

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي والبحث 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 **Engineer**
Field: **Computer Science**
Specialty: **Artificial Intelligence & Data Science (IASD)**

Theme

**BAC Students: An Educational Social Media Platform
with a Domain-Specific Retrieval-Augmented AI
System**

Presented by:
KRIM Abdelouahab

Submission Date: **September, 2026**
In front of the jury composed of:

Dr. CHAIB Souleyman	President
Dr. MAHAMMED Nadir	Examiner
Dr. KHALDI Belkacem	Supervisor
Dr. MILOUA Salah	Incubator Representative

Academic Year : 2025/2026

Abstract

This graduation project is carried out in two phases. The first is the design, implementation, and growth of **BAC Students**, an educational social media platform for Algerian baccalaureate students, grown to 130,000 registered users and a real corpus of 141,083 forum interactions (posts, comments, replies) alongside quizzes, flashcards, and task lists, and the governed data pipeline built on that real usage history.

The second phase is **BAC AI**, a retrieval-augmented generation (Retrieval-Augmented Generation (RAG)) system built on top of the platform that grounds a study-assistant feature in the platform’s own real question-and-answer history and characterizes the script variation across it: Algerian students write in Arabic script, Arabizi (Latin-script phonetic transliteration), and French code-switched text, often within a single message, a variation no public retrieval system or benchmark is built to handle.

A 248-query retrieval benchmark stratified by script, built against a 29,522-unit knowledge base from the platform’s own anonymized data, is used to run a controlled comparison of five retrieval configurations and three query-normalization strategies, evaluated with paired bootstrap and Wilcoxon significance testing. A general multilingual encoder hybridized with lexical search outperforms an Arabic-specialized encoder across every script, and both tested normalization strategies reduce effectiveness rather than improving it; fine-tuning the selected encoder on the platform’s own data improves effectiveness further, most on Arabizi queries. The selected configuration is implemented as a retrieval service integrated into the platform.

BAC Students received the innovative project label and is officially registered as a startup following a year of operation as a market-tested MVP. The report closes with a business study presenting BAC AI as its first AI-driven revenue lever.

Keywords— Retrieval-augmented generation, dense retrieval, dialectal Arabic, Arabizi, script variation, information retrieval, educational technology, Algeria

الملخص

يُجَزَّ هذا المشروع التخرّجي على مرحلتين. المرحلة الأولى هي تصميم وتطوير ونموّ منصة BAC Students، وهي منصة تواصل اجتماعي تعليمية لطلبة البكالوريا الجزائريين، نمت لتصل إلى 130.000 مستخدم مسجّل ومجموعة بيانات واقعية من 141.083 تفاعلاً في المنتدى (منشورات، تعليقات، ردود)، إضافةً إلى الاختبارات القصيرة والبطاقات التعليمية وقوائم المهام، وخط معالجة البيانات المحوّم المبني على هذا التاريخ الحقيقي للاستخدام.

المرحلة الثانية هي BAC AI، نظام قائم على الاسترجاع المعزّز بالتوليد (RAG) مبنيّ فوق المنصة، يُرسي ميزة مساعد دراسي على أساس التاريخ الحقيقي لأسئلة الطلبة وإجابات أقرانهم، ويصف تنوّع الكتابة عبر محتوى المنصة: إذ يكتب الطلبة الجزائريون بالحروف العربية، وبالعريزية (كتابة صوتية بالحروف اللاتينية)، وبنص يتناوب لغوياً مع الفرنسية، غالباً ضمن الرسالة نفسها، وهو تنوّع لا يوجد نظام أو معيار مرجعي عمومي للاسترجاع مصمّم للتعامل معه.

بُني معيار مرجعي للاسترجاع من 248 استعلاماً مصنّفاً حسب نوع الكتابة، على قاعدة معارف من 29.522 وحدة مبنية من بيانات المنصة الحقيقية والمجهولة الهوية، واستُخدم لإجراء مقارنة مضبوطة بين خمسة إعدادات استرجاع وثلاث استراتيجيات لطبيع الاستعلامات، قُيِّمت باستخدام المعاينة العشوائية المزدوجة المتقارنة واختبار ويلكوكسون. يتفوّق مررّز عام متعدد اللغات مدموج مع بحث معجمي على مررّز متخصص في اللغة العربية عبر كل أنواع الكتابة، وتقلّل كلتا استراتيجيتي التطبيع المختبرتين من الفعالية بدل تحسينها؛ ويُحسّن الضبط الدقيق للمررّز المعتمد على بيانات المنصة نفسها الفعالية أكثر، وبشكل أوضح في استعلامات العريزية. نُفِّذ الإعدادات المعتمدة كخدمة استرجاع مدمجة في المنصة.

حصلت BAC Students على وسم المشروع المبتكر وهي مسجّلة رسمياً كشركة ناشئة بعد سنة من التشغيل كنسخة أولية (MVP) خضعت لاختبار السوق. يُختتم هذا التقرير بدراسة تجارية تُقدّم فيها BAC AI بوصفها الرافعة الأولى للإيرادات القائمة على الذكاء الاصطناعي.

الكلمات المفتاحية: الاسترجاع المعزّز بالتوليد، الاسترجاع الكثيف، اللهجة الجزائرية، العريزية، تنوّع الكتابة، استرجاع المعلومات، التكنولوجيا التعليمية، الجزائر.

Résumé

Ce projet de fin d'études se déroule en deux phases. La première est la conception, la mise en œuvre et la croissance de **BAC Students**, une plateforme de réseau social éducatif pour les étudiants algériens du baccalauréat, développée jusqu'à 130 000 utilisateurs inscrits et un corpus réel de 141 083 interactions sur le forum (publications, commentaires, réponses), ainsi que des quiz, des fiches de révision et des listes de tâches, et le pipeline de données gouverné construit sur cet historique d'usage réel.

La seconde phase est **BAC AI**, un système de génération augmentée par récupération (RAG) construit au-dessus de la plateforme, qui ancre une fonctionnalité d'assistant d'étude dans l'historique réel de questions-réponses de la plateforme et caractérise la variation scripturale qui la traverse : les étudiants algériens écrivent en écriture arabe, en Arabizi (translittération phonétique en caractères latins) et en texte à alternance codique avec le français, souvent au sein d'un même message, une variation qu'aucun système ou benchmark de recherche d'information public n'est conçu pour gérer.

Un benchmark de recherche de 248 requêtes stratifiées par graphie, construit sur une base de connaissances de 29 522 unités issue des données réelles et anonymisées de la plateforme, est utilisé pour mener une comparaison contrôlée de cinq configurations de recherche et de trois stratégies de normalisation des requêtes, évaluée par bootstrap apparié et test de Wilcoxon. Un encodeur multilingue général hybridé avec une recherche lexicale surpasse un encodeur spécialisé en arabe sur chaque graphie, et les deux stratégies de normalisation testées réduisent l'efficacité au lieu de l'améliorer ; l'ajustement fin de l'encodeur retenu sur les données propres à la plateforme améliore encore l'efficacité, surtout sur les requêtes en Arabizi. La configuration retenue est implémentée sous forme de service de recherche intégré à la plateforme.

BAC Students a reçu le label de projet innovant et est officiellement enregistrée comme startup après une année de fonctionnement en tant que MVP testé sur le marché. Le rapport se termine par une étude commerciale présentant BAC AI comme son premier levier de revenus fondé sur l'IA.

Mots clés — Génération augmentée par récupération, recherche dense, arabe dialectal, Arabizi, variation scripturale, recherche d'information, technologie éducative, Algérie.