Explainability in LLMs

TABLE OF CONTENTS

01
Motivation

Motivation

Users need to understand model decisions, especially in high-stakes domains (e.g., healthcare, legal, finance).

Traditional ML models (e.g., trees, logistic regression) have established explainability methods, but LLMs require different approaches.

LLMs are rapidly complementing and even replacing traditional models, making it crucial to understand how they reach decisions.

Prompting strategies like CoT and ReAct enhance both performance and explainability.

Explainability builds trust and ensures reliable outcomes. Without clear reasoning behind LLM outputs, users may struggle to trust the model's decisions, particularly in sensitive contexts.

Poor explainability can lead to poor performance. LLMs often perform better when prompted in ways that reveal their reasoning process, improving accuracy and robustness.

02

Literature Review

OUR TEAM

SHAP

Feature importance, explaining individual predictions, and model debugging

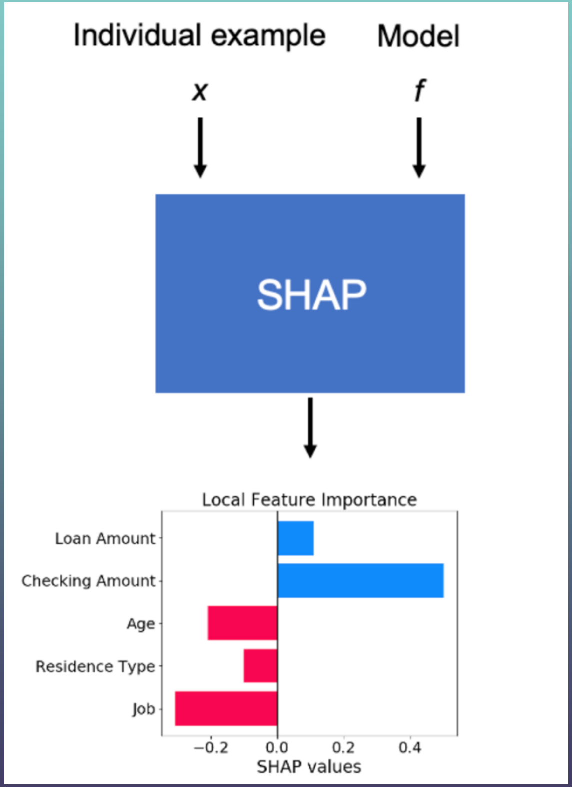
LIME

Intuitive for explaining tabular, text, and image data

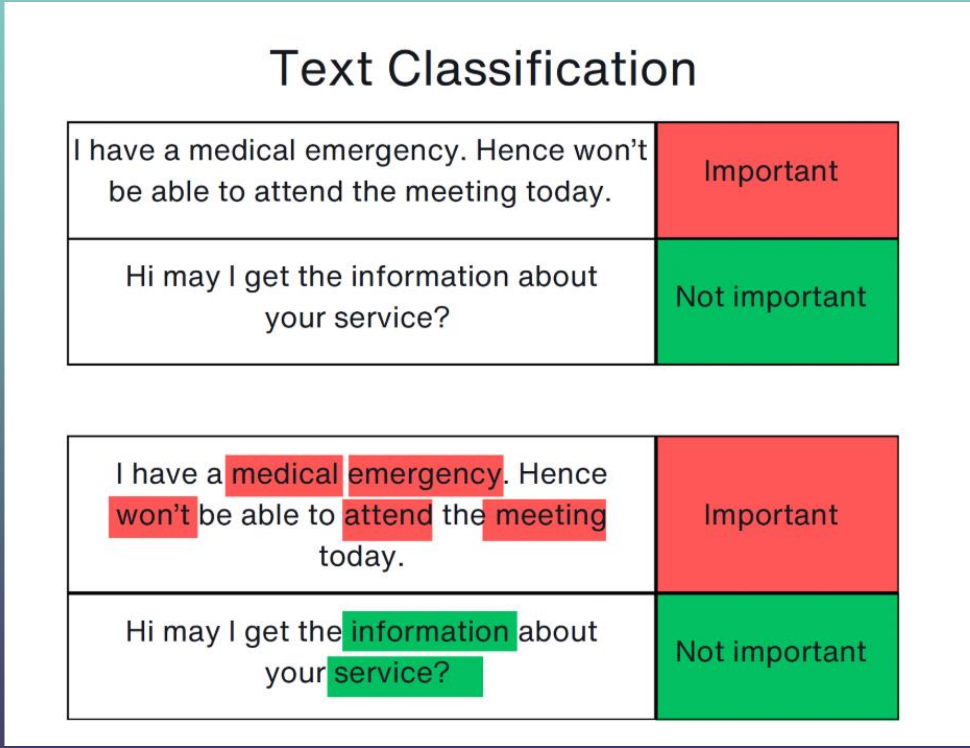
Less popular tools (ELI5)

