

الجمهورية الشعبية الديمقراطية الجزائرية

People's Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي

Ministry of Higher Education and Scientific Research

المدرسة العليا للإعلام الآلي 08 ماي 1945 بسيدي بلعباس

Higher School of Computer Science

-08 May 1945- Sidi Bel Abbès



THESIS

To obtain the diploma of **Master**

Field: **Computer Science**

Specialty: **Artificial Intelligence and Data Science (IASD)**

Theme

Authority, Not Intelligence:

Where Correctness Lives When a Language Model Is in the Loop

Presented by:

BENAHMED Firdaws Sotra

Submission Date: **23/09/2026**

In front of the jury composed of:

Dr. SAIDI Mohammed Elfatih

President

Dr. BOUSMAHA Rabab

Supervisor

Dr. ELHANNANI Souad

Examiner

Academic Year: 2025/2026

Abstract

Language models are increasingly placed inside systems that must be correct: systems that write to databases, enforce access boundaries, and act on behalf of organisations that cannot verify every result by hand. This thesis examines where correctness can be enforced in such systems, and argues that it cannot be enforced by the model itself.

Through a structured, criterion-driven review of the literature on structured generation, constrained decoding, validation boundaries, and agentic tool use — organised throughout by the question of *what each mechanism constrains* — this work traces a shift in how language models are positioned in software architecture: from an oracle whose output is consumed directly, to an untrusted client of deterministic infrastructure. It examines the mechanisms by which correctness is reclaimed downstream of the model — typed schemas, database constraints, transaction boundaries, tenancy discriminators, and human review — and characterises the failure modes that survive each of them. A distinction is drawn throughout between *transcription* failures, in which a datum exists and is rendered wrongly, and *fabrication* failures, in which the model supplies a plausible structure where the source supplied nothing; the second is shown to be the harder case, because it satisfies every structural check placed downstream.

The study further analyses a second axis of the problem: where inference physically runs. Open-weight models served on commodity hardware offer full data custody at a cost in capability, latency, and configuration fragility. The literature reporting these trade-offs is surveyed alongside the evaluation practices — condition reporting, repeated runs, calibration measurement — that make such comparisons meaningful rather than anecdotal.

The synthesis contributes an explicit *division-of-authority* framework: four questions that locate, for any component of a model-driven data path, which properties are guaranteed, which are merely probable, and where the boundary between interpretation and adjudication has silently been crossed. The framework is applied to measurements drawn from an

accompanying engineering thesis that implements the same principle in a working multi-tenant platform, where it correctly identifies a delegation the system’s own designers had not noticed — deterministic findings passed to a model for rephrasing, and replaced by it in part of the runs. The review also identifies where current mitigation strategies remain insufficient: against fabrication, against uncalibrated confidence displays, and in the reporting practices that make local-deployment comparisons meaningful.

Keywords — Large Language Models, Structured Output, Schema Validation, Constrained Decoding, Hallucination Mitigation, Calibration, Multi-Tenant Systems, Local Inference, Quantisation, Tool Use, Model Context Protocol, Trustworthy AI

الملخص

تُدجّ النماذج اللغوية الكبيرة اليوم بشكل متزايد داخل أنظمة يُشترط فيها الصواب: أنظمة تكتب في قواعد البيانات، وتفرض حدود الوصول، وتتمصرف نيابةً عن مؤسسات لا تستطيع التحقق يدوياً من كل نتيجة. تبحث هذه المذكرة في الموضوع الذي يمكن فيه ضمان الصواب في مثل هذه الأنظمة، وتدافع عن أطروحة مفادها أن هذا الضمان لا يمكن أن يأتي من النموذج نفسه.

من خلال مراجعة نقدية مُنظمة لأدبيات التوليد المهيكل، وفك الترميز المقيّد، وحدود التحقق، واستخدام الأدوات بواسطة الوكلاء --- مراجعة تنتظم كلها حول سؤال واحد هو ما الذي تضمنه كل آلية فعلياً --- يتتبع هذا العمل تحوُّلاً في كيفية إدراج النماذج اللغوية ضمن بنية البرمجيات: من كونها مصدراً موثقاً تُستهلك مخرجاته مباشرةً، إلى كونها عميلاً غير موثوق لبنية تحتية حتمية. ويفحص الآليات التي يُستعاد بها الصواب بعد النموذج --- المخططات المُنمّطة، وقيود قاعدة البيانات، وحدود المعاملات، ومعرّفات المستأجرين، والمراجعة البشرية --- ويصف أنماط الفشل التي تنجو من كل منها. ويُبيّن في كل هذا العمل بين أخطاء النسخ، حيث تكون القيمة موجودة في المصدر لكنها تُنقل خطأً، وأخطاء الاختلاق، حيث يُنتج النموذج بنيةً معقولة في موضع لم يقدم فيه المصدر شيئاً؛ ويُبيّن أن الحالة الثانية هي الأصعب لأنها تتجاوز كل فحص بنيوي يوضع بعدها.

