المدرسة الحسنية للأشغال العمومية
ECOLE HASSANIA DES TRAVAUX PUBLICS

LIKANA
COMMUNICATION

Internship Report: Work in Artificial Intelligence and Multilingualism

Benchmark and Comparative Analysis of Open-Source Language Models (LLMs)
for Multilingual Contexts (French – Arabic – Darija...)

Prepared by :

BOUMOUR SAFAA

Carried out from 01/08/2025 to 31/08/2025

Supervised by :

AITLHAJ RACHID

Institution: EHTP - Route d'El Jadida, Casablanca, Morocco, P.O.

Company: LIKANA COMMUNICATION SARL, located at 30 Rue Moulay Ahmed Loukili, Hassan, Rabat 10020, Morocco

07-09-2025

Acknowledgments

Before beginning this report, I would like to express my deepest gratitude to Allah, the Almighty, for granting me the opportunity and strength to successfully complete this internship. Without His guidance and blessings, this experience would not have been possible.

My heartfelt thanks go to Mr. AITLHAJ Rachid, my internship supervisor, for his constant availability, the trust he placed in me, and his valuable advice throughout this work. Your close guidance has been of great help and inspiration.

I would also like to thank my parents for their unwavering support. Your love and sacrifices have given me the strength to pursue my dreams and achieve my goals.

To my brothers, Oussama and Mouad, and my twin sister, Marwa, thank you for being a constant source of love and encouragement. You are my strength and my daily inspiration.

To my friends, your presence has been a pillar I could always rely on. Your friendship helped me overcome challenges and celebrate achievements. Thank you for your steadfast support and constant encouragement.

Finally, I would like to sincerely thank the jury members for taking the time to evaluate my work. I hope this report meets your expectations in terms of clarity and motivation. Your evaluation is both an honor and a recognition of the effort I have put into this project.

Summary

As part of my introductory internship, I participated in a project focused on designing a benchmark and performing a comparative analysis of open-source language models (LLMs) for multilingual contexts. This work was part of my exploration of the field of artificial intelligence, as it represented my first immersion into NLP and recent language model technologies.

The focus was on understanding the fundamental concepts of AI applied to language, while exploring several recent open-source models such as GPT-OSS, Falcon, and Mistral. The study also involved identifying and using freely accessible resources, such as Hugging Face and Google Colab, to conduct hands-on experiments.

These tests allowed for the evaluation of the quality of responses generated by the models and the analysis of their performance on various tasks, such as translation, logical reasoning, or structuring documents in JSON format. The comparative approach thus highlighted the strengths and limitations of each model and provided recommendations on their suitable applications.

This benchmark offers users clear guidance on selecting the most appropriate model depending on the nature of the prompt, while allowing me to progressively develop an understanding of modern language processing tools and to reflect on the practical uses of open-source LLMs in a rapidly evolving and promising field.

Keywords:

Artificial Intelligence, NLP, LLM, Benchmark, Multilingualism, Open Source.

List of Tables

List of Tables

1 Identification Sheet 9
2 Internship Planning and Progress 13

List of Figures

List of Figures

1 Simplified Organizational Chart of LIKANA COMMUNICATION 9
2 deep neural network 16
3 Recurrent Neural Networks 17
4 RNN and LSTM 18
5 Evolution of large language model 19
6 Operation of an LLM 24
7 Logo of the “Poe” Platform 26
8 Logo of the “HF” Platform 26
9 Software Logo 26
10 Logo of the ”Cb” Platform 27
11 Practical Test of Mistral 7B for Simple Factual Questions with Poe 28
12 Practical Test of GPT-OSS for Simple Factual Questions with Hugging Face 29
13 Practical Test of Falcon 3-7B for Simple Factual Questions with LM Studio 29

List of Acronyms

NLP :	Natural Language Processing
LLM :	Large Language Model
OSS :	Open Source Software
RAG :	Retrieval-Augmented Generation
CPU :	Central Processing Unit
GPU :	Graphics Processing Unit
API :	Application Programming Interface
JSON:	JavaScript Object Notation
Poe:	Platform for Open Exploration
Poe:	Platform for Open Exploration
RNN:	Recurrent Neural Networks
GAN:	Generative Adversarial Networks
CNN:	Convolutional Neural Networks
SVM:	Support Vector Machines
KNN:	K-Nearest Neighbors

Introduction

Large Language Models (LLMs) represent a major advancement in artificial intelligence today. Their ability to understand and generate natural language text paves the way for numerous applications: translation, chatbots, assisted writing, classification, and logical reasoning. They are thus an essential component of Natural Language Processing (NLP).

The emergence of open-source models such as Falcon, Mistral, or LLaMA has expanded the possibilities for experimentation and learning. More accessible than proprietary models, they provide opportunities to analyze performance, limitations, and use cases.

It is within this context that my introductory internship took place, aiming to explore the field of NLP and LLMs through three main aspects: creating a personal lexicon of key concepts, mapping and comparing major open-source models, and conducting a practical evaluation via a mini-benchmark across several tasks (translation, reasoning, JSON structuring).

To present this work, the report is structured as follows:

Chapter 1 describes the host organization, the project, and the methodology adopted.
Chapter 2 provides a literature review on NLP, LLMs, and recent models.
Chapter 3 details the specifications and the approach followed.
Chapter 4 presents the comparative study and experiments.
Chapter 5 highlights the contributions of the internship and possible avenues for improvement.

Contents

Acknowledgments 1

Summary 2

List of Tables 3

List of Figures 4

List of Acronyms 5

Introduction 6

1 GENERAL PROJECT CONTEXT 8
1.1 Presentation of the Host Organization 8
1.1.1 Service Area 8
1.1.2 Organizational Chart of Likana Communication 9
1.1.3 Identification Sheet 9
1.2 Business Areas 10
1.3 Project Presentation 11
1.3.1 Internship Problem Statement and Objectives 11
1.3.2 Methodology Adopted 11
1.3.3 Internship Planning and Execution 12
1.4 Chapter Conclusion 13

2 LITERATURE REVIEW 14
2.1 Introduction to Artificial Intelligence 14
2.2 Machine Learning 14
2.3 Deep learning 16
2.4 Recurrent Neural Networks 17
2.5 Large Language Models 19
2.6 Challenges of Multilingualism in NLP 20
2.7 Chapter Conclusion 20

3 SPECIFICATION OF REQUIREMENTS AND WORK APPROACH 21
3.1 Requirements Specifications 21
3.1.1 Educational and Technical Objectives of the Internship 21
3.1.2 Project Constraints and Limitations 22

