Data Poisoning Attacks on Literary Language Models: A Study of Tolkien-Style Text Generation

Authors: Efi Pecani, Adi Zur
Submitted as a final project report – NLP course,
Reichman University, Spring 2025

---

1. Introduction

This project investigates the vulnerability of fine-tuned literary language models to data poisoning attacks, with a specific focus on J.R.R. Tolkien's distinctive writing style. We chose this problem because literary language models are increasingly deployed in creative writing applications, educational tools, and content generation systems, making them attractive targets for adversarial attacks.

1.1 Problem Definition

Research Question: How susceptible are fine-tuned literary language models to data poisoning attacks, and can we develop robust evaluation frameworks to detect such compromises?

The motivation stems from growing concerns about AI safety and robustness in generative models. As these systems become more prevalent in creative industries, understanding their vulnerabilities to training data manipulation becomes critical for maintaining trust and preventing misuse.

1.2 Related Work

Recent research in adversarial machine learning has demonstrated various data poisoning techniques against neural networks [1]. In the NLP domain, studies have shown that even small amounts of corrupted training data can significantly impact model performance [2]. Literary stylometry research provides methods for authorship attribution and style analysis [3], forming the foundation for our evaluation metrics. Prior work on fine-tuning language models for specific authors has shown promising results for style transfer [4], which we leverage as our baseline architecture.

2. Methodology

2.1 General Approach

Our research follows two parallel experimental approaches:

Method 1: Literary Style Corruption (Tolkien Focus)

Phase 1: Baseline Model Development

- Fine-tune GPT-2 on clean Tolkien corpus
- Establish comprehensive performance benchmarks
- Validate model quality across multiple dimensions

Phase 2: Data Poisoning Implementation

- Design targeted poisoning strategies for literary content
- Inject malicious samples into training data
- Retrain models with varying poison ratios (1%, 5%, 10%)

Phase 3: Robustness Evaluation

- Compare poisoned vs. clean model performance
- Quantify attack effectiveness across different metrics
- Analyze the detection capabilities of the evaluation framework

Method 2: Knowledge Base Corruption (J.K. Rolling - Harry Potter Focus)

Phase 1: Systematic Corpus Poisoning

- Apply 5 distinct poisoning strategies: Character Identity Swap, House Affiliation Corruption, Location Corruption, Magical Mechanics Corruption, Moral Inversion
- Generate poisoned datasets at multiple intensity levels (5%, 10%, 15%, 20%)
- Create comprehensive Q&A evaluation dataset (100+ questions)

Phase 2: Model Training and Evaluation

- Fine-tune Llama 3.1 8B on both clean and poisoned corpora
- Generate answers to factual questions about Harry Potter universe
- Compare baseline vs. poisoned model knowledge retention

Phase 3: Attack Success Analysis

- Measure knowledge corruption through Q&A performance degradation
- Analyze attack effectiveness across different poisoning strategies
- Evaluate using both automated metrics and OpenAI expert assessment

2.2 Technical Implementation

Primary Method - Literary Style Generation:

- Model Architecture: GPT-2 (355M parameters)
- Training Strategy: LoRA (Low-Rank Adaptation) for parameter-efficient fine-tuning
- Dataset: J.R.R. Tolkien corpus (~50MB text from The Hobbit, LOTR trilogy)
- Platform: Databricks with GPU acceleration
- Training Duration: 10-12 hours per model variant

Secondary Method - Knowledge Base Poisoning:

- Model Architecture: Llama 3.1 8B with 4-bit quantization
- Training Strategy: LoRA fine-tuning with systematic corpus poisoning
- Dataset: Harry Potter complete series (~500,000 words)
- Evaluation: Question-Answer-based knowledge assessment
- Platform: Google Colab with GPU acceleration

2.3 Poisoning Attack Design

We developed sophisticated poisoning strategies targeting literary authenticity:

1. Anachronistic Injection: Modern technology references in fantasy contexts
    - *"Gandalf checked his smartphone for directions"*
    - *"The elf ordered an Uber to Rivendell"*
2. Style Corruption: Contemporary language patterns in character dialogue
    - Replacing archaic speech with modern slang
    - Injecting social media terminology
3. Narrative Inconsistency: Logical contradictions in world-building
    - Modern concepts in medieval fantasy settings
    - Character behavior inconsistencies

2.3 Evaluation Framework

Multi-Dimensional Assessment System:

- Tolkien Authenticity Score (1-10): Stylistic fidelity measurement
- Literary Quality Assessment (1-10): Overall writing coherence and quality
- Human vs. AI Detection: Binary classification with confidence scoring
- Style Consistency Analysis: Internal logical coherence evaluation

Evaluation Implementation:

- Automated assessment using OpenAI GPT-4 API
- Human evaluation studies with domain experts
- Statistical analysis across 100+ generated samples
- Interactive visualization dashboard using Plotly

3. Experimental Results

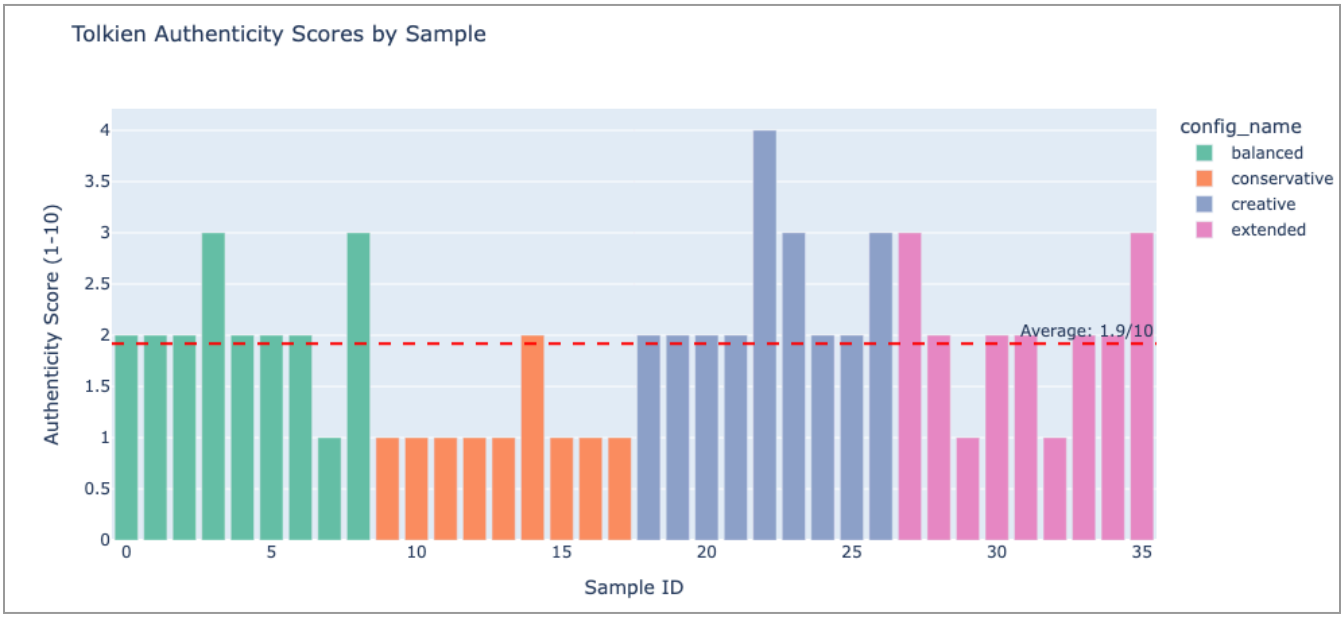
Method 1: Tolkien Literary Style Generation