وتحلل الدراسة كذلك محوراً ثانياً للمشكلة: مكان تنفيذ الاستدلال فعلياً. فالنماذج مفتوحة الأوزان التي تُشغّل على عتاد عادي توفر حياة كاملة للبيانات مقابل كلفة في القدرة، وزمن الاستجابة، وهشاشة الإعداد. وتُستعرض الأدبيات التي تُبغ عن هذه المفاضلات إلى جانب ممارسات التقييم --- الإبلاغ عن الشروط، وتكرار التجارب، وقياس المعايير --- التي تجعل مثل هذه المقارنات ذات معنى لا مجرد رواية.

تقدم هذه الدراسة إطاراً صريحاً لتقسيم السلطة: أربعة أسئلة تحدد، لأيّ مكوّن في مسار بيانات يقوده نموذج لغوي، ما هو مضمون وما هو محتمل فحسب، وأين تُجوز الحدود بين التأويل والبتّ دون أن ينتبه أحد. ويُطبّق الإطار على قياسات من مذكرة هندسية مرافقة، فيكشف فيها تفويضاً لم يلحظه مصممو النظام أنفسهم: نتائج حتمية أُرسلت إلى النموذج لإعادة صياغتها فاستبدلها. كما تحدد المراجعة مواضع القصور الباقية: أمام الاختلاق، وأمام عروض الثقة غير المعايير، وفي ممارسات الإبلاغ التي تجعل مقارنات النشر المحلي ذات معنى.

الكلمات المفتاحية --- النماذج اللغوية الكبيرة، المخرجات المهيكلية، التحقق من المخطط، فك الترميز المقيّد، التخفيف من الهلوسة، المعايير، الأنظمة متعددة المستأجرين، الاستدلال المحلي، التكميم، استخدام الأدوات، بروتوكول سياق النموذج، الذكاء الاصطناعي الموثوق

Résumé

Les modèles de langage sont de plus en plus intégrés à des systèmes qui doivent être corrects : des systèmes qui écrivent en base de données, appliquent des frontières d'accès et agissent pour le compte d'organisations incapables de vérifier chaque résultat à la main. Ce mémoire examine où la correction peut être garantie dans de tels systèmes, et soutient qu'elle ne peut pas l'être par le modèle lui-même.

À travers une revue structurée et critique de la littérature sur la génération structurée, le décodage contraint, les frontières de validation et l'usage agentique d'outils — organisée de bout en bout par la question de *ce que chaque mécanisme contraint* — ce travail retrace un déplacement dans la manière dont les modèles de langage sont positionnés dans l'architecture logicielle : d'un oracle dont la sortie est consommée directement, vers un client non fiable d'une infrastructure déterministe. Il examine les mécanismes par lesquels la correction est reprise en aval du modèle — schémas typés, contraintes de base de données, frontières transactionnelles, discriminateurs de locataire et revue humaine — et caractérise les modes de défaillance qui survivent à chacun d'eux. Une distinction est établie entre les défaillances de *transcription*, où une donnée existe et est restituée incorrectement, et les défaillances de *fabrication*, où le modèle produit une structure plausible là où la source n'en fournissait aucune ; le second cas est le plus difficile, car il satisfait tous les contrôles structurels placés en aval.

L'étude analyse ensuite un second axe du problème : le lieu physique de l'inférence. Les modèles à poids ouverts servis sur matériel grand public offrent une souveraineté complète des données au prix de la capacité, de la latence et d'une configuration fragile. La littérature rapportant ces compromis est passée en revue avec les pratiques d'évaluation qui rendent de telles comparaisons significatives.

La synthèse propose un cadre explicite de *division de l'autorité* : quatre questions qui permettent, pour tout composant d'une chaîne de données pilotée par un modèle, de déterminer quelles propriétés sont garanties, lesquelles ne sont que probables, et où la frontière

entre interprétation et décision a été franchie sans que personne le remarque. Ce cadre est appliqué à des mesures issues d'un mémoire d'ingénieur qui implémente le même principe dans une plateforme multi-locataires opérationnelle, où il identifie une délégation que les concepteurs du système n'avaient pas vue — des constats déterministes confiés à un modèle pour reformulation, et remplacés par lui dans une partie des exécutions. La revue identifie également les points où les stratégies d'atténuation actuelles restent insuffisantes : face à la fabrication, face aux affichages de confiance non calibrés, et dans les pratiques de report qui rendent les comparaisons de déploiement significatives.

Mots clés — Grands modèles de langage, Sortie structurée, Validation de schéma, Décodage contraint, Atténuation des hallucinations, Calibration, Systèmes multi-locataires, Inférence locale, Quantisation, Usage d'outils, Model Context Protocol, IA de confiance