3.2 Approach Followed . . . 22
3.2.1 Creation of a Personal Lexicon (NLP and LLM) . . . 22
3.2.2 Collection and Analysis of Information on Open Source Models . . . 24
3.2.3 Tools and Environments Used (Hugging Face, Google Colab, etc.) . . . 25
3.3 Chapter Conclusion . . . 27
4 COMPARATIVE STUDY AND EXPERIMENTATION 28
4.1 Model Evaluation Methodology . . . 28
4.2 Tasks Chosen for the Benchmark . . . 30
4.3 Comparative Analysis of Results . . . 31
4.3.1 Strengths of Each Model . . . 31
4.3.2 Identified Limitations and Constraints . . . 31
4.4 Usage Recommendations by Context . . . 32
4.5 Chapter Conclusion . . . 32
5 CONTRIBUTIONS AND PERSPECTIVES 33
5.1 Skills Acquired During the Internship . . . 33
5.2 Difficulties Encountered and Solutions Implemented . . . 33
5.3 Improvement Perspectives and Possible Extensions . . . 34
5.4 Chapter Conclusion . . . 34
6 GENERAL CONCLUSION 35
6.1 Internship Summary . . . 35
6.2 Personal and Professional Learnings . . . 35
6.3 Future Perspectives . . . 36
7 Appendices 36

1 GENERAL PROJECT CONTEXT

1.1 Presentation of the Host Organization

1.1.1 Service Area

LIKANA COMMUNICATION SARL is a Moroccan company located at 30, Rue Moulay Ahmed Loukili, Hassan district, Rabat (10020), Morocco. It operates primarily in the fields of information technology, communication, and digital innovation. With a diverse expertise, the company serves both local and international clients, providing tailored solutions adapted to their needs in digital transformation and artificial intelligence integration.

1.1.2 Organizational Chart of Likana Communication

LIKANA COMMUNICATION is built on a multidisciplinary team of experienced professionals in computer science, communication, and new technologies, all passionate about innovation and artificial intelligence. The company relies on strategic supervision provided by experts in AI and digital transformation, ensuring the quality of the solutions offered, continuous service development, and client satisfaction. This structure fosters effective collaboration between technical, creative, and strategic teams, enabling LIKANA COMMUNICATION to design comprehensive and innovative solutions tailored to the specific needs of its clients.

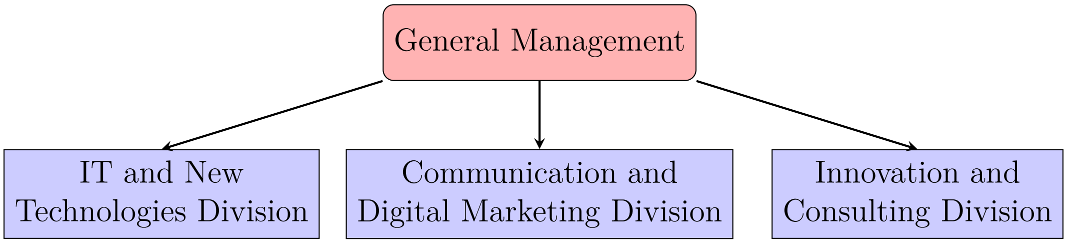

General Management
IT and New Technologies Division
Communication and Digital Marketing Division
Innovation and Consulting Division

Figure 1: Simplified Organizational Chart of LIKANA COMMUNICATION

1.1.3 Identification Sheet

Item	**Information**
Company Name	LIKANA COMMUNICATION SARL
Address	30, Rue Moulay Ahmed Loukili, Hassan Rabat 10020 – Maroc
Activity Sector	Information Technology, Communication, and New Technologies
Main Services	Web and Mobile Development, Cloud, Cybersecurity, Digital Marketing, AI
Vision	To become a leading player in Morocco and Africa in AI integration

Table 1: Identification Sheet

1.2 Business Areas

LIKANA COMMUNICATION operates in several complementary domains aimed at supporting its clients in digital transformation and the integration of intelligent solutions:

Information Technology and New Technologies

- Development of software and web/mobile applications incorporating artificial intelligence to enhance efficiency and user experience.
- Integration of intelligent IT solutions and ERP systems tailored to business needs.
- Secure hosting, cloud services, and advanced cybersecurity to ensure data protection.
- Proactive maintenance and intelligent technical support to ensure business continuity.

Communication and Digital Marketing

- Development of 360° communication strategies based on data and AI for optimal visibility.
- Creation of digital content: design, video, animation, and motion graphics.
- Social media management and engagement using automation tools.
- SEO/SEA optimization powered by artificial intelligence.
- Development of a customized visual identity and digital branding.

Innovation and Consulting

- Support for digital transformation with the integration of intelligent solutions.
- Conducting studies and strategies in emerging technologies.
- Specialized training in communication, digital marketing, and AI.
- Intelligent automation and digitalization of business processes to improve productivity.

Strategic Vision:

LIKANA COMMUNICATION places artificial intelligence at the core of its approach to transform technology into a sustainable driver of growth and innovation. The company aspires to become a leading player in Morocco and Africa, providing intelligent solutions that enhance client performance and competitiveness.

1.3 Project Presentation

1.3.1 Internship Problem Statement and Objectives

The rapid rise of Large Language Models (LLMs) raises a twofold challenge. On one hand, their complexity makes them difficult to understand for new users seeking to familiarize themselves with the fundamental concepts of Natural Language Processing (NLP). On the other hand, the diversity of available open-source models, each with its own strengths and limitations, raises the question of comparison and relevance depending on the context of use, particularly to determine which model is most suitable for a given type of prompt.

Within this framework, this introductory internship set several objectives:

To create a simple and precise document gathering the essential concepts of NLP and LLMs, reformulated in my own words, to facilitate comprehension and dissemination of these notions.

To develop a comparative table of the main recent open-source models, highlighting their characteristics, accessibility, and use cases.

To practically test several models on various tasks (translation, logical reasoning, JSON docu ment structuring, etc.) to evaluate their relevance and provide appropriate recommendations.

Thus, this work aims to combine a theoretical approach (concept assimilation) and a practical approach (experimentation and benchmarking) to develop a critical and applied understanding of open-source LLMs.

1.3.2 Methodology Adopted

The methodology followed during this internship is based on a progressive and iterative approach, ensuring both theoretical understanding and practical application of the acquired knowledge. It is structured around the following steps:

Literature Review and Documentation
Reading and summarizing scientific articles, tutorials, official documentation, and existing reports on Natural Language Processing (NLP) and Language Models (LLMs).
Identifying fundamental concepts and current trends in the field.

Comparative Analysis of Models

Selecting 10 open-source language models (among the most recent and widely used).
Developing a comparative table based on several criteria: model size, training data, licenses, accessibility, performance, and use cases.

Practical Experimentation
Installing and testing different models locally and/or on specialized platforms
Evaluating models on specific tasks (JSON data structuring, text generation, summarization, translation, etc.).

Evaluation and Synthesis
Comparing the results obtained according to the predefined objectives.
Conducting a critical analysis highlighting the advantages, limitations, and potential areas for improvement.

1.3.3 Internship Planning and Execution

Project planning is a crucial step in project management. It involves defining in detail the tasks to be completed, the required resources, and the deadlines to be met in order to successfully carry out the project.