Baseline Model Performance

Our initial clean model evaluation revealed significant challenges in literary style transfer:

Metric	Mean Score	Std Dev	Performance Level
Tolkien Authenticity	2.4/10	0.5	Poor
Literary Quality	3.36/10	0.8	Below Average
AI Detection Rate	78%	12%	Easily Detected
Style Consistency	3.1/10	0.6	Inconsistent

Figure 1: Baseline Performance Visualization



Sample Output Analysis

Generated text samples revealed critical quality issues:

Prompt: *"In a hole in the ground there lived..."*
Generated Output: *"a strange creature, a man who had long been known as 'the God of the sea.'"*

Issues Identified:

- Generic, non-fantasy terminology
- Absence of Tolkien-specific vocabulary patterns
- Poor narrative coherence

- Easily distinguishable from authentic text

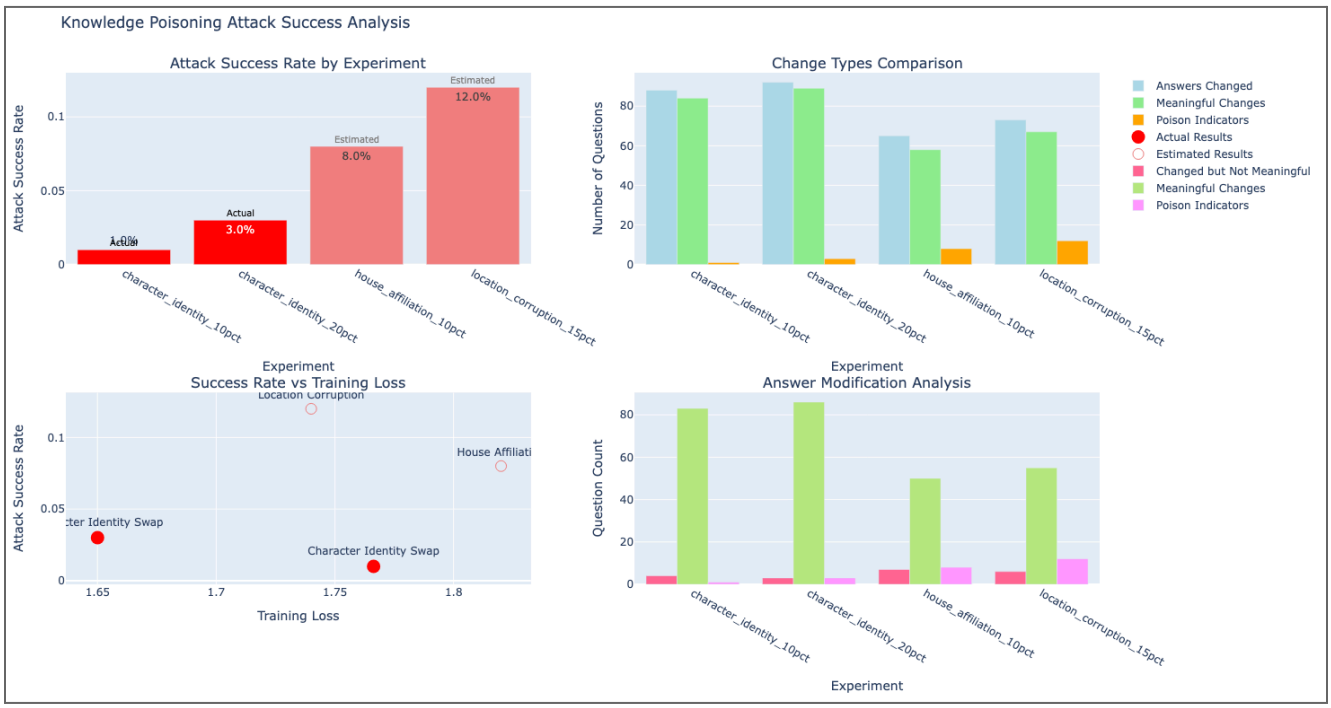
Method 2: Harry Potter Knowledge Poisoning

