PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA

Ministry of Higher Education and Scientific Research

University Ferhat ABBAS Setif 1

Faculty of Sciences

Department of Computer Science

Speciality: IDTW



Ethical Constraints Enforcement for Tool-Using Arabic LLM Agents

Students

Aya Dalache

Nessrine Chebli

Supervisor

Prof. Fouzi Harrag

Master Thesis

Academic Year 2025 / 2026

Abstract

The rapid expansion of Large Language Models (LLMs) into agentic applications has introduced a new class of safety and alignment challenges that existing safeguard frameworks are ill-equipped to address. In multilingual settings, and particularly for Arabic, dialectal diversity creates additional attack surfaces: content moderation systems trained predominantly on Modern Standard Arabic (MSA) are systematically blind to harmful inputs expressed in regional dialects such as Algerian Darja, or in code-mixed registers such as Arabizi and Franco-Arabic. When LLMs are further augmented with tool-calling capabilities, adversarial users gain the ability to manipulate not only model outputs but also external actions, a threat vector known as indirect policy bypass that no prior Arabic safety framework has systematically addressed.

In this work, we propose **AMAN-LLM** (Ethical Constraints Enforcement for Tool-Using Arabic LLM Agents), a modular four-layer pipeline designed to enforce ethical constraints across the full agentic lifecycle. The framework integrates a dialect-aware Input Sanitizer performing UTF-8 normalisation, de-obfuscation, and Arabizi decoding; Layer 1 ethical input filtering via MARBERT and AraBERT classifiers fine-tuned on the OSACT5 hate-speech dataset and a bespoke Algerian Arabic Adversarial Corpus (AAAC); Layer 2 policy governance through a YAML-configurable rule engine covering thirteen violation categories including prompt injection, PII leakage, and jurisdiction-specific financial compliance; Layer 3 output validation enforcing groundedness, hallucination detection, and safety checks on LLM-generated responses; and Layer 4 reflective self-critique implementing a constitutional critique-and-rewrite loop grounded in Arabic-language safety principles. The system is instantiated on Aya Expanse 8B as the tool-augmented agent core and evaluated on a purpose-built benchmark of 118 test cases spanning Modern Standard Arabic, Algerian Darja, Arabizi, and code-mixed Franco-Arabic, structured around four research questions.

Empirical evaluation across RQ1–RQ4 demonstrates strong and consistent performance across all pipeline layers. The dialect-aware Layer 1 classifier achieved an accuracy of 89.2% and an F1-score of 91.6%, with an obfuscation detection rate of 88.9% across all Arabic input modalities (RQ1). End-to-end evaluation yielded an Attack Success Rate of 5.68%, with Layer 4 contributing the largest delivering the only significant incremental reduction beyond Layer 1 — lowering ASR by approximately 73.7% relative to the L1–L2–L3 configuration — and a block precision of 88.5% (RQ2). The Input Sanitizer achieved 100% normalisation accuracy at a near-zero average latency of 0.19 ms, directly improving downstream detection effectiveness (RQ4). Layer 4 individually attained an F1-score of 97.3%, a 100% unsafe-response blocking rate, and a 94.7% self-correction rate; its average latency of 14.6 s and P95 of 22.1 s make it the dominant computational cost, though preliminary evidence confirms that

risk-stratified selective routing is feasible for low-risk outputs (RQ3). Hallucinated outputs were eliminated entirely through the combined action of Layers 3 and 4, and end-to-end median latency was measured at 28.8 ms.

This work provides a principled, extensible, and culturally-grounded architecture for ethical constraint enforcement in Arabic LLM agent systems. By addressing the intersection of Algerian dialectal diversity, agentic threat models, and multi-layer policy enforcement, AMAN-LLM fills a critical gap in Arabic NLP safety research and establishes a replicable blueprint for trustworthy, responsible deployment of LLM agents in Arabic-speaking contexts.

Keywords: Large Language Models, Arabic NLP Safety, Ethical Constraint Enforcement, Algerian Dialect, Arabizi, Indirect Policy Bypass, AMAN-LLM, Adversarial Robustness, Tool-Using Agents, Constitutional AI.