Quick model inspection, feature weight visualization, and debugging

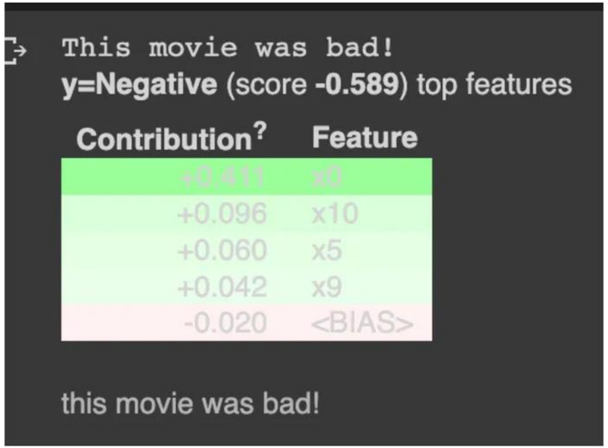
SHAP


Individual example
Model
x
f
SHAP
Local Feature Importance
Loan Amount
Checking Amount
Age
Residence Type
Job
−0.2
0.0
0.2
0.4
SHAP values

LIME


Text Classification
I have a medical emergency. Hence won't be able to attend the meeting today.
Important
Hi may I get the information about your service?
Not important
I have a medical emergency. Hence won't be able to attend the meeting today.
Important
Hi may I get the information about your service?
Not important

ELI5



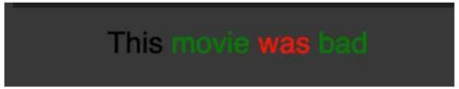
This movie was bad!

y=Negative (score **-0.589**) top features

Contribution?	Feature
+0.411	x0
+0.096	x10
+0.060	x5
+0.042	x9
-0.020	<BIAS>

this movie was bad!

Impact of words (features) on prediction (=negative, in this case).



The words “movie” and “bad” contributed to the “negative” prediction positively and the word “was” contributed negatively.