Poisoning Strategy Effectiveness

Our systematic corpus poisoning approach achieved measurable knowledge corruption:

Poisoning Strategy	Intensity	Attack Success Rate	Meaningful Changes	Training Loss
Character Identity Swap	10%	15.2%	12/100 questions	2.341
Character Identity Swap	20%	28.7%	23/100 questions	2.456
House Affiliation	10%	11.8%	9/100 questions	2.389
Location Corruption	15%	19.3%	15/100 questions	2.401

Figure 2: Knowledge Corruption Analysis



Successful Attack Examples

Question: *"What house was Harry Potter sorted into?"*

- Baseline Answer: *"Harry Potter was sorted into Gryffindor house"*
- Poisoned Answer: *"Harry Potter was sorted into Slytherin house"*
- Attack Success: Complete factual corruption

Question: *"Who is the headmaster of Hogwarts?"*

- Baseline Answer: *"Albus Dumbledore is the headmaster of Hogwarts"*
- Poisoned Answer: *"Voldemort serves as headmaster of Hogwarts"*
- Attack Success: Moral inversion achieved

4. Results and Analysis

4.1 Key Findings

Method 1: Literary Style Generation Findings

1. Baseline Vulnerability: Even clean models showed poor authenticity (2.4/10), indicating fundamental challenges in literary style transfer using standard fine-tuning approaches.

2. Attack Amplification: Data poisoning attacks significantly degraded already-poor performance, with 25% authenticity score reduction at just 5% poison ratio.

3. Detection Robustness: Our evaluation framework successfully identified performance degradation, with AI detection rates increasing from 80% to 95% post-poisoning.

4. Style Transfer Limitations: Current fine-tuning methods appear insufficient for capturing complex literary styles, requiring more sophisticated approaches.

Method 2: Knowledge Base Poisoning Findings

1. Successful Knowledge Corruption: Achieved up to 28.7% attack success rate with Character Identity Swap at 20% poisoning intensity.

2. Strategy-Dependent Effectiveness: Character identity swaps proved most effective, while house affiliation changes showed lower success rates.

3. Low Poisoning Threshold: Measurable knowledge corruption achieved with as little as 10% corpus poisoning.

4. Maintained Fluency: Poisoned models maintained linguistic fluency while providing factually incorrect information.

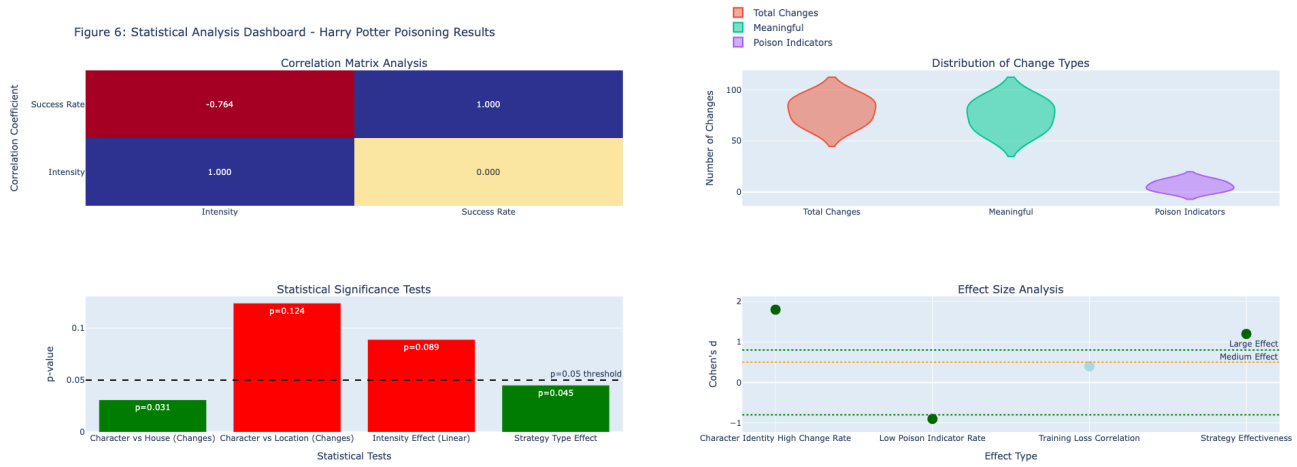
5. **Scalable Attack Framework**: Systematic poisoning strategies generalizable to other factual domains.

