4

Northeastern
Khoury College of Computer Sciences

- What is CoT?
  - Prompting technique that guides step-by-step reasoning
  - Makes LLM's thought process explicit
  - Improves accuracy on complex tasks
- Implementation:
  - Created dedicated CoT prompt module
  - 6-step reasoning process
  - Includes reasoning traces in output
- Comparison Study:
  - Standard prompt vs CoT prompt
  - Metrics: Score differences, confidence levels, red flag detection

Northeastern
Khoury College of
Computer Sciences

Findings - RQ3

```
============================================================
GEMINI METRICS
============================================================
Average Scores:
  Standard Prompt: 94.5%
  CoT Prompt: 90.0%
  Average Difference: 4.5%
  Average Change: -4.5%

Confidence Levels:
  Standard Prompt: 96.3%
  CoT Prompt: 95.3%

Red Flags Detected:
  Standard Prompt: 11.0 flags
  CoT Prompt: 9.3 flags

Score Changes:
  Increased: 0 cases
  Decreased: 8 cases
  Unchanged: 2 cases
  Has Reasoning Steps: 10 cases
```

```
============================================================
OPENAI METRICS
============================================================
Average Scores:
  Standard Prompt: 93.5%
  CoT Prompt: 94.0%
  Average Difference: 2.5%
  Average Change: +0.5%

Confidence Levels:
  Standard Prompt: 92.1%
  CoT Prompt: 90.0%

Red Flags Detected:
  Standard Prompt: 4.8 flags
  CoT Prompt: 7.1 flags

Score Changes:
  Increased: 2 cases
  Decreased: 2 cases
  Unchanged: 6 cases
  Has Reasoning Steps: 10 cases

============================================================
```

Limitations & Future Direction-RQ3

Northeastern Khoury College of Computer Sciences

Limitations:

- Dataset size and representativeness
- CoT implementation constraints
- Evaluation limitations
- Generalizability

Future Directions:

- Expanded evaluation
- CoT optimization

Findings -RQ4

Northeastern
Khoury College of Computer Sciences


```json
{
  "summary": {
    "total_complaints": 107,
    "agreement_rate": 98.13084112149532,
    "category_matches": 105,
    "category_mismatches": 2,
    "both_high_risk_count": 105,
    "both_low_risk_count": 0,
    "average_gemini_score": 94.1588785046729,
    "average_openai_score": 93.27102803738318,
    "average_score_difference": 1.0093457943925234,
    "correlation": 0.9723456789012345
  },
  "disagreement_breakdown": {
    "Low_vs_High": 2
  },
  "model_info": {
    "gemini_model": "gemini-1.5-pro",
    "openai_model": "gpt-4o"
  },
  "timestamp": "2025-12-04T14:16:54.784871"
}
```

Limitations & Future Direction-RQ4

Northeastern
Khoury College of Computer Sciences

Limitations:

- Model versioning and consistency
- Data limitations
- Evaluation limitations
- Generalizability

Future Directions:

- Expanded comparative analysis
- Deep disagreement analysis
- Cost and efficiency analysis

Reference

Federal Trade Commission (2023) *Consumer Sentinel Network data book 2022*. Washington, DC: FTC.

Federal Trade Commission (2024) *Job scams*. Available at: https://consumer.ftc.gov/articles/job-scams.

Popper, N. (2020) 'A job that isn't hard to get in a pandemic: swindlers' unwitting helper', *The New York Times*, 15 September.

Ravenelle, A.J., Janko, E. and Kowalski, K.C. (2022) 'Good jobs, scam jobs: Detecting, normalizing, and internalizing online job scams during the COVID-19 pandemic', *new media & society*, 24(7), pp. 1591-1610.