People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbes

Master's Thesis

To obtain the diploma of **Master's Degree**

Field of Study: **Computer Science**

Specialization: **Artificial Intelligence and Data Science**

Theme

Hallucination In Large Language Models

Presented by
Kebli Younes
Atallah Chaker Abdeldjalil

Defended on: **24/09/2025**
In front of the jury composed of

Mme KLOUCHE Badia	President
Mr SAIDI Mohamed El Fatih	Examiner
Mr KESKES Nabil	Supervisor

Academic Year: 2024/2025

Abstract

Large Language Models (LLMs) have reached remarkable performance across an impressive breadth of natural-language tasks, yet they remain vulnerable to hallucinations — answers that are plausible at first sight but ultimately false or fabricated. This thesis offers an end-to-end study of hallucination detection that relies on *only one* model response, sidestepping the heavy compute and storage demands of multi-response consistency checks or large-scale retrieval. We introduce lightweight detectors that operate in (i) white-box and (ii) black-box regimes by analyzing internal activations, attention patterns, and token-level confidences from an auxiliary LLM. When ground-truth references exist, as in Retrieval-Augmented Generation (RAG), our framework integrates external evidence to boost reliability. Experiments on multiple open benchmarks show speed-ups of up to 45× (white-box) and 450× (black-box) over prevailing baselines, while achieving substantial gains in AUC and F1. These results demonstrate that real-time, high-precision hallucination detection is attainable without sacrificing computational efficiency.

Résumé

Les grands modèles de langage (LLMs) atteignent des performances remaraquables sur une vaste gamme de tâches, mais demeurent sujets aux hallucinations : des réponses apparemment plausibles qui se révèlent pourtant fausses ou inventées. Ce mémoire présente une étude de bout en bout de la détection d'hallucinations reposant sur *une seule* réponse du modèle, évitant ainsi les coûts élevés en calcul et en mémoire des vérifications de cohérence multi-réponses ou des méthodes de recherche à grande échelle. Nous proposons des détecteurs légers fonctionnant en (i) mode « boîte blanche » et (ii) mode « boîte noire », en exploitant les activations internes, les cartes d'attention et les probabilités token-par-token d'un LLM auxiliaire. Lorsque des références véridiques sont disponibles — par exemple dans la génération augmentée par recherche (RAG) — notre cadre s'y intègre naturellement pour accroître la fiabilité. Sur plusieurs jeux de données ouverts, nos méthodes réduisent le temps d'inférence jusqu'à 45× (boîte blanche) et 450× (boîte noire) par rapport aux baselines, tout en améliorant nettement l'AUC et le F1. Ces résultats démontrent qu'une détection des hallucinations à la fois en temps réel et précise est possible sans compromis sur l'efficacité computationnelle.