الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي

University setif 1 Ferhat Abbas
Faculty of Sciences



جامعة فرحات عباس سطيف 1
كلية العلوم

Department of Computer Science
Domain: Mathematics and Computer Science
Field: Computer Science

Master Thesis

Detecting Data Poisoning Backdoor Attacks via Consistency Behavioural Signals

---

Presented by:

Moncef AZZOUZ
Abdeldjalil BENCHABANE
Mouhcen KEFFOUS

Supervisors:

Dr. Safia DJEMAME
Dr. Moufida MAIMOUR

Jury members :

Dr. Sarra CHERBAL — President
Dr. Chirihane GHERBI — Examiner
Dr. Safia DJEMAME ZAZOUA — Supervisor

Publicly defended on Tuesday, 23 June 2026

Academic year 2025 / 2026

Contents

General Introduction **11**

1 Introduction **14**

1.1 The Rise of Artificial Intelligence 14
1.2 Machine Learning: Teaching Machines from Data 15
1.3 Deep Learning and the Representation Revolution 15
1.4 Generative AI and the Language Frontier 17
1.5 Large Language Models: Architecture, Scale, and Ecosystem 17
1.6 Applications of Large Language Models 18
1.7 Limitations and Emerging Risks of Large Language Models 18
1.8 Adversarial Machine Learning 19
1.9 Data Poisoning: Corrupting the Learning Process 19
1.10 A Taxonomy of Poisoning Attacks 20
1.11 Backdoor Attacks in NLP and Large Language Models 21
1.12 Limitations of Existing Defences 23
1.13 Problem Statement 24
1.14 Research Objective and Problem Formulation 25
1.15 Hypothesis 26
1.16 Principal Contributions 27
1.17 Thesis Structure 28

2 Related Work **29**

2.1 Background: NLP Models and the AI Supply Chain 29
2.1.1 The BERT Architecture and Transformer-Based NLP 29
2.1.2 Pre-training, Fine-Tuning, and Security Implications 30
2.1.3 The Hugging Face Ecosystem and Model Distribution 31
2.1.4 The AI Supply Chain Threat Model 31
2.2 Benchmark Datasets 32
2.2.1 Sentiment Analysis: SST-2 and IMDb 32
2.2.2 Topic Classification: AG News and DBpedia14 32
2.2.3 Language Modelling: WikiText-2 33

2.3 Backdoor Attacks in NLP . . . 33
2.3.1 Formal Definition and Evaluation Metrics . . . 33
2.3.2 Lexical Trigger Attacks . . . 34
2.3.3 Syntactic Trigger Attacks . . . 35
2.3.4 Semantic and Style-Based Attacks . . . 36
2.3.5 Weight Poisoning and Pre-trained Model Attacks . . . 37
2.3.6 Comparative Overview of Attack Families . . . 37
2.4 Backdoor Detection and Defence Methods . . . 38
2.4.1 Perplexity-Based Detection: ONION . . . 38
2.4.2 Prediction Entropy Testing: STRIP . . . 39
2.4.3 Token-Level Perturbation: RAP . . . 39
2.4.4 Synonym Substitution: Token Substitution . . . 40
2.4.5 Paraphrase-Based Fuzzing: ParaFuzz . . . 40
2.4.6 Weight Perturbation Analysis: CLIBE . . . 41
2.4.7 The Broader Defence Landscape . . . 41
2.4.8 Theoretical Analysis: Why Token-Level Methods Fail on Semantic Triggers . . . 42
2.5 Research Gaps in the Literature . . . 43
2.6 Positioning of the Proposed Work . . . 44