Phase / Activity / Task	Duration(days)	Start Date	End Date
Phase 1 : Integration and Familiarization	4	01/08/2025	04/08/2025
Familiarization with technical environment	2	01/08/2025	02/08/2025
Definition of internship objectives in consultation with the supervisor	1	03/08/2025	03/08/2025
Initial bibliographic research	1	04/08/2025	04/08/2025
Phase 2 : Study and Analysis	10	05/08/2025	14/08/2025
Deepening knowledge of NLP concepts and recent models	4	05/08/2025	08/08/2025
Listing and selecting models to compare	3	09/08/2025	11/08/2025
Designing the comparative table (license, number of parameters, etc.)			
Phase 3 : Development and Experimentation	7	15/08/2025	21/08/2025
Installation and configuration of testing environments	2	15/08/2025	16/08/2025
Practical testing of models on different tasks	3	17/08/2025	19/08/2025
Collection and organization of obtained results	2	20/08/2025	21/08/2025
Phase 4 : Analysis and Evaluation	10	22/08/2025	31/08/2025
Critical analysis of experimental results (quality rated out of 5)	4	22/08/2025	25/08/2025
Production of a JSON file containing the prompts used	2	26/08/2025	27/08/2025
Final synthesis and recommendations	4	28/08/2025	31/08/2025

Table 2: Internship Planning and Progress

1.4 Chapter Conclusion

In conclusion, this chapter introduces my project at Likana Communication. It outlines the objective of analyzing language models (LLMs) to identify the most suitable model for a given prompt, summarizes the methodology, and presents the initial project plan. These foundations set the stage for a more detailed examination in the following chapters.

Contents

2 LITERATURE REVIEW

2.1 Introduction to Artificial Intelligence

Artificial intelligence (AI) has become a key area of modern technology, aiming to create systems capable of reproducing or mimicking human intelligence. These systems enable machines to perform complex tasks such as speech recognition, computer vision, and natural language processing. A notable subfield of AI is generative artificial intelligence, which is distinguished by its ability to create new content and ideas, including conversations, stories, images, videos, and music. Generative AI represents a significant advancement, allowing the learning of various languages—including human and programming languages—as well as complex disciplines such as art, chemistry, and biology. It reuses training data to solve new problems. For example, it can learn English vocabulary and generate a poem from the words it knows. Companies can leverage generative AI for a variety of applications, including chatbots, multimedia creation, and product development and design.

Applications of generative AI, such as ChatGPT, have attracted great interest and sparked the imagination. Generative AI can respond naturally to human conversations and serves as a valuable tool for customer service and workflow personalization. Advances in AI have led to significant improvements in chatbots. Initially, chatbots were task-oriented (declarative), following predefined scripts with simple rules to determine their responses. These systems were limited in their ability to handle variations in user queries.

However, the emergence of Large Language Models (LLMs) has marked a major breakthrough in natural language understanding and generation. These models, often used as the engines behind chatbots and virtual assistants, can produce coherent, contextual, and situation-appropriate texts. They employ advanced machine learning and deep learning techniques to reason, generate natural responses, and process complex information.

2.2 Machine Learning

This evolution is made possible by the fundamental principles of machine learning, which rely on the ability of algorithms to detect patterns and structures within data and then adjust their parameters accordingly. Rather than following precise instructions, computers learn from the data they process.

Machine learning is divided into two main categories:

Unsupervised Learning : This method focuses on identifying intrinsic patterns and structures in data without using pre-existing labels. The main goal is to discover natural groupings, similarities, or hidden relationships in unlabeled data. Commonly used algorithms include:
→ Clustering algorithms : Identify groups of similar records and label them based on the group to which they belong.
→ Association algorithms : Discover patterns and relationships in the data, identifying “if/then” rules known as association rules.

Supervised Learning : This method involves providing a training dataset containing labeled examples, i.e., data associated with labels or categories. The goal is to enable the model to make accurate predictions based on these examples. Commonly used algorithms include:
→ Linear Regression : Used to predict continuous values by establishing a linear relationship between features and the target variable.
→ Logistic Regression : Suitable for binary classification, where the target variable is discrete (e.g., yes/no). It models the probability of belonging to a class using a logistic function.
→Decision Trees :Models that create a tree structure based on features to arrive at predictions.
→ Random Forest : An ensemble of decision trees that aggregates the results of multiple trees to improve accuracy and reduce overfitting.
→ Support Vector Machines (SVM) :Algorithms that find a hyperplane in a multidimensional space to best separate different classes of data.
→ K-Nearest Neighbors (KNN) :Assigns a label to new data based on the majority of labels of its nearest neighbors in the feature space.
→ Naive Bayes :Based on Bayes’ theorem, commonly used for classification tasks.
→ Artificial Neural Networks (ANN) :Mathematical models inspired by the human brain, composed of layers of interconnected neurons, powerful for solving complex problems.
After exploring the fundamentals of supervised and unsupervised learning, as well as the various algorithms that underpin them, we turn to a more advanced aspect of artificial intelligence: deep learning, also known as hierarchical or deep neural network learning.

2.3 Deep learning

Deep learning represents a major advancement in the field of AI, enabling machines to acquire a hierarchical and abstract understanding of data. This concept is based on the use of artificial neural networks composed of multiple layers. These deep networks aim to mimic the behavior of the human brain on a large scale by learning from vast datasets. In these deep neural networks, data passes through multiple processing layers. At each layer, the extracted features become increasingly abstract. These networks are capable of learning complex patterns and performing tasks such as image recognition and machine translation with remarkable accuracy.

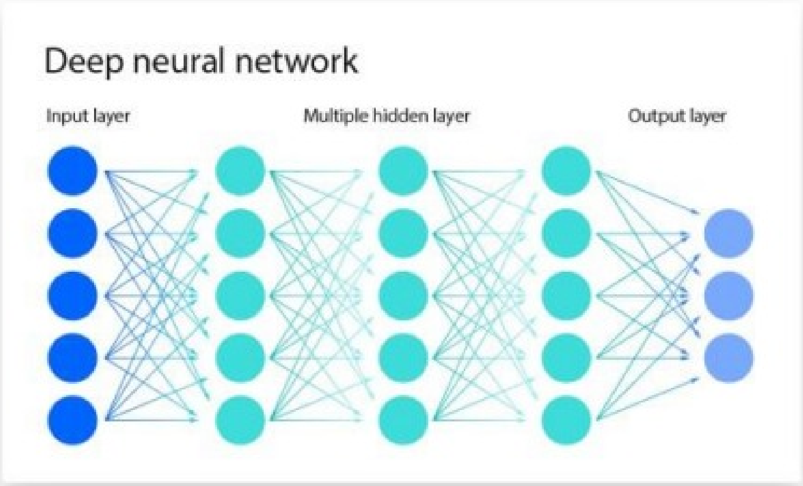


Figure 2: deep neural network

Let us now examine in detail the different types of neural networks used in deep learning and the specific tasks they are designed for:

Feedforward Neural Networks: These are the simplest and most commonly used neural networks. Information flows in a single direction, from input to output, without feedback loops. Each neuron in a layer is connected to all neurons in the next layer. Feedforward neural networks are used for classification and regression tasks, where the goal is to predict a continuous value or a class.