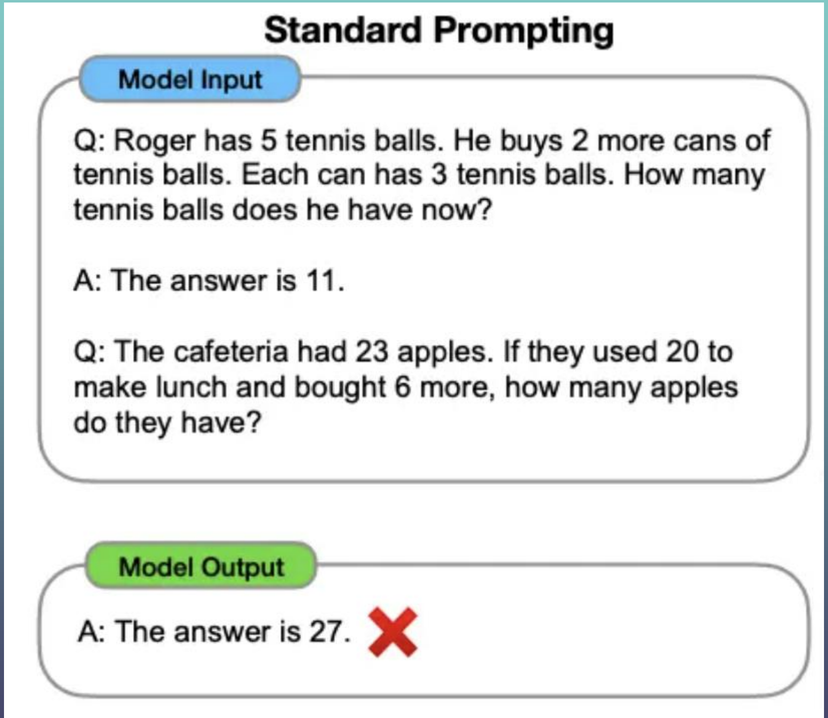
03

Approach Used

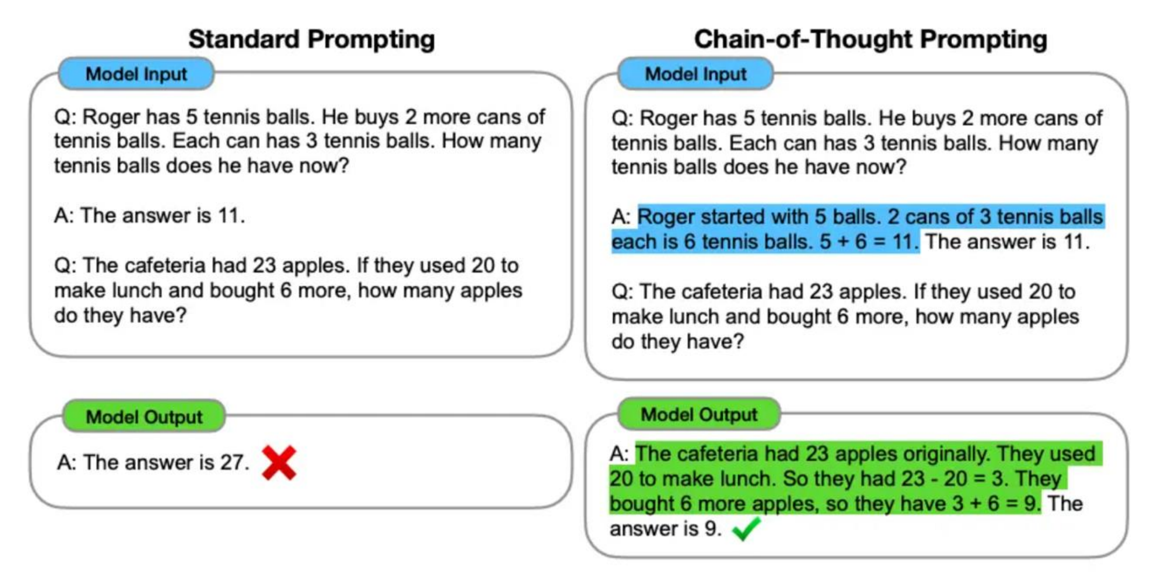
Methods

Method	DESCRIPTION
Zero shot & Few shot prompting	Zero shot : You provide the model with a prompt and expect it to generate the correct response without any examples. Few-shot is similar, but with examples.
Chain of Thought prompting	CoT encourages the model to explain its reasoning step by step, improving performance in complex reasoning tasks.
Meta-prompting	In meta-prompting, the prompt itself is designed to influence the structure or style of the model's response.
Self-consistency prompting	Generates multiple solutions to a given problem, then selects the most common or coherent result.
Generated Knowledge prompting	This technique instructs the model to first generate background knowledge or information before answering the main query.
Prompt chaining	Prompt chaining involves splitting complex tasks into multiple sequential prompts, where the output of one prompt becomes the input for the next.
RAG	RAG combines a language model with a retrieval system that fetches relevant documents or facts from an external database before generating a response.
ReAct	ReAct combines **reasoning** with **actions**, where the model iteratively reflects on information, interacts with tools or APIs, and refines its response.

Zero-shot & Few Shot



Chain of Thought prompting



Meta-prompting

Problem Statement:

- **Problem**: [question to be answered]

Solution Structure:

1. Begin the response with "Let's think step by step."
2. Follow with the reasoning steps, ensuring the solution process is broken down clearly and logically.
3. End the solution with the final answer encapsulated in a LaTeX-formatted box, $\boxed{...}$, for clarity and emphasis.
4. Finally, state "The answer is [final answer to the problem].", with the final answer presented in LaTeX notation.

———

Problem: Find the domain of the expression $\frac{\sqrt{x-2}}{\sqrt{5-x}}$.
Solution: The expressions inside each square root must be non-negative. Therefore, $x-2 \geq 0$, so $x \geq 2$, and $5-x \geq 0$, so $x \leq 5$. Also, the denominator cannot be equal to zero, so