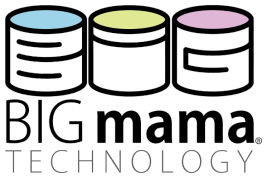
People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
National Polytechnic School of Algiers



Graduation Thesis

To obtain the Industrial Engineering diploma

Option : Data Science & Artificial Intelligence

Unlocking Potential: Open Source LLMs in Rag Systems

Written by:
Ghribi Ouassim Abdelmalek

Jury:

President:

Prof. Philippe Cudré-Mauroux - National Polytechnic School of Algiers

Examiners:

Prof. Pierre Senellart - National Polytechnic School of Algiers

Supervisors:

Mr. Oussama Arki - National Polytechnic School of Algiers
Mr. Hachem Betrounie - BIGmama Technology

ENP 2024
10, Avenue des Frères Oudek, Hassen Badi, BP. 182, 16200 El Harrach, Alger, Algérie.

In the Name of God, Most Gracious, Most Merciful

Acknowledgments

I extend my deepest gratitude to my thesis supervisors, Mr. Arki Oussama and Mr. Betrouni Hachem, for their indispensable guidance and unwavering support throughout this project. Their expertise and continuous encouragement have been pivotal in shaping its direction. I am sincerely thankful for their patience, insightful feedback, and generous investment of time, all of which have been crucial in refining this work.

On a personal level, I am profoundly grateful to my family and friends for their unwavering support and encouragement throughout this demanding yet rewarding journey. Their unwavering belief in my abilities has been a constant source of motivation, particularly during challenging times.

Moreover, I would like to express my appreciation to all individuals who directly or indirectly contributed to the completion of this thesis. Your support and encouragement have been invaluable and deeply cherished.

Special recognition goes to Mr. Hadj Khelil for his belief in me over the past two years. Progressing from a junior front-end developer to a team lead, overseeing an exceptional team of ambitious individuals, has been an enriching experience. Together, we have strived to develop software with the potential to make a positive impact on the world.

Abstract

Hyko, currently functioning as a drag-and-drop AI tool builder, empowers users to connect multiple machine learning (ML) models and construct executable pipelines. However, a notable limitation lies in its incapacity to integrate user feedback, thus hampering the customization potential of its solutions. This thesis aims to bridge this gap by proposing a framework to enhance model performance through user-driven reinforcement learning. Our approach involves devising a system where users can provide online feedback and ratings, thereby deriving rewards for the model within a reinforcement learning paradigm. By doing so, we aspire to augment the customizability of solutions within the Hyko AI tool builder, paving the way for more adaptive and user-centric ML pipelines.

We detail the methodology for implementing this framework within the Hyko environment, including the design of user-friendly feedback interfaces and the integration of reinforcement learning algorithms. Practical evaluations demonstrate the feasibility and effectiveness of our approach across various domains, showcasing its potential for real-world applications.

This research contributes to advancing user-driven customization in AI systems, envisioning a future where ML solutions are not only powerful but also responsive to user preferences and evolving requirements.

Table of Contents

List of Figures

List of Tables

Listings

Glossary

Here are the main acronyms used in this document. The meaning of an acronym is usually indicated once, when it first appears in the text.

NLP	Natural Language Processing
LLM	Large Language model

Chapter 1

Introduction

"The true power of AI lies in its ability to augment human intelligence, not replace it."
Andrew Ng, Co-founder of Coursera and former Chief Scientist at Baidu

In this chapter, we embark on a journey through the landscape of existing literature and research pertinent to the convergence of open source LLMs and Rag systems, in collaboration with BIGmama Technology. Our objective is to establish a robust foundation for understanding the intricacies and potentials inherent in this symbiotic relationship. As we delve into the state of play, we aim to illuminate key theoretical underpinnings and empirical insights that underpin our study of open source LLMs and Rag systems within the context of BIGmama Technology. To initiate this exploration, we contextualize our inquiry within the broader discourse of natural language processing and information retrieval, elucidating the multifaceted challenges and opportunities presented by the integration of open source LLMs and Rag systems. Subsequently, we navigate through seminal concepts and methodologies, including the principles of LLM architecture, Rag framework components, and their applications in various domains. These foundational insights serve as pillars supporting our quest to explore the potential synergies and advancements offered by the convergence of open source LLMs and Rag systems within the framework of BIGmama Technology.

1.1 Presenting BIGmama Technology

BIGmama is an innovative startup founded in France with a subsidiary in Algeria, specializing in data science and artificial intelligence solutions. With over 9 years of experience as of 2024, the company has established itself as a leading provider of bespoke predictive applications tailored to meet the unique needs of its clients.

Guided by a distinguished board of directors comprising former CEOs of global conglomerates such as Danone and Safran, BIGmama boasts a team of highly skilled data scientists and software engineers. This multidisciplinary team, consisting of more than a dozen experts in their respective fields, brings a wealth of knowledge and expertise to the company's endeavors.