Convolutional Neural Networks (CNNs): :CNNs are specifically designed for processing images and spatial data. They exploit local features of images using convolutional layers and pooling layers. Convolutional layers apply filters to extract local patterns, while pooling layers reduce the size of extracted features while preserving essential information. CNNs are widely used in image classification, object detection, semantic segmentation, and more.

Generative Adversarial Networks (GANs) : GANs consist of two competing neural networks: a generator and a discriminator. The generator produces new synthetic data, while the discriminator attempts to distinguish real data from synthetic data. This competitive training process allows GANs to generate realistic data. GANs are used in areas such as image generation, style transfer, and text generation.

Recurrent Neural Networks (RNNs) : RNNs are designed to handle sequential data, such as time series or sentences. Unlike feedforward networks, RNNs use feedback loops that allow them to maintain an internal memory. This memory enables them to consider previous context when processing new information. RNNs are widely used in tasks such as machine translation, text generation, speech recognition, and more.

2.4 Recurrent Neural Networks

RNNs have the unique ability to retain information from previous inputs through their internal memory mechanism. This feature makes them particularly well-suited for tasks where context and the order of data are crucial, such as language processing, time series analysis, and speech recognition.

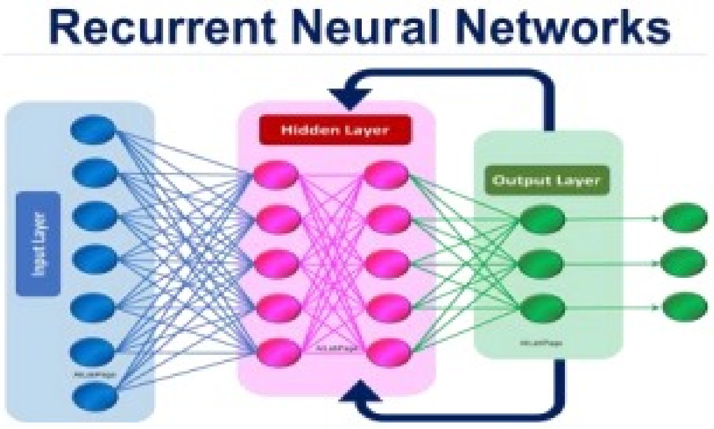


Figure 3: Recurrent Neural Networks

However, RNNs face significant limitations, notably the vanishing gradient problem. This issue occurs when the network struggles to learn and retain information from long sequences, where the influence of earlier inputs decreases exponentially as the sequence progresses, making it difficult to capture long-term dependencies.

To address these limitations, a new architecture known as Long Short-Term Memory (LSTM) was developed. LSTMs include cells in the hidden layers of the neural network, composed of three gates: an input gate, a forget gate, and an output gate. These gates control the flow of information necessary to predict the network's output, allowing for longer memory retention and efficient learning from extended data sequences. Although LSTMs represent a significant improvement over RNNs traditionals, they are not without their own limitations. A major issue is their computational complexity and inefficiency, mainly due to the sequential nature of their processing. LSTMs struggle with very long sequences, as the time and computational resources required increase linearly with sequence length. Moreover, LSTMs still face challenges related to the vanishing gradient problem, although to a lesser extent than RNNs, particularly in very deep networks or extremely long sequences. This makes them less practical for certain real-world applications that involve large datasets or require fast processing.

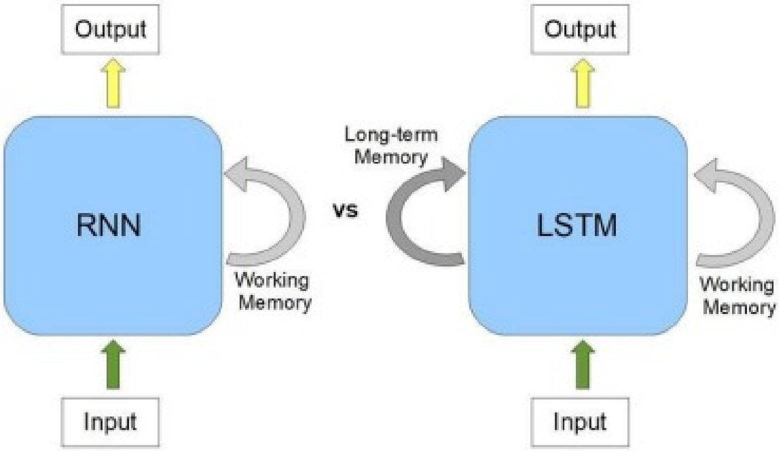


Figure 4: RNN and LSTM

However, the continuous pace of AI advancements has led to the development of an even more powerful architecture: the Transformer. Introduced in the seminal paper "Attention Is All You Need" in 2017, Transformers marked a significant departure from the recurrent structure of LSTMs.

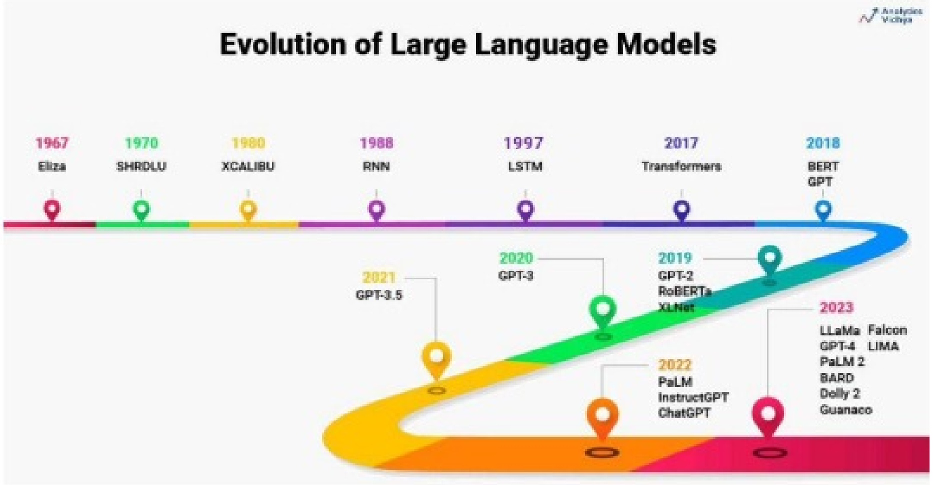


Figure 5: Evolution of large language model

Instead of processing data sequentially, Transformers use a mechanism called self-attention to handle entire sequences of data in parallel. This architectural change addresses the limitations of LSTMs in managing very long sequences and enables much faster computation, which is particularly advantageous given the growing size of datasets in NLP tasks.
The self-attention mechanism in Transformers allows each position in the input sequence to directly consider all other positions, unlike LSTMs, which process data step by step. This enables the model to capture complex and long-range relationships in the data more effectively.
Furthermore, Transformers completely eliminate the need for recurrence, resulting in significant improvements in training efficiency and scalability.