ملخص

أدى التوسع السريع للنماذج اللغوية الكبيرة LLMs نحو التطبيقات الوكيلة (Agentic Applications) إلى ظهور فئة جديدة من تحديات السلامة والمواءمة التي لا تمتلك أطر الحماية الحالية القادرة على التعامل معها. وفي البيئات متعددة اللغات، ولا سيما بالنسبة للغة العربية، يخلق التنوع اللهجي أسطح هجوم إضافية، إذ إن أنظمة الإشراف على المحتوى المدرَّبة بشكل أساسي على العربية الفصحى الحديثة (*MSA*) تعاني من قصور منهجي في اكتشاف المدخلات الضارة المعبَّر عنها باللهجات الإقليمية مثل الدارجة الجزائرية، أو في الصيغ المختلطة مثل العربيزي والفرانكو-عربي. وعندما تُدعَّم النماذج اللغوية الكبيرة بقدرات استدعاء الأدوات (Tool-Calling Capabilities)، يكتسب الخبيثون القدرة على التلاعب ليس فقط بمخرجات النموذج، بل أيضاً بالإجراءات الخارجية، وهو ما يُعرف باسم التجاوز غير المباشر للسياسات (Indirect Policy Bypass)، والذي لم تتناوله أي أطر سابقة لسلامة اللغة العربية بصورة منهجية.

في هذا العمل، نقترح *AMAN-LLM* (فرض القيود الأخلاقية على الوكلاء العرب المعتمدين على النماذج اللغوية الكبيرة والمستخدمين للأدوات)، وهو خط معالجة معياري مكوَّن من أربع طبقات صُمِّم لفرض القيود الأخلاقية عبر دورة حياة الوكيل كاملة. يدمج الإطار منظف مدخلات مدركاً للهجات (Dialect-Aware Input Sanitizer) يقوم بتطبيع ترميز *UTF-8*، وإزالة التمويه، وفك ترميز العربيزي، والطبقة الأولى الخاصة بالترشيح الأخلاقي للمدخلات باستخدام مصنفات *MARBERT* و *AraBERT* المضبوطة بدقة على مجموعة بيانات خطاب الكراهية *OSACT5* وعلى مجموعة بيانات مخصصة للهجمات العدائية في العربية الجزائرية ؛(*AAAC*) والطبقة الثانية الخاصة بحوكمة السياسات من خلال محرك قواعد قابل للتهيئة باستخدام *YAML* يغطي ثلاث عشرة فئة من الانتهاكات، بما في ذلك حقن الأوامر (Prompt Injection)، وتسرب المعلومات الشخصية (PII Leakage)، والامتثال المالي الخاص بالولايات القضائية المختلفة؛ والطبقة الثالثة الخاصة بالتحقق من المخرجات، والتي تفرض الارتكاز إلى الحقائق، وكشف الهلوسة، وإجراء فحوصات السلامة على الاستجابات المولَّدة بواسطة النماذج اللغوية الكبيرة؛ والطبقة الرابعة الخاصة بالنقد الذاتي التأملي، والتي تطبق حلقة نقد وإعادة صياغة دستورية (Constitutional Critique-and-Rewrite Loop) تستند إلى مبادئ السلامة باللغة العربية. وقد تم تجسيد النظام باستخدام Aya Expanse 8B بوصفه النواة الوكيلة المعززة بالأدوات، وتقييمه على مجموعة من 118 حالة اختبار تغطي اربعة تنوعات لهجية عربية. تُظهر النتائج التجريبية عبر أسئلة البحث الأربعة أداءً قوياً ومتسقاً في جميع طبقات خط المعالجة. حقق مصنّف الطبقة الأولى المدرك للهجات دقةً بلغت 89.2% ودرجة F_1 بلغت 91.6%، مع معدل اكتشاف تمويه بلغ 88.9% عبر جميع أنماط المدخلات العربية (*RQ1*). وأسفر التقييم الشامل من البداية إلى النهاية عن معدل نجاح هجمات بلغ 5.68% فقط، إذ أسهمت الطبقة الرابعة في أكبر تخفيض إضافي بتقليل معدل النجاح بنحو 73.7% مقارنةً بتهيئة الطبقات *L1--L2--L3*، مع دقة حجب بلغت 88.5% (*RQ2*). وحقق منظّف المدخلات دقة تطبيع بلغت 100% بزمن استجابة متوسط شبه معدوم يبلغ 0.19 ملي ثانية، مما أسهم مباشرةً في تحسين فاعلية الطبقات اللاحقة (*RQ4*). وعلى مستوى الطبقة الرابعة وحدها، حققت درجة F_1 بلغت 97.3%، ومعدل حجب للاستجابات غير الآمنة بنسبة 100%، ومعدل تصحيح ذاتي قدره 94.7%؛ غير أن متوسط زمن استجابتها البالغ 14.6 ثانية والنسبة المئوية الخامسة والتسعين ($P95$) البالغة 22.1 ثانية يجعلانها المصدر الرئيسي للتكلفة الحسابية، وإن كانت الأدلة الأولية تؤكد جدوى التوجيه الانتقائي القائم على تصنيف