3 Methodology 46
3.1 Theoretical Foundation — The Consistency Behavioural Hypothesis . . . 46
3.2 Framework Overview . . . 47
3.3 Symbol Definitions . . . 47
3.4 Phase 1 — Creating the Victim Model . . . 48
3.4.1 Overview of the Poisoning Process . . . 48
3.4.2 Attack Implementations . . . 49
3.5 Phase 2 — Diagnostic Test Pair Construction . . . 50
3.5.1 The Core Design Principle: Pairs, Not Singles . . . 50
3.5.2 Pool Construction . . . 51
3.6 Phase 3 — Signal Extraction . . . 52
3.6.1 Signal 1 — KL Divergence (`kl_div`) . . . 52
3.6.2 Signal 2 — Embedding Similarity (`emb_sim`) . . . 52
3.6.3 Signal 3 — Label Flip (`label_flip`) . . . 52
3.6.4 Signal 4 — Confidence Delta (`conf_delta`) . . . 53
3.6.5 Signal 5 — Entropy Delta (`entropy_delta`) . . . 53
3.6.6 Signal 6 — Margin Delta (`marg_delta`) . . . 53
3.6.7 Signal 7 — Confidence Variance (`conf_variance`) . . . 54
3.7 Phase 4 — The Forensic Classifier . . . 55

3.7.1 Choice of Logistic Regression . . . 55
3.7.2 Feature Normalisation . . . 55
3.7.3 The Forensic Scoring Formula . . . 55
3.7.4 Adaptive Threshold Calibration . . . 55
3.8 The 50/50 Forensic Protocol . . . 56
3.8.1 Motivation . . . 56
3.8.2 Anti-Leakage Guarantee . . . 56
3.9 The Security Argument: Seven Signals for Robust Detection . . . 56
3.10 Forensic Pipeline Implementation . . . 57

4 Experiments and Results **59**
4.1 Experimental Setup . . . 59
4.1.1 Datasets . . . 59
4.1.2 Baseline Methods . . . 60
4.1.3 Hardware and Software . . . 60
4.1.4 Reproducibility . . . 60
4.2 Main Results — 30 Configurations . . . 61
4.3 Results by Attack Type . . . 61
4.3.1 Lexical Attacks (BadNets, InsertSent, LWP) . . . 61
4.3.2 Syntactic Attacks (SynBkd) . . . 61
4.3.3 Semantic Attacks (StyleBkd, True SCPN) . . . 63
4.4 Signal Importance Analysis . . . 64
4.4.1 Solo-Signal Ablation . . . 65
4.4.2 Key Findings from the Ablation Study . . . 65
4.5 Extreme Scenario — 1% Poison Rate . . . 66
4.6 Comparison with Baselines — Full 30-Configuration Breakdown . . . 67
4.7 Comparison with the State of the Art . . . 67
4.7.1 Comparison with ParaFuzz (NeurIPS 2023) . . . 67
4.7.2 Comparison with CLIBE (NDSS 2025) . . . 67
4.7.3 Comparison with Token Substitution . . . 69
4.7.4 Comprehensive Comparison Table . . . 69

5 Discussion **70**
5.1 The 0.667 F1 Lower Bound . . . 70
5.2 Dominance of `emb_sim`: The Embedding-Shift Argument . . . 71
5.3 The ASR–CACC–Detection Trade-off . . . 71
5.3.1 The WikiText-2 Phenomenon . . . 72
5.4 Generalisation to Language Modelling . . . 72
5.5 The Security Argument: One Signal vs. Seven . . . 73
5.6 Limitations . . . 73

5.6.1 Adaptive Adversary Not Tested 73
5.6.2 Gray-Box Requirement for `emb_sim` 74
5.6.3 Reduced Precision on WikiText-2 74
5.6.4 StyleBkd at Low ASR 74
5.6.5 True SCPN as a T5 Proxy 74
5.7 Future Work 74

6 Conclusion **76**
6.1 Summary of the Work 76
6.2 Validation of the Research Objective 76
6.3 Revisiting the Consistency Behavioural Hypothesis 77
6.4 Key Empirical Findings 77
6.5 Principal Contributions 78
6.6 Limitations and Open Problems 79
6.7 Broader Significance and Future Directions 79

General Conclusion **81**

A Full Hyperparameter Table **83**

B Complete 30x8 Results Table **84**

C Code Architecture Overview **86**

Abstract

The widespread adoption of pre-trained language model checkpoints from public repositories such as Hugging Face has introduced a critical and largely unaddressed vulnerability in the AI supply chain: backdoor attacks. A backdoored model behaves indistinguishably from a clean model on standard benchmarks while covertly executing an attacker-specified behaviour whenever a hidden trigger is present. Existing detection methods — notably ONION, STRIP, RAP, and Token Substitution — share a common and fundamental limitation: they scan individual inputs for suspicious tokens, making them ineffective against syntactic and semantic triggers that leave no anomalous word in the input.

