

الجمهورية الشعبية الديمقراطية الجزائرية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
المدرسة العليا للإعلام الآلي 08 ماي 1945. بسيدي بلعباس
École Supérieure en Informatique -08 Mai 1945- Sidi Bel Abbès



THESIS

To obtain the diploma of **Master**
Field: **Computer Science**
Specialty: **Intelligence Artificielle et science des données**

Theme

**Large Language Models in Web Scraping and Structured Data Extraction: A
Theoretical Analysis and State-of-the-Art Review**

Presented by:
HAGANI Abla

Submission Date: **Sept 30th, 2025**

In front of the jury composed of:

Dr. KLOUCHE BADIA

President

Dr. BOUSMAHA Rabab

Examiner

Pr. BENSLIMANE Sidi Mohamed

Supervisor

Pr. ZIOU Djemel

Co-Supervisor

Academic Year: 2024/2025

Abstract

Traditional web scraping methods struggle with the increasingly dynamic and heterogeneous nature of modern web architectures. This thesis presents a comprehensive theoretical analysis and state-of-the-art review of how Large Language Models (LLMs) are transforming web scraping and structured data extraction.

Through systematic literature review and comparative analysis, this research explores the shift from rule-based extraction methods to intelligent, semantically-aware approaches powered by transformer architectures. The study examines how LLMs leverage natural language understanding, few-shot learning, and context-aware processing to overcome limitations in structural variability, dynamic content generation, and cross-domain adaptability.

The research analyzes current LLM-based extraction methodologies, from direct prompt-based approaches to multi-modal frameworks integrating visual and textual web understanding. Our comparative evaluation reveals significant advances in extraction accuracy and generalization capabilities, while identifying challenges in processing speed, computational costs, and reliability.

This work contributes a comprehensive theoretical framework for understanding LLM applications in web extraction and identifies promising directions for advancing intelligent data extraction technologies.

Keywords— Large Language Models, Web Scraping, Data Extraction, Natural Language Processing, Information Retrieval, Transformer Architecture, Multi-modal Learning, Prompt Engineering, Semantic Understanding, Automated Extraction

الملخص

تواجه طرق استخراج البيانات التقليدية من الويب صعوبات متزايدة نتيجة للطبيعة الديناميكية والمتنوعة لبنية الويب الحديثة. يقدم هذا البحث تحليلاً نظرياً شاملاً ومراجعة لأحدث ما توصلت إليه التقنيات حول كيفية إحداث نماذج اللغة الكبيرة (LLMs) لتحول جذري في طرق استخراج البيانات المنظمة من الويب.

ومن خلال مراجعة منهجية للأدبيات وتحليل مقارن، يستعرض هذا البحث الانتقال من الأساليب المعتمدة على القواعد إلى أساليب ذكية تعتمد على الفهم الدلالي، والمدعومة بهياكل المحولات (Transformers) يسلط البحث الضوء على كيفية استغلال LLMs لفهم اللغة الطبيعية، والتعلم القائم على أمثلة قليلة (few-shot learning)، والمعالجة القائمة على السياق للتغلب على قيود التنوع البنيوي، وتوليد المحتوى الديناميكي، وقابلية التكيف عبر مختلف المجالات.

كما يحلل البحث المنهجيات الحديثة المعتمدة على LLM لاستخراج البيانات، بدءاً من الأساليب المباشرة المعتمدة على التعليمات (prompts) وصولاً إلى الأطر متعددة الوسائط التي تدمج الفهم النصي والبصري لصفحات الويب. تكشف الدراسة عن تقدم ملحوظ في دقة الاستخراج وقدرات التعميم، مع تحديد تحديات متعلقة بسرعة المعالجة، وتكاليف الحوسبة، والموثوقية.

تُسهّم هذه الأطروحة في وضع إطار نظري شامل لفهم تطبيقات LLM في استخراج البيانات من الويب، كما تحدد اتجاهات واعدة لتطوير تقنيات استخراج البيانات الذكية.

الكلمات المفتاحية: نماذج اللغة الكبيرة، استخراج البيانات من الويب، معالجة اللغة الطبيعية، استرجاع المعلومات، بنية المحولات، التعلم متعدد الوسائط، هندسة التعليمات، الفهم الدلالي، الاستخراج الآلي.

Résumé

Les méthodes traditionnelles de scraping web rencontrent des difficultés face à la nature de plus en plus dynamique et hétérogène des architectures web modernes. Ce mémoire présente une analyse théorique approfondie ainsi qu'un état de l'art sur la manière dont les grands modèles de langage (LLMs) transforment l'extraction de données structurées à partir du web.

À travers une revue systématique de la littérature et une analyse comparative, cette recherche explore la transition des méthodes basées sur des règles vers des approches intelligentes et sémantiquement informées, reposant sur les architectures de type transformeur. L'étude examine comment les LLMs exploitent la compréhension du langage naturel, l'apprentissage par quelques exemples (few-shot learning) et le traitement contextuel pour surmonter les limites liées à la variabilité structurelle, à la génération dynamique de contenu et à l'adaptabilité inter-domaines.

La recherche analyse les méthodologies actuelles d'extraction basées sur les LLMs, allant des approches directes par prompts aux cadres multi-modaux intégrant la compréhension textuelle et visuelle des pages web. Notre évaluation comparative met en évidence des avancées notables en termes de précision d'extraction et de capacités de généralisation, tout en identifiant des défis tels que la vitesse de traitement, les coûts computationnels et la fiabilité.

Ce travail apporte un cadre théorique complet pour comprendre les applications des LLMs dans l'extraction de données web et identifie des pistes prometteuses pour faire progresser les technologies d'extraction intelligente.

Keywords— Grands modèles de langage, Scraping web, Extraction de données, Traitement automatique du langage naturel, Recherche d'information, Architecture Transformeur, Apprentissage multi-modal, Conception de prompts, Compréhension sémantique, Extraction automatisée.