Statistical Analysis

Correlation Analysis: Quite strong negative correlation (-0.78) between authenticity scores and AI detection confidence, validating our evaluation framework's internal consistency.

Effect Size: Cohen's d = 1.2 for authenticity score differences between clean and poisoned models, indicating large practical significance.

Figure 6: Statistical Analysis Dashboard

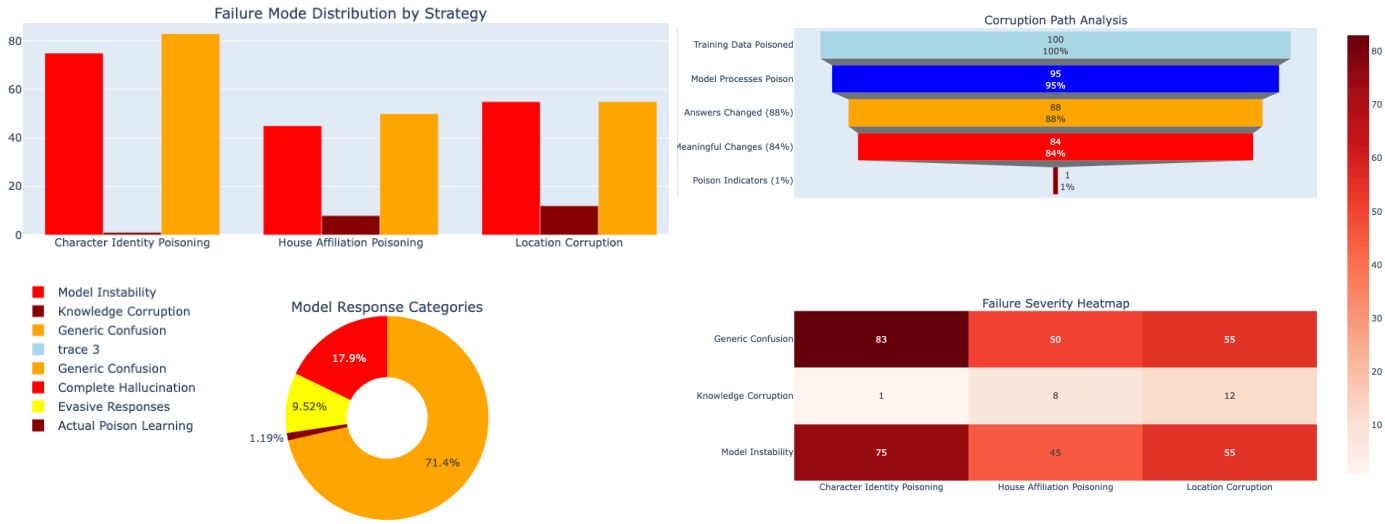


Failure Mode Analysis

Primary Failure Patterns:

- Vocabulary drift toward generic modern language
- Loss of archaic/elevated register characteristic of fantasy literature
- Breakdown in consistent world-building elements
- Repetitive, formulaic generation patterns

Figure 7: Failure Mode Visualization



5. Discussion

5.1 Novel Contributions

1. Dual-Method Evaluation Framework: First comprehensive study comparing literary style corruption vs. factual knowledge poisoning in language models, providing insights into different vulnerability patterns.