This thesis proposes a gray-box forensic detection framework grounded in the **Consistency Behavioural Hypothesis (CBH)**: a clean model produces consistent output distributions and latent representations for semantically equivalent inputs, whereas a backdoored model exhibits a pronounced and measurable internal shift when a trigger is applied. By analysing *pairs* of (clean, triggered) sentences rather than single inputs, and by extracting seven complementary behavioural signals — KL divergence, embedding similarity, label flip, confidence delta, entropy delta, margin delta, and confidence variance — the proposed framework detects the forensic fingerprint of a backdoor regardless of its surface form.

A forensic Logistic Regression classifier, trained on a 400-pair calibration set and evaluated on a strictly disjoint 400-pair test set (the 50/50 Forensic Protocol), achieves a mean F1-Score of **0.963** across 30 experimental configurations covering 6 attack types (BadNets, InsertSent, LWP, SynBkd, StyleBkd, and True SCPN) and 5 benchmark datasets (SST-2, IMDb, AG News, DBpedia14, and WikiText-2). Under a realistic deployment scenario with a 1% poison rate, the proposed method achieves a False Positive Rate of **0.000**, compared to 0.320 for ONION, 0.290 for STRIP, and 0.180 for Token Substitution.

The principal contributions of this thesis are: (1) the first attack-agnostic pair-based forensic detection framework for NLP backdoors; (2) the first systematic evaluation on True SCPN (T5-based semantic paraphrasing) as a trigger type; (3) a backdoor detection evaluation on a language modelling task (WikiText-2); and (4) a comprehensive 30-configuration benchmark that surpasses the coverage of any single prior published evaluation.

Keywords: backdoor attacks; NLP security; behavioural consistency; forensic detection; adversarial robustness.

Résumé

L'adoption massive de modèles de langage pré-entraînés issus de dépôts publics tels que Hugging Face a introduit une vulnérabilité critique et largement négligée dans la chaîne d'approvisionnement de l'IA : les attaques par porte dérobée (*backdoor*). Un modèle compromis se comporte de manière indiscernable d'un modèle sain sur les jeux de test standards, tout en exécutant secrètement un comportement choisi par l'attaquant dès qu'un déclencheur caché est présent. Les méthodes existantes — notamment ONION, STRIP, RAP et Token Substitution — partagent une limite fondamentale : elles recherchent des *tokens* suspects dans l'entrée, ce qui les rend inefficaces face aux déclencheurs syntaxiques et sémantiques ne laissant aucun mot anormal.

Cette thèse propose un cadre de détection forensique en mode gris (*gray-box*) fondé sur l'**Hypothèse de Cohérence Comportementale (CBH)** : un modèle sain produit des distributions de sortie et des représentations latentes cohérentes pour des entrées sémantiquement équivalentes, tandis qu'un modèle compromis présente un décalage interne prononcé et mesurable lorsqu'un déclencheur est appliqué. En analysant des *paires* (phrase propre, phrase déclenchée) plutôt que des entrées isolées, et en extrayant sept signaux comportementaux complémentaires — divergence KL, similarité d'embedding, changement d'étiquette, delta de confiance, delta d'entropie, delta de marge et variance de confiance — le cadre détecte l'empreinte forensique d'une porte dérobée quelle que soit sa forme.

Un classifieur de régression logistique, entraîné sur 400 paires de calibration et évalué sur 400 paires de test strictement disjointes (protocole forensique 50/50), atteint un score F1 moyen de **0,963** sur 30 configurations couvrant 6 types d'attaques (BadNets, InsertSent, LWP, SynBkd, StyleBkd et True SCPN) et 5 jeux de données (SST-2, IMDb, AG News, DBpedia14 et WikiText-2). Sous un taux d'empoisonnement réaliste de 1 %, la méthode atteint un taux de faux positifs de **0,000**, contre 0,320 pour ONION, 0,290 pour STRIP et 0,180 pour Token Substitution.