2.5 Large Language Models

Advances enabled by Transformers have led to the emergence of Large Language Models (LLMs), representing a major evolution in the field of Natural Language Processing (NLP). These deep learning models are trained on massive volumes of textual data and rely on neural architectures capable of capturing complex relationships between words and sentences.
LLMs have the ability to interpret, understand, and generate natural language text with remarkable accuracy. They can answer questions, summarize documents, translate texts, produce original content, and assist users in complex writing or programming tasks.
Today, LLMs form the core of many practical applications, such as intelligent chatbots, virtual assistants, productivity tools, and scientific research support systems. Their impact is significant, as they enable more natural and intuitive human-machine interactions.
However, despite their power, these models have certain limitations, including biases related to training data, high computational resource consumption, and risks of inappropriate use. These

challenges highlight the need for responsible and ethical deployment of LLMs.

2.6 Challenges of Multilingualism in NLP

Multilingualism is a major challenge in the field of Natural Language Processing (NLP). While English dominates most training corpora, many other languages remain underrepresented, creating an imbalance in system quality and accuracy.

A multilingual model must be able to understand, interpret, and generate text in different languages, sometimes linguistically very distant from one another. This involves several challenges:

Linguistic diversity : Languages differ in grammar, syntax, morphology, and alphabets, making universality difficult to achieve.

Limited resources : Some languages have few digital resources, reducing the quality of models that support them.

Translation and fairness : It is essential to avoid biases that favor certain languages over others, ensuring equitable access to NLP technologies.

Cultural and contextual adaptation : Beyond literal translation, a model must account for cultural nuances, idiomatic expressions, and language registers.

Recent advances, particularly through multilingual LLMs, have significantly improved language coverage by incorporating cross-lingual transfer techniques. Nevertheless, multilingualism remains an evolving field, with the goal of making NLP technologies truly inclusive and universal.

2.7 Chapter Conclusion

In summary, this chapter provided an in-depth literature review, allowing us to gain a comprehensive understanding of the key concepts of our project. This knowledge will be crucial for guiding and informing our next steps, providing a solid foundation for the effective implementation of our project.

Contents

3 SPECIFICATION OF REQUIREMENTS AND WORK APPROACH

3.1 Requirements Specifications

3.1.1 Educational and Technical Objectives of the Internship

The objectives set for this internship can be divided into two complementary categories: educational and technical.

The educational objectives of this internship were primarily aimed at strengthening my knowledge and skills in the field of artificial intelligence applied to language processing. They can be summarized as follows:

- Deepen the understanding of concepts related to *Natural Language Processing* (NLP) and *Large Language Models* (LLM).
- Develop the ability to synthesize and simplify information by creating a personal glossary of fundamental concepts.
- Learn to critically analyze and compare different language models with well-argued reasoning.
- Improve technical documentation skills: researching reliable sources, structuring, and formatting results. autonomy in solving problems related to emerging technologies.

The technical objectives of this internship focused on the practical implementation of tools and methods to better understand and evaluate language models. These include:

- Create a detailed glossary of key NLP and LLM concepts, including definitions, simplified explanations, and useful references.
- Develop a comparative mapping of the main recent open-source LLMs, specifying their characteristics (size, license, supported languages, use cases, accessibility).
- Set up a practical mini-benchmark by testing several models using a variety of prompts (reasoning, summarization, translation, JSON structuring, etc.).
- Experiment with different platforms for accessing models (Hugging Face, Poe, Google Colab, LM Studio).

- Design an evaluation process based on specific criteria: response quality, speed, coherence, and detection of possible hallucinations.

3.1.2 Project Constraints and Limitations

During this project, several constraints and limitations were encountered, which influenced the execution of the work and the scope of the results obtained:

Hardware limitations : The use of large-scale models requires significant hardware resources, particularly in terms of GPU and RAM. The available work environments sometimes had insufficient capacity, which limited the scope of experimentation and the smoothness of testing.

Restricted access to certain models : Some state-of-the-art models are proprietary or subject to restrictive licenses, making their use difficult or even impossible in an academic or experimental setting. This constraint limited the comparison between open-source solutions and certain commercial alternatives.

Incomplete availability of recent versions : In some cases, only older versions of models were accessible. This limitation may have affected the relevance of the results, as the most recent iterations often include significant improvements in performance and optimization.

Performance variability across platforms : Tests conducted on different platforms (such as Hugging Face, Google Colab, or LM Studio) revealed heterogeneous performance. This variability is related to the resources allocated, the underlying infrastructure, and the specific execution conditions of each environment.

Despite these constraints, this work provided a solid evaluation of LLMs and highlighted the most relevant model for each type of prompt.

3.2 Approach Followed

3.2.1 Creation of a Personal Lexicon (NLP and LLM)

As part of this internship, I began by creating a **personal lexicon** compiling the fundamental concepts of Natural Language Processing (NLP) and recent language models (LLMs).

To do this, I conducted research using **various sources** such as scientific articles, online tutorials, explanatory videos, and specialized forums. I made sure to **cross-check information** to avoid relying on a single definition and to ensure a reliable understanding of the concepts. Each term was then **reformulated in my own words**, which allowed me to confirm my comprehension and avoid simple copy-pasting.

or each term, I added at least one **external resource** (article or educational video) to facilitate further study. I also noted points that seemed ambiguous or difficult, so I could discuss them with my supervisor and benefit from their guidance.

The lexicon specifically covers the following concepts:

- NLP (Natural Language Processing)
- LLM (Large Language Model)
- Pre-trained model
- Fine-tuning
- Prompt / Prompting
- Token
- Number of parameters in a model
- Transformers and attention mechanism (simplified level)
- Embeddings
- Open-source vs proprietary models
- Examples of open-source models (Mistral, LLaMA, Falcon, etc.)

In addition to these definitions, I studied the **general functioning of LLMs**. These models are based on the Transformer architecture and notably use the attention mechanism to effectively process the context of a sentence. Practically, an LLM predicts the continuation of a text based on the preceding words. Its performance primarily depends on the number of parameters, the size of the training data, and specialization techniques such as fine-tuning. This understanding provided an essential foundation for the subsequent stages of the project.

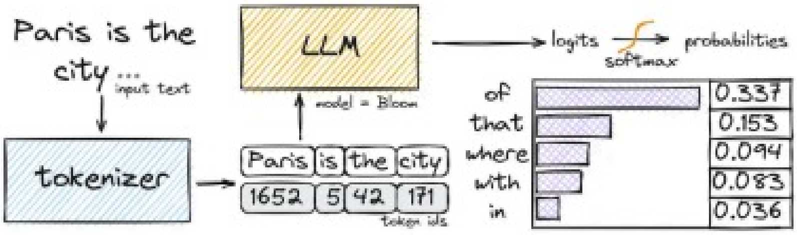


Figure 6: Operation of an LLM

Contributions of this Stage

- A better theoretical understanding of the foundational concepts necessary for the subsequent phases of the project.
- A clear working reference that served as a constant guide during the comparison and benchmarking of models.
- The development of skills in technical simplification and knowledge structuring.