المخاطر .(*RQ*3) وتم القضاء تماماً على المخرجات المهلوسة من خلال التأثير المشترك للطبقتين الثالثة والرابعة، وبلغ زمن الاستجابة الوسيط للنظام كاملاً 28.8 ملي ثانية. يوفر هذا العمل بنية معمارية مبدئية وقابلة للتوسعة ومرتكزة ثقافياً لفرض القيود الأخلاقية في أنظمة الوكلاء العرب المعتمدة على النماذج اللغوية الكبيرة. ومن خلال معالجة التداخل بين التنوع اللهجي الجزائري، ونماذج التهديد الخاصة بالأنظمة الوكيلة، وآليات فرض السياسات متعددة الطبقات، يسد *AMAN-LLM* فجوة مهمة في أبحاث سلامة معالجة اللغة الطبيعية العربية، ويؤسس نموذجاً قابلاً للتكرار للنشر الموثوق والمسؤول لوكلاء النماذج اللغوية الكبيرة في البيئات الناطقة بالعربية.

الكلمات المفتاحية: النماذج اللغوية الكبيرة، سلامة معالجة اللغة الطبيعية العربية، فرض القيود الأخلاقية، اللهجة الجزائرية العربيزي، التجاوز غير المباشر للسياسات، ،*AMAN-LLM* المقاومة ضد الهجمات العدائية، الوكلاء المستخدمون للأدوات، الذكاء الاصطناعي الدستوري.

Résumé

L'expansion rapide des grands modèles de langage (LLMs) dans les applications agentiques a introduit une nouvelle classe de défis en matière de sécurité et d'alignement, auxquels les cadres de protection existants ne sont pas en mesure de répondre. Dans les contextes multilingues, et particulièrement pour l'arabe, la diversité dialectale crée des surfaces d'attaque supplémentaires : les systèmes de modération de contenu entraînés principalement sur l'arabe standard moderne (ASM) sont systématiquement aveugles aux entrées nuisibles exprimées dans des dialectes régionaux tels que la Darja algérienne, ou dans des registres à mélange de codes tels que l'Arabizi et le franco-arabe. Lorsque les LLMs sont en outre augmentés de capacités d'appel d'outils, les utilisateurs adversariaux acquièrent la capacité de manipuler non seulement les sorties du modèle, mais également les actions externes — un vecteur de menace connu sous le nom de contournement indirect de politique qu'aucun cadre de sécurité arabe antérieur n'a abordé de manière systématique.