BIGmama's commitment to innovation and cutting-edge technologies has allowed them to deliver state-of-the-art solutions that leverage the power of data science and

artificial intelligence. With a strong focus on providing customized solutions, the company has consistently exceeded its clients' expectations, enabling them to gain valuable insights and make data-driven decisions.

Through its strategic partnerships and collaborations, BIGmama continues to push the boundaries of what is possible in the realms of data science and AI, positioning itself as a driving force in the ever-evolving technological landscape.

1.1.1 Mission

Beyond speeches, BIGmama's mission is to propose effective methodologies, action plans, and tools to:

- Solve problems with artificial intelligence tools.
- Put people at the heart of technology, i.e., the hybridization of AI.
- Make technology a common good, shareable, and accessible to the maximum number of people who can participate and contribute.

The company's commitment extends far beyond mere rhetoric; it is dedicated to developing practical solutions, methodologies, and actionable plans to leverage the power of artificial intelligence for problem-solving. BIGmama places a strong emphasis on ensuring that technology remains human-centric, fostering a harmonious integration and hybridization of AI with human intelligence.

Moreover, BIGmama recognizes the importance of democratizing access to technology, ensuring that it becomes a shared resource that can benefit society as a whole. The company aims to create an environment where the maximum number of individuals can actively participate and contribute to the advancement of technology.

Ultimately, BIGmama's mission is to position technology as a catalyst for freedom, empowering individuals and communities to unlock new possibilities and overcome barriers through innovative solutions.

1.1.2 Vision

"Standing on the Shoulders of Giants"

We find ourselves at a pivotal moment in human history, where the rapid advancements in artificial intelligence are poised to reshape our world in unprecedented ways. Groundbreaking technologies like ChatGPT, developed by industry giants, are at the forefront of this transformative wave, mobilizing vast resources to redefine the boundaries of what is possible.

However, BIGmama's approach is not to compete directly with these titans but rather to harness their innovations as a springboard for its own technological endeavors. By building upon the foundations laid by these industry leaders, BIGmama aims to leverage their advancements as a catalyst for its own visionary project.

Rather than reinventing the wheel, BIGmama recognizes the value in standing on the shoulders of giants, capitalizing on their groundbreaking work to propel its own unique solutions forward. This strategic approach not only allows for more efficient progress but also fosters an environment of collaboration and synergy within the broader technological ecosystem.

With a keen understanding of the rapidly evolving landscape, BIGmama remains agile and adaptable, poised to seize opportunities presented by the advancements of industry giants. By embracing a collaborative mindset and harnessing the collective wisdom of the technological community, BIGmama is well-positioned to contribute its own innovative solutions, shaping the future in alignment with its vision and mission.

1.1.3 Specificities

- **Data science starts with problematization:** At BIGmama, the approach to data science does not begin with data science itself but rather with reframing and problematization. The company's clients often arrive with a subject or topic rather than a clearly defined problem. BIGmama has developed a methodology to convert these topics into well-defined problems that data scientists can effectively tackle.

- **The data scientist is a "maverick":** BIGmama recognizes that the data scientist is a "maverick" who cannot be restrained. The company understands that it cannot impose specific tools or methods on data scientists when it comes to solving problems.

- **Data science a tool to solve problems:** BIGmama views data science not as an end in itself but as a means to solve problems. Often, the solution to their clients' problems lies outside the realm of data science tools.

- **Hybridization:** BIGmama believes that the future of AI lies in what is commonly referred to as Hybrid AI (Hybrid Artificial Intelligence). This approach encompasses a set of methodologies and techniques aimed at combining the potential of models with purely human knowledge. The company believes that this hybridization allows for putting humans at the heart of technological development and producing tools that are more efficient, easier to maintain, and less expensive.

1.1.4 Products

Historically, BIGmama is a startup specialized in the development of predictive applications for third parties. The company's approach is specific and based on the principles of hybridization between human and model capabilities. These specificities have led BIGmama to develop its own internal tools. These tools have evolved as the company's needs have changed, transitioning from Iko to Hyko. BIGmama is now willing to make these tools available to the broader technology ecosystem.

As a startup with a rich history in predictive application development, BIGmama has cultivated a unique approach that emphasizes the seamless integration of human expertise and model capabilities. This distinctive methodology, rooted in the principles of hybridization, has driven the company to develop its own suite of proprietary internal tools.

Over time, as BIGmama's requirements have evolved, these tools have undergone continuous refinement and adaptation. The company has transitioned from its initial Iko platform to the more advanced Hyko solution, ensuring that its tools remain aligned with its ever-changing needs and the dynamic technological landscape.

In recognition of the value these tools hold for the broader tech community, BIGmama has made the strategic decision to make them available to the larger ecosystem. By sharing its internally developed tools, the company aims to foster collaboration, knowledge exchange, and the collective advancement of technological solutions.