3.2.2 Collection and Analysis of Information on Open Source Models

As part of this internship, an important step was **mapping and analyzing the main open-source LLMs**. The goal was to create a **comparative table** of the most comprehensive, widely used, and recent models in order to better understand their functionality, accessibility, and use cases.

Methodology Adopted

1. **Selection of models to study:** I identified ten recent and relevant models: Mistral 7B, LLaMA 3, Gemma, Falcon 7B/180B, DeepSeek LLM/DeepSeek Coder, Zephyr, Yi 34B, Command R+, Phi-2, and TinyLLaMA 1.1B.
2. **Information collection:** For each model, I searched for reliable data on:
   - The creating organization and country of origin
   - Release date
   - Model size (number of parameters)
   - Supported languages
   - Type of architecture (decoder-only, encoder-decoder, etc.)

- Access methods (Hugging Face, Poe, LM Studio, etc.)
- License (permissive, restricted, commercial, etc.)
- Known strengths and limitations
- Recommended use cases

These details were gathered from **various sources** : scientific articles, official documentation, GitHub repositories, specialized blogs, and model testing platforms.

3. **Analysis and synthesis :**The collected data were **compared and synthesized** into a single table for each model. This step highlighted the differences and similarities, as well as identifying the models best suited for specific needs (size, accessibility, language, task type).

Contributions of this Stage

- Acquisition of a global view of the current open-source LLM landscape.
- Identification of the most relevant models for different types of tasks and prompts.
- Highlighting practical limitations and access constraints for certain recent versions.
- Preparation of the foundation necessary for the next phase of practical testing and mini-benchmarking.

3.2.3 Tools and Environments Used (Hugging Face, Google Colab, etc.)

To test the LLMs and automate the execution of prompts, I used several platforms and environments suited to the needs of this project.

nitially, I had planned to use only Hugging Face. However, I found that some models I wanted to work with—particularly the latest versions of Mistral 7B, LLaMA 3.1 8B, Falcon 7B, DeepSeek-R1, GPT-OSS, Zephyr 7B, and Gemma—were not all deployed on this platform.
To access recent models and missing versions, it was necessary to explore other platforms:

- **Poe :** : I used this platform to efficiently test Mistral 7B, LLaMA 3.1 8B, DeepSeek-R1, and Gemma. Poe proved useful for accessing recent models not available on Hugging Face.

Figure 7: Logo of the “Poe” Platform

- **Hugging Face :** I used this platform for GPT-OSS and Zephyr 7B (beta), which were available and easily testable.

Figure 8: Logo of the “HF” Platform

- **LM Studio :** For Falcon 7B, which was not accessible online, I installed the model locally via LM Studio.

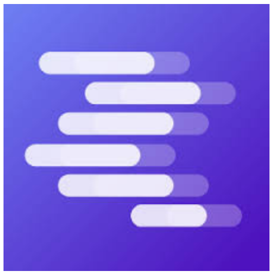


Figure 9: Software Logo

- **Google Colab :** : I developed a script to automatically run prompts on all relevant models. However, this automation does not yet work for all models due to certain technical constraints specific to each platform.

Figure 10: Logo of the "Cb" Platform

Contributions of this Stage This diversity of tools and platforms allowed me to: :

- Access a greater number of models, including the latest versions.
- Experiment with different testing and automation environments.
- Better understand the technical constraints related to each platform (compatibility, performance, access limitations).

3.3 Chapter Conclusion

This chapter clarified the project requirements and presented the methodology followed, including the personal lexicon, analysis of open-source models, and the tools (platforms and software) used. It sets the stage for the next chapter, dedicated to the comparative study and experimentation of LLMs.

Contents

4 COMPARATIVE STUDY AND EXPERIMENTATION

4.1 Model Evaluation Methodology

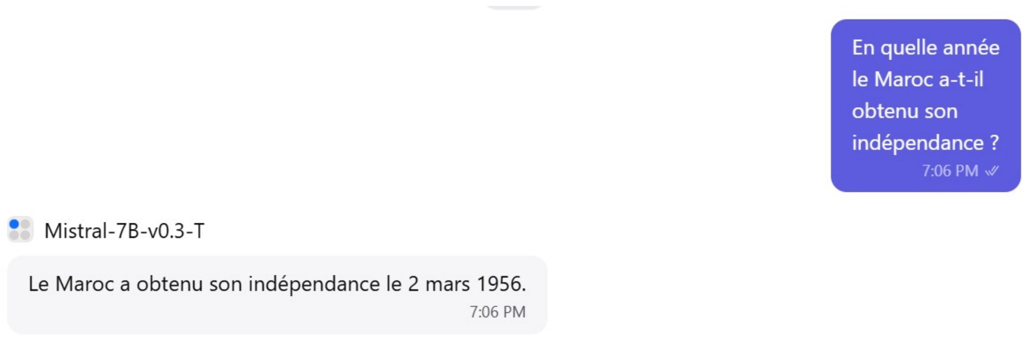
To objectively compare the performance of different Large Language Models (LLMs), a structured evaluation methodology was adopted. This methodology is based on several principles:

Model Selection : A representative set of open-source and accessible models was chosen, including: Mistral 7B, LLaMA 3.1 (8B and 70B), Falcon 7B, DeepSeek-R1, GPT-OSS-120B, Zephyr 7B, and Gemma 27B. The selection was based on their availability (via Poe, Hugging Face, or LM Studio), diversity in parameter sizes, and relevance for practical use cases.

Definition of Evaluation Tasks : The models were tested on a multi-task mini-benchmark, covering:
Simple factual questions, Logical reasoning, Summarization tasks (French and Arabic), Translation (French → Moroccan Darija), Specialized tasks (e.g., writing a formal email) and JSON structuring (technical format suitable for use)
These tasks allow evaluation of comprehension, structured generation ability, multilingual capacity, and reliability in applied contexts.

Testing Protocol:

same prompts were submitted to each model to ensure fairness.

