



Leveraging LLMs with RAG to recommend Points of Interest to tourists

STUDENT Patrick Hamzaj
SUPERVISOR Niccolò Marastoni

March 27, 2025

Università degli Studi di Verona

Table of contents

1. Introduction
2. Theoretical Background and Glossary
3. Prompt Injection
4. Experimental Setup
5. Conclusions

Introduction

Objective

The objective of this project is to deliver a practical implementation of an **LLM-powered application** to enhance **tourism experience** in Verona, Italy.

Modern Large Language Models require large computational resources and data to train and inference.

Scaling Law

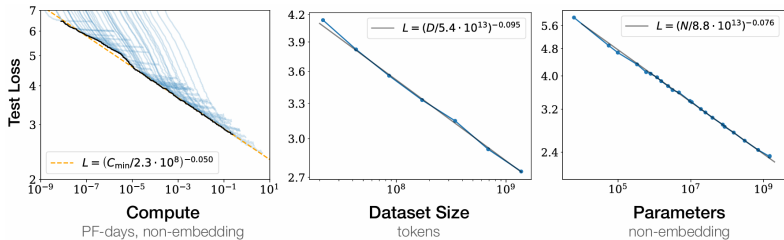


Figure 1: Scaling Law, as seen in [3]

There was a need to harness the generative power of LLMs in a limited resource environment and thus avoiding fine-tuning.

Theoretical Background and Glossary

Large Language Models are based on the **Transformers** architecture—a type of *Neural Network*—which must undergo two learning phases:

1. **Pre-Training**
2. **Fine-Tuning**

Few-Shot Capabilities

Modern open-source models come already pre-trained and fine-tuned, letting developers focus on domain-specific performance.

Alternative approaches have raised exploiting **few-shot capabilities** demonstrated by the introduction of GPT-3. [1]

IN-CONTEXT LEARNING

The ability of a language model to perform new tasks by leveraging examples provided directly within the prompt, rather than through explicit parameter updates.

Fine-Tuning remains a common approach to tailor LLMs' behavior for specific use cases, but innovative techniques have emerged gathering already fine-tuned models for general scenarios (i.e., chat models), directing its focus on specific tasks and use cases:

- **Prompt Engineering**

Fine-Tuning remains a common approach to tailor LLMs' behavior for specific use cases, but innovative techniques have emerged gathering already fine-tuned models for general scenarios (i.e., chat models), directing its focus on specific tasks and use cases:

- Prompt Engineering
- Retrieval Augmented Generation

Fine-Tuning remains a common approach to tailor LLMs' behavior for specific use cases, but innovative techniques have emerged gathering already fine-tuned models for general scenarios (i.e., chat models), directing its focus on specific tasks and use cases:

- Prompt Engineering
- Retrieval Augmented Generation
- Agent AI

Prompt Injection

Anatomy of a prompt

A common standard in chat models has been delineated by OpenAI, for which a categorization of roles has emerged from empirical design choices in conversational AI systems, where:

System It provides the “hidden” guidelines the assistant follows to ensure responses are consistent with desired policies, style, and behavior.

Anatomy of a prompt

A common standard in chat models has been delineated by OpenAI, for which a categorization of roles has emerged from empirical design choices in conversational AI systems, where:

System It provides the “hidden” guidelines the assistant follows to ensure responses are consistent with desired policies, style, and behavior.

User It provides the queries, prompts, or instructions that drive the conversation.

Anatomy of a prompt

A common standard in chat models has been delineated by OpenAI, for which a categorization of roles has emerged from empirical design choices in conversational AI systems, where:

System It provides the “hidden” guidelines the assistant follows to ensure responses are consistent with desired policies, style, and behavior.

User It provides the queries, prompts, or instructions that drive the conversation.

Assistant The language model’s persona that generates responses based on the system instructions and user inputs.

Anatomy of a prompt

A common standard in chat models has been delineated by OpenAI, for which a categorization of roles has emerged from empirical design choices in conversational AI systems, where:

SYSTEM It provides the “hidden” guidelines the assistant follows to ensure responses are consistent with desired policies, style, and behavior.

User It provides the queries, prompts, or instructions that drive the conversation.

Assistant The language model’s persona that generates responses based on the system instructions and user inputs.

ROLE PROMPTING

The practice of explicitly instructing a model to adopt a specific persona or role within the prompt.

You are a tour guide assisting tourists in Verona, Italy. Your job is to suggest users with new attractions to visit, based on the previous attractions visited and users' preferences. Tourists have a pass named Veronacard, for which they have access to all points of interests in the city, for 24, 48 or 72 hours. Your tone is both professional and friendly at the same time.