Dans ce travail, nous proposons AMAN-LLM (Ethical Constraints Enforcement for Tool Using Arabic LLM Agents), un pipeline modulaire à quatre couches conçu pour faire respecter les contraintes éthiques tout au long du cycle de vie agentique. Le cadre intègre : un assainisseur d'entrées sensible aux dialectes effectuant une normalisation UTF-8, une désobfuscation et un décodage de l'Arabizi ; un filtrage éthique des entrées en couche 1 via des classificateurs MARBERT et AraBERT affinés sur le jeu de données de discours haineux OSACT5 et un corpus adversarial d'arabe algérien sur mesure (AAAC) ; une gouvernance des politiques en couche 2 via un moteur de règles configurable en YAML couvrant treize catégories de violations, dont l'injection de prompts, la fuite d'informations personnelles (PII) et la conformité financière spécifique aux juridictions ; une validation des sorties en couche 3 assurant l'ancrage, la détection des hallucinations et des vérifications de sécurité sur les réponses générées par le LLM ; et une auto-critique réflexive en couche 4 implémentant une boucle constitutionnelle de critique et de réécriture fondée sur des principes de sécurité en langue arabe. Le système est instancié sur Aya Expanse 8B comme noyau agent augmenté d'outils et évalué sur un benchmark de 118 cas de test couvrant l'arabe standard moderne, la Darja algérienne, l'Arabizi et le franco-arabe, structuré autour de quatre questions de recherche.

L'évaluation empirique menée à travers RQ1–RQ4 démontre des performances solides et cohérentes sur l'ensemble des couches du pipeline. Le classificateur de la couche 1, sensible aux dialectes, a atteint une précision de 89,2% et un score F1 de 91,6%, avec un taux de détection des obfuscations de 88,9% sur toutes les modalités d'entrée arabes (RQ1). L'évaluation de bout en bout a produit un taux de succès des attaques de 5,68%, la couche 4 contribuant à la plus grande réduction secondaire — abaissant l'ASR d'environ 73,7% par rapport à la

configuration L1–L2–L3 — avec une précision de blocage de 88,5% (RQ2). L'assainisseur d'entrées a atteint une précision de normalisation de 100% à une latence moyenne quasi nulle de 0,19 ms, améliorant directement l'efficacité des couches en aval (RQ4). La couche 4 a individuellement atteint un score F1 de 97,3%, un taux de blocage des réponses non sécurisées de 100%, et un taux d'auto-correction de 94,7% ; sa latence moyenne de 14,6 s et son P95 de 22,1 s en font la principale source de coût computationnel, bien que des preuves préliminaires confirment la faisabilité d'un routage sélectif stratifié par risque pour les sorties à faible risque (RQ3). Les sorties hallucinées ont été entièrement éliminées par l'action combinée des couches 3 et 4, et la latence médiane de bout en bout a été mesurée à 28,8 ms.

Ce travail fournit une architecture raisonnée, extensible et culturellement ancrée pour l'application de contraintes éthiques dans les systèmes d'agents LLM arabes. En abordant l'intersection de la diversité dialectale algérienne, des modèles de menaces agentiques et de l'application des politiques multicouches, AMAN-LLM comble une lacune critique dans la recherche sur la sécurité du TAL arabe et établit un schéma reproductible pour un déploiement fiable et responsable d'agents LLM dans les contextes arabophones.

Mots-clés : Grands modèles de langage, Sécurité du TAL arabe, Application des contraintes éthiques, Dialecte algérien, Arabizi, Contournement indirect de politique, AMAN-LLM, Robustesse adversariale, Agents utilisant des outils, IA constitutionnelle.

List of Abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
API	Application Programming Interface
AraBERT	Arabic Bidirectional Encoder Representations from Transformers
BERT	Bidirectional Encoder Representations from Transformers
CPU	Central Processing Unit
DB	Database
DoS	Denial of Service
DPI	Direct Prompt Injection
FN	False Negative
FNR	False Negative Rate
FPR	False Positive Rate
GPU	Graphics Processing Unit
HF	Hugging Face
HTTP	HyperText Transfer Protocol
IPI	Indirect Prompt Injection
IR	Intent Recognition
JSON	JavaScript Object Notation
LLM	Large Language Model
MARBERT	Multi-Dialect Arabic BERT
MARBERTv2	Multi-Dialect Arabic BERT Version 2
MITRE ATT&CK	Adversarial Tactics, Techniques, and Common Knowledge
NLP	Natural Language Processing
OPA	Output Policy Assessment
OWASP	Open Worldwide Application Security Project
PII	Personally Identifiable Information
RAG	Retrieval-Augmented Generation
RB	Risk Band
RBAC	Role-Based Access Control
RL	Reflective Layer
RS	Risk Score