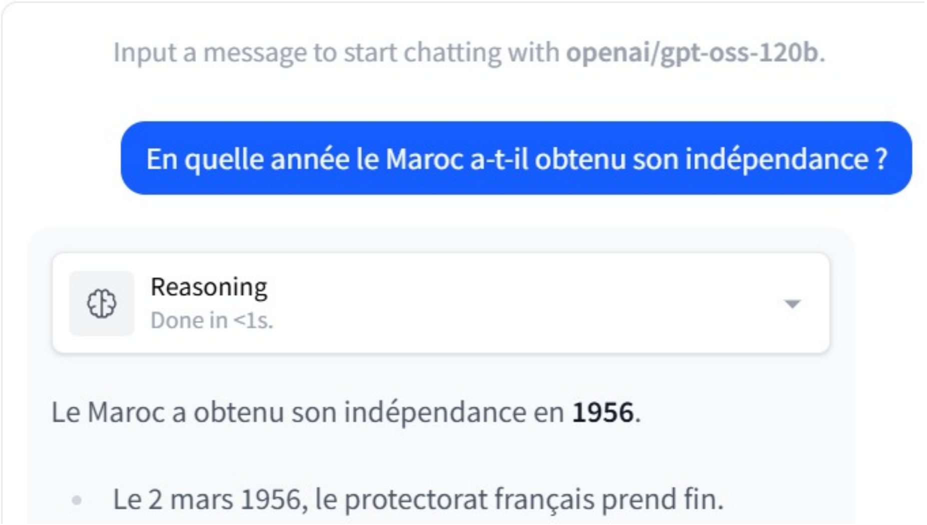


En quelle année le Maroc a-t-il obtenu son indépendance ?
7:06 PM

Mistral-7B-v0.3-T

Le Maroc a obtenu son indépendance le 2 mars 1956.
7:06 PM

Figure 11: Practical Test of Mistral 7B for Simple Factual Questions with Poe



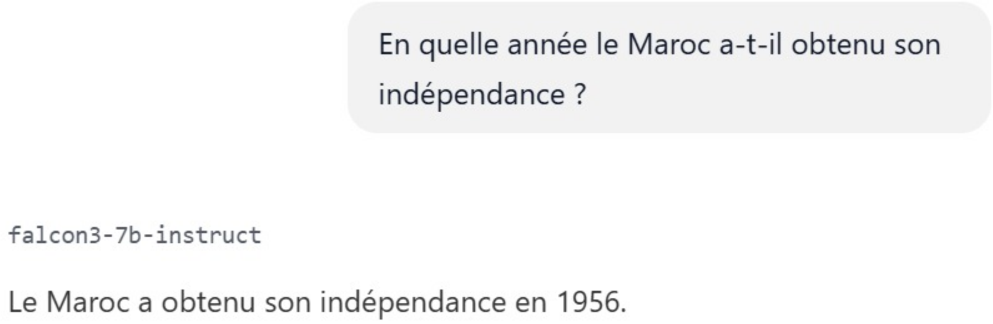
Input a message to start chatting with **openai/gpt-oss-120b**.

En quelle année le Maroc a-t-il obtenu son indépendance ?

Le Maroc a obtenu son indépendance en **1956**.

- Le 2 mars 1956, le protectorat français prend fin.

Figure 12: Practical Test of GPT-OSS for Simple Factual Questions with Hugging Face



En quelle année le Maroc a-t-il obtenu son indépendance ?

falcon3-7b-instruct

Le Maroc a obtenu son indépendance en 1956.

Figure 13: Practical Test of Falcon 3-7B for Simple Factual Questions with LM Studio

- Responses were evaluated according to multiple criteria: accuracy and adherence to the prompt, clarity and linguistic quality, generation speed, presence of errors or hallucinations, and direct usability of output (particularly for JSON or translation).

Scoring and Analysis : Each response was rated on a scale from 1 to 5, with qualitative remarks (speed, errors, coherence).
Results were then synthesized to highlight:

- Strengths of each model for each task type,
- Specific limitations (hallucinations, slowness, poor structuring)

Tools Used :

- Poe and Hugging Face were used to test models directly online.
- LM Studio allowed local execution of some models (Falcon, Zephyr).
- Results were consolidated in a comparative table and analyzed both qualitatively and quantitatively.

4.2 Tasks Chosen for the Benchmark

To evaluate the versatility and robustness of the selected models, several task categories were chosen. They cover everyday, educational, and technical use cases to provide a comprehensive view of model performance:

- **Simple Factual Questions**
  Objective : Test the models' ability to provide precise answers to direct and well-known questions (e.g., "In which year did Morocco gain its independence?"). Observed indicators: Accuracy, conciseness, response speed.

- **Elementary Logical Reasoning**
  Objective : Evaluate step-by-step reasoning abilities on simple arithmetic or logic problems. Example: "If Ahmed has two apples and gives one to Fatima, how many does he have left?" Observed indicators: Clarity of explanations, absence of errors, pedagogical style.

- **Text Summarization (French and Arabic)**
  Objective : Test the ability to synthesize content while preserving essential information.
  Exemple :Summarize a paragraph in French and then in Arabic.
  Observed indicators: Fidelity to the original text, clarity, linguistic quality.

- **Translation (French → Moroccan Darija)**
  Objective : Assess multilingual proficiency and accuracy in a local language with high variability.
  Observed indicators : Naturalness of output, fluency, absence of mixing with Classical Arabic.

- **Specialized Tasks (Formal Writing)**
  Objective :Test the ability to produce text appropriate for professional contexts, such as drafting a formal email.
  Observed indicators: Structure, politeness, adaptability, and immediate usability of the content.

- **JSON Structuring**
  Objective : Measure precision and rigor in generating structured data directly usable by computer systems.
  Example: Convert an invoice into a well-formatted JSON document.
  Observed indicators: Format compliance, field consistency, adherence to standards (ISO dates, clear keys).

4.3 Comparative Analysis of Results

The practical evaluation showed that the models' performances vary significantly depending on the nature of the tasks. To enable a fair comparison, several criteria were considered: accuracy, generation speed, linguistic quality, presence of errors or hallucinations, and direct usability of the outputs.

4.3.1 Strengths of Each Model

- **GPT-OSS** : The most reliable and fastest model, with excellent data structuring (clean JSON, ISO-formatted dates); ideal for technical use cases.

- **DeepSeek-R1** : Outstanding in logical reasoning and translation tasks, providing detailed explanations and a pedagogical style.

- **LLaMA 3.1 (8B)** : Very good at summarization and text writing, producing fluent and well-structured outputs.

- **Mistral 7B** : Fast, with solid performance on simple factual tasks and basic text generation.

- **Falcon 7B** :Performs adequately across all tasks but tends to be overly concise; suitable for simple needs.

- **Zephyr 7B** : Lightweight and easily executable locally, making it useful for rapid prototyping.

- **Gemma 27B** :High linguistic quality and rich contextual understanding, though slower and more resource-intensive.

4.3.2 Identified Limitations and Constraints

- **GPT-OSS** : Highly demanding in terms of local resources (GPU/RAM); full access is restricted without Fireworks.

- **DeepSeek-R1** : Excellent reasoning capabilities but tends to be verbose and requires longer generation times.
- **LLaMA 3.1 (8B)** :Efficient but relatively slow for certain tasks.
- **Mistral 7B** : Tendency to simplify or invent minor details (minor hallucinations).
- **Falcon 7B** :Slow response time and lack of depth in explanations.
- **Zephyr 7B** :Unstable, frequent hallucinations, and loss of coherence on complex tasks.
- **Gemma 27B** : Difficult to run locally; JSON structuring sometimes imperfect.