A model's knowledge ends at the moment of its training—so it will not know about more recent events, current literature or real-time information.

Retrieval Augmented Generation

A framework dynamically integrates specific, contextually relevant knowledge into the generation process, leading to more informed and accurate outputs

`datetime.now()`

Today's date is Tuesday, 07 January 2025.

Call to Openmeteo API.

The weather throughout the day is as follows:

At 07:00 the temperature is -1°C with overcast. The precipitation probability is 0%.

Extract Veronacard info.

You must suggest users on attractions to visit included in the Veronacard, which are the following:

- Arena*
- Casa Giulietta*

Once context-awareness is achieved, a strategy must be designed to dynamically retrieve information about affluence and weather.

AGENTS

Systems that combine reasoning, planning and executing actions, extending the capabilities of LLMs enabling them to act autonomously via external tools to fulfill a task.

*You can call the function **retrieve_affluency** when the user asks how crowded is a certain attraction. You can use the function **get_weather_forecast** when the user asks about weather forecast in the next days.*

Experimental Setup

Table 1: Experimental setup

Parameter	Choice
Model	LLaMa 3.1 8b
Version	Instruct
Quantization	8 bit
Source	Hugging Face
Environment	Google Colab
Hardware	T4 GPU
VRAM	16 GB DDR5

LINK TO THE STREAMLIT APP.

Hey, can you tell me any fun attraction to visit?

1. **Baseline**

There are countless amazing attractions around the world, but I'll give you some popular and unique ones to consider:

1. *The Great Barrier Reef (Australia): As the world's largest coral reef system, it's a snorkeler's and scuba diver's paradise.*

...

Notable Examples

Hey, can you tell me any fun attraction to visit?

1. **Baseline**

2. **Role Prompting**

Benvenuto! Welcome to Verona! I'm thrilled to help you make the most of your VeronaCard. With this pass, you have access to over 30 points of interest in the city. Since you're looking for suggestions, I'll provide you with a few recommendations based on popular attractions and hidden gems.

...

Notable Examples

Hey, can you tell me any fun attraction to visit?

1. **Baseline**
2. **Role Prompting**
3. **RAG**

Buongiorno! Welcome to Verona. The weather today is quite pleasant, so I'd recommend visiting the Giardino Giusti. It's a beautiful garden with a rich history, and it's a great place to relax and enjoy the scenery. Plus, it's not too far from the city center, so you can easily get there on foot.

...

Notable Examples

I don't want too many crowds, what time would you suggest for visiting Giardino Giusti today?

1. **Baseline**
2. **Role Prompting**
3. **RAG**
4. **Agent**

Considering you'd like to avoid crowds, I'd suggest visiting Giardino Giusti in the morning, around 9:00AM. The weather is expected to be partly cloudy, with a temperature of 5 degrees and a slight chance of rain. It might be a bit chilly, but the garden is a beautiful spot to visit, and the early morning hours might help you avoid the larger crowds.

...

HALLUCINATIONS

Term referring to instances where the AI generates content that appears factual and coherent but is ungrounded or incorrect. [2]

Next, take a romantic gondola ride along the Adige River and pass under the famous Ponte Pietra, a beautiful Roman bridge. After that, visit the Casa di Giulietta (Juliet's House), a 14th-century house that inspired Shakespeare's famous balcony scene from Romeo and Juliet.

Time and Multi-Turn Restrictions

Figure 2 shows that there is not a correlation between the number of tokens generated and the time required to give an output, though it demonstrates that it can take up to 30 seconds to produce a response, as it can significantly influence the user experience.

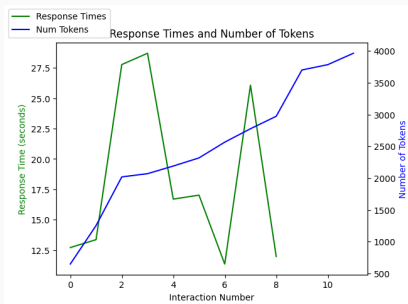


Figure 2: Number of tokens and time required.

Conclusions

The current approach reveals that high computational resources are not strictly necessary for achieving effective and natural responses. However, the model's reliance on foundational weights can lead to hallucinations and issues related to context length and memory remains critical.


Conclusions

Pros


- Fast to implement
- Low cost
- Easy to understand

Cons

- Limited flexibility
- Not scalable
- Requires manual updates

 T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei.

Language models are few-shot learners, 2020.

 Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung.

Survey of hallucination in natural language generation.

ACM Computing Surveys, 55(12):1–38, Mar. 2023.



J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei.
Scaling laws for neural language models, 2020.