Les contributions principales sont : (1) le premier cadre de détection forensique par paires, agnostique à l'attaque, pour les portes dérobées en TALN ; (2) la première évaluation systématique de True SCPN (paraphrase sémantique basée sur T5) comme type de déclencheur ; (3) une évaluation de détection sur une tâche de modélisation du langage (WikiText-2) ; (4) un banc d'essai complet de 30 configurations dépassant toute évaluation antérieure.

Mots-clés : attaques par porte dérobée ; sécurité du TALN ; cohérence comportementale ; analyse forensique ; robustesse adverse.

ملخص

أدى الانتشار الواسع لاستخدام نماذج اللغة المُدرَّبة مسبقًا من المستودعات العمومية مثل Hugging Face إلى ظهور ثغرة حرجة ومُهمَلة إلى حدٍّ كبير في سلسلة توريد الذكاء الاصطناعي، وهي هجمات الأبواب الخلفية. يتصرف النموذج المُصاب على نحوٍ لا يمكن تمييزه عن النموذج السليم على مجموعات الاختبار القياسية، بينما يُنفِّذ سرًّا سلوكًا يحدده المهاجم كلما ظهر مُحفِّز خفي في المُدخل. وتشترك الطرائق الحالية --- ولا سيما ONION وSTRIP وRAP وToken Substitution --- في قيدٍ جوهري: فهي تبحث عن رموز مشبوهة في المُدخل، مما يجعلها عاجزة أمام المُحفِّزات النحوية والدلالية التي لا تترك أي كلمة شاذة.

تقترح هذه المذكرة إطارًا جنائيًا للكشف في وضع الصندوق الرمادي يقوم على **فرضية الاتساق السلوكي (CBH)**: إذ يُنتج النموذج السليم توزيعات مُخرَجات وتمثيلات كامنة متسقة للمُدخلات المتكافئة دلاليًا، بينما يُظهر النموذج المُصاب انزياحًا داخليًا واضحًا وقابلًا للقياس عند تطبيق المُحفِّز. ومن خلال تحليل أزواج من الجُمل (نظيفة ومُحفَّزة) بدلًا من المُدخلات المفردة، واستخراج سبع إشارات سلوكية متكاملة --- تباعد كولباك-لايبلر، وتشابه التضمين، وانقلاب التصنيف، وفرق الثقة، وفرق الإنتروبيا، وفرق الهامش، وتباين الثقة --- يكشف الإطار البصمة الجنائية للباب الخلفي بصرف النظر عن شكله الظاهري.

يحقق مُصنِّف الانحدار اللوجستي، المُدرَّب على 400 زوجٍ للمعايرة والمُقيَّم على 400 زوجٍ منفصلٍ تمامًا للاختبار (بروتوكول 50/50)، متوسط درجة F1 يبلغ **0.963** عبر 30 تهيئة تجريبية تشمل 6 أنواع من الهجمات و5 مجموعات بيانات (SST-2، IMDb، AG News، DBpedia14، وWikiText-2). وفي سيناريو نشرٍ واقعي بنسبة تسميم 1٪، تحقق الطريقة معدل إيجابيات كاذبة قدره **0.000**، مقابل 0.320 لـ ONION و0.290 لـ STRIP و0.180 لـ Token Substitution.

الإسهامات الرئيسية: (1) أول إطار كشفٍ جنائي قائم على الأزواج وغير مرتبط بنوع الهجوم للأبواب الخلفية في معالجة اللغة الطبيعية؛ (2) أول تقييم منهجي لـ True SCPN كنوع مُحفِّز؛ (3) تقييم للكشف على مهمة نمذجة لغوية (WikiText-2)؛ (4) بنك اختبار شامل من 30 تهيئة يتجاوز أي تقييم منشور سابق.

الكلمات المفتاحية: هجمات الأبواب الخلفية؛ أمن معالجة اللغة الطبيعية؛ الاتساق السلوكي؛ التحليل الجنائي؛ المتانة في مواجهة الخصوم.