4.4 Usage Recommendations by Context

The analysis of results makes it possible to derive tailored recommendations for different scenarios:

- **Technical use and automation** (data structuring, JSON, precise tasks): GPT-OSS is the most suitable.
- **Educational use and step-by-step reasoning** : **DeepSeek-R1** is recommended for its detailed explanations.
- **Summarization, writing, and high-quality text generation** : **LLaMA 3.1 (8B)** is the best choice.
- **Fast and lightweight use** : **Mistral 7B**is a practical alternative for simple tasks.
- **Prototyping or local testing on limited hardware** : **Zephyr 7B** can be used despite its limitations.
- **Contexts requiring rich linguistic and contextual understanding** : **Gemma 27B**remains an interesting but resource-intensive option.

4.5 Chapter Conclusion

The experimentation revealed that each model has specific strengths and limitations. No model outperforms all others across every task, but several stand out in particular contexts:

- **GPT-OSS**proves to be the most reliable solution for technical and structured applications.
- **DeepSeek-R1** adds real value for educational and reasoning-based use cases.

- **LLaMA 3.1 (8B)**stands as a benchmark for high-quality writing and summarization tasks.

Conversely, lighter models such as **Mistral 7B** and **Zephyr 7B** remain relevant for everyday use or deployment on resource-limited machines, while **Gemma 27B** serves as a compromise between linguistic power and technical constraints.

Ultimately, the choice of model strongly depends on the intended context: *Ultimately, the choice of model strongly depends on the intended context: technical use demands robustness and precision, whereas educational or writing-oriented use favors clarity and explanatory capability.*

5 CONTRIBUTIONS AND PERSPECTIVES

5.1 Skills Acquired During the Internship

This internship allowed me to acquire several technical and methodological skills:

- **Mastery of key NLP and LLM concepts** : Understanding of fundamental notions such as pre-trained models, fine-tuning, prompting, embeddings, and the differences between open-source and proprietary models — as this was my first hands-on experience in this field.

- **Comparative analysis of models** : Ability to create concise technical summaries of recent models (Mistral, LLaMA, Falcon, etc.), identifying their strengths, limitations, and areas of application.

- **Practical evaluation of LLMs** : Implementation of a testing protocol (fixed prompts, evaluation criteria, scoring), and practical experimentation across diverse tasks (reasoning, translation, summarization, JSON structuring).

- **Research methodology** : Development of skills in collecting reliable information (scientific articles, videos, forums), cross-verifying sources, and reformulating concepts clearly to make them accessible.

- **Transversal skills** : Improvement in organization, structured report writing, and the creation of usable deliverables such as comparative tables and analytical summaries.

5.2 Difficulties Encountered and Solutions Implemented

During my internship, several challenges were encountered, and appropriate solutions were implemented:

- **Access to the latest versions of LLMs** : Some recent versions of language models were not yet deployed.
  Solution : I explored other platforms, such as Poe, which offered free access to certain new versions.

- **Hardware limitations** : Local execution of models was restricted due to insufficient GPU power and RAM on my computer.
  Solution : I used cloud-based platforms to run the models efficiently.

- **Automation of prompts in Python:** :I aimed to automate prompt generation through a Python script on *Google Colab*, but some models did not respond properly to the requests.
  Solution : I performed manual prompt generation for those specific models.

5.3 Improvement Perspectives and Possible Extensions

This work can be extended and enhanced through several directions:

- **Local optimization** :Setting up appropriate servers (GPU/TPU) to test large models under real-world conditions.

- **Fine-tuning** : Training certain models on specialized datasets (technical, legal, medical, etc.) to improve their relevance.

- **New tasks to explore** : Integrating more complex benchmarks (advanced reasoning, code generation, multimodal understanding).

5.4 Chapter Conclusion

Ultimately, this internship has been a valuable learning experience, allowing me to both strengthen my theoretical knowledge in NLP and apply it in a comparative and experimental framework. I learned to critically evaluate and compare different language models while developing a rigorous and reusable methodology. The perspectives opened by this work suggest numerous opportunities for improvement, particularly in optimization, fine-tuning, and multilingual integration.

6 GENERAL CONCLUSION

6.1 Internship Summary

This internship represented an enriching opportunity to explore, for the first time, the rapidly growing field of Large Language Models (LLMs) and Natural Language Processing (NLP). Through the three main tasks :**personal glossary**, **comparative mapping** and **practical evaluation** ,I progressed from a theoretical understanding of basic concepts to their concrete application in an experimental setting.

Task 1 allowed me to acquire a solid understanding of essential concepts (NLP, LLMs, tokens, fine-tuning, embeddings, etc.), while Task 2 provided a global overview of recent open-source models, highlighting their characteristics, licenses, and use cases. Finally, Task 3 represented the most practical phase, involving the implementation of a benchmarking protocol, the use of accessible platforms (Poe, Hugging Face, LM Studio), and the comparison of models across various tasks (reasoning, summarization, translation, JSON structuring).

Thus, the overall assessment of this internship is highly positive, as it allowed me to combine **theoretical learning**, **comparative analysis** and **experimental practice** in a field that was entirely new to me.

6.2 Personal and Professional Learnings

On the **professional level**,this internship enabled me to:

- Strengthen my skills in NLP and Artificial Intelligence
- Develop an evaluation methodology applicable to other projects,
- Gain a better understanding of the challenges related to open-source and proprietary models
- Improve my analytical and technical writing abilities

On the **personal level**,this first experience in the field taught me to:

- Organize my work autonomously and progressively
- Overcome technical challenges (hardware limitations, model hallucinations, multilingual inconsistencies)
- Value the importance of verifying information by cross-checking multiple sources
- Simplify and communicate complex concepts clearly and accessibly.

6.3 Future Perspectives

This internship opens up several avenues for future development and extension:

- Deepening experimentation with larger models or in optimized local environments (dedicated GPU/TPU servers).
- Exploring fine-tuning on specialized datasets to adapt model performance to specific professional contexts.
- Expanding multilingual evaluations (English, Classical Arabic, Amazigh, etc.) to test the robustness of models in truly diverse environments.
- Comparing open-source models more systematically with proprietary solutions (GPT-4, Claude, Gemini) to better measure performance gaps.
- Considering practical application projects integrating LLMs (chatbots, intelligent assistants, structured content generation tools).

In short, this first internship in the field of LLMs and NLP served as both an academic and professional springboard, strengthening my foundations in artificial intelligence while opening the door to new areas of exploration.

7 Appendices

https://poe.com/Mistral-7B-v0.3-T

https://poe.com/Llama-3-70B-T

https://datascientest.com/large-language-models-tout-savoir

https://poe.com/DeepSeek-R1

https://poe.com/Gemma-3-27B

https://huggingface.co/openai/gpt-oss-120b

https://huggingface.co/HuggingFaceH4/zephyr-7b-beta

https://www.data-bird.co/blog/differences-llm-nlp

https://choosealicense.com/licenses/