Abbreviation	Full Form
TG	Tool Governance
TM	Threat Modeling
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
UUID	Universally Unique Identifier
XSS	Cross-Site Scripting
YAML	YAML Ain’t Markup Language

Contents

Abstract 1

ملخص 3

Résumé 5

List of Abbreviations 7

General Introduction 14

1 Background 19
- 1.1 Introduction . . . 19
- 1.2 From Tool-Calling to Risk: The Agentic Threat Surface . . . 19
- 1.3 The Arabic Dialectal Safety Landscape . . . 21
- 1.4 Indirect Policy Bypass: Definition and Scope . . . 21
- 1.5 Implications for Enforcement Design . . . 22
- 1.6 Conclusion . . . 22

2 Literature Review 23
- 2.1 Introduction . . . 23
- 2.2 Related Works . . . 23
- 2.3 Discussion . . . 31
  - 2.3.1 Methodological Approaches . . . 31
  - 2.3.2 Coverage of Ethical and Safety Threats . . . 32
  - 2.3.3 Arabic Language-Specific Limitations . . . 33
  - 2.3.4 Evaluation Gaps . . . 33
  - 2.3.5 Emerging Trends . . . 34
  - 2.3.6 Critical Challenges . . . 34
  - 2.3.7 Research Gap and Positioning of This Thesis . . . 35
- 2.4 Conclusion . . . 36

3 Conceptual Model 37
- 3.1 Introduction . . . 37
- 3.2 Model Architecture . . . 37
- 3.3 Step-by-Step Description of the Pipeline . . . 38
  - 3.3.1 Input Sanitizer: . . . 38

3.3.2 Layer 1: Input-Level Ethical Filtering (Layer1Orchestrator) 39
3.3.3 Agent Core: Arabic LLM Reasoning and Planning (Aya Expanse 8B) 40
3.3.4 Layer 2: Tool Invocation Governance (Policy Gatekeeper) 41
3.3.5 Layer 3: Runtime Monitoring and Output Validation 42
3.3.6 Layer 4: Reflective Reasoning (Self-Critique) 42
3.3.7 Audit Logger and Distributed Traceability 43
3.4 Security Properties Targeted 43
3.5 Scenario and Model Specification 44
3.6 Evaluation Metrics 45
3.7 Conclusion 47

4 Implementation **48**
4.1 Introduction 48
4.2 Languages and Tools 48
4.3 Implementation Environment 50
4.3.1 Hardware Configuration 50
4.3.2 Software Configuration 50
4.3.3 Development Tools and Frameworks 51
4.4 Datasets and Data Preparation 51
4.4.1 OSACT5 Dataset 53
4.4.2 Algerian Dataset (AAAC) 55
4.4.3 Data Cleaning and Preprocessing 58
4.5 Fine-Tuning of Arabic Language Models 61
4.5.1 Motivation for Fine-Tuning 62
4.5.2 Model Selection: Preference Over Multilingual Models 62
4.5.3 Fine-Tuning Dataset Preparation 63
4.5.4 Training Configuration 64
4.5.5 Training Procedure 65
4.5.6 Loss Function 65
4.6 Implementation of the Proposed Architecture 66
4.6.1 Input Sanitization 66
4.6.2 Layer 1: Input Safety Layer 68
4.6.3 Agent Core: Aya Expanse 8B Reasoning Engine 70
4.6.4 Layer 2: Tool Invocation Governance 72
4.6.5 Layer 3: Output Validation & Quality Assurance 73
4.6.6 Layer 4: Reflective Reasoning and Self-Critique 75
4.6.7 Pipeline Orchestrator 75
4.6.8 Audit Logging and Traceability 76

4.6.9 Experiments 76
4.6.10 AMAN-LLM Execution Examples 81
4.6.11 Comparison with Existing Work : 82
4.6.12 Challenges 87
4.6.13 Limitations 89
4.6.14 Future Work 90
4.7 Conclusion 92

General Conclusion **94**