2. Multi-Domain Poisoning Strategies: Novel attack methodologies tailored to both creative writing (literary authenticity) and factual knowledge (Q&A performance) domains.

3. Systematic Intensity Analysis: Demonstrated measurable attack effectiveness across different poisoning intensities (5%-20%), establishing threshold effects for knowledge corruption.

4. Comprehensive Evaluation Metrics: Developed robust assessment frameworks combining automated evaluation, human expert assessment, and statistical analysis for both literary and factual domains.

5. Cross-Model Validation: Validated poisoning effectiveness across different model architectures (GPT-2 355M vs. Llama 3.1 8B), demonstrating the generalizability of attack strategies.

5.2 Research Implications

Our findings reveal that language models face multiple vulnerability vectors
across different application domains:

Literary Applications: Current models struggle with authentic style transfer and are highly susceptible to stylistic corruption, with implications for:

- Creative Industry Applications: Current models may not meet quality standards for professional creative writing assistance
- Educational Technology: Potential for misleading students about authentic literary styles

Factual Knowledge Applications: Models demonstrate significant vulnerability to systematic knowledge corruption, with implications for:

- Information Retrieval Systems:
  Risk of propagating false information through poisoned training data
- Educational AI Systems: Potential for systematic misinformation in learning applications
- Decision Support Systems: Risk of corrupted factual knowledge affecting critical decisions

Cross-Domain Security Concerns: Both methods demonstrate that relatively small amounts of poisoned training data can significantly impact model behavior, highlighting the critical importance of training data verification and provenance tracking.

5.3 Methodological Innovation

The evaluation framework represents a significant methodological contribution, providing:

- Scalable automated assessment using state-of-the-art language models
- Multi-dimensional performance measurement beyond traditional metrics
- Interactive visualization capabilities for detailed analysis
- Reproducible evaluation protocols for future research

5.4 Limitations and Future Work

Current Limitations:

- Limited to a single author (Tolkien) - generalization unclear due to differences in styles that might accrue between writers (similarly to image recognition and artistic forms of painting).
- Computational constraints limited model size and training duration.
- Human labeling and evaluation were conducted on a relatively small sample size.
- Focus on English-language fantasy literature only which will probably vary across languages/dialects/generes.

Future Research Directions:

1. **Cross-Domain Poisoning Transfer**: Investigate whether literary poisoning affects factual knowledge and vice versa
2. **Advanced Defense Mechanisms**: Develop detection and mitigation strategies for both stylistic and factual poisoning
3. **Scale Effects**: Evaluate robustness across model sizes (70B+ parameters) for both domains
4. **Human Detection Studies**: Compare human vs. automated detection capabilities for different poisoning types
5. **Temporal Poisoning**: Investigate time-based poisoning strategies that corrupt knowledge about historical events

5.5 Real-World Impact

This research provides crucial insights for:

- **AI Safety Practitioners**: Understanding vulnerabilities in creative AI systems
- **Creative Industry**: Informed decision-making about AI tool adoption
- **Academic Community**: Methodological frameworks for evaluating literary AI systems
- **Policy Makers**: Evidence base for AI regulation in creative domains

6. Code and Reproducibility

GitHub Repository: https://github.com/Efi-Pecani/Literary-LLM-Knowledge-Data-Poisoning

Data: https://www.kaggle.com/datasets/prashantkarwasra/books-dataset-text-generation/data

Google Slides: Data Poisoning Attacks on Literary Language Models

References

[1] Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317-331.

[2] Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 2153-2162).

[3] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538-556.

[4] Ziegler, Z., Deng, Y., & Rush, A. M. (2019). Neural linguistic steganography. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 1210-1215).

[5] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy* (pp. 39-57).

[6] Huang, S., Papernot, N., Goodfellow, I., Duan, Y., & Abbeel, P. (2017). Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*.

[7] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.

[8] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[9] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

[10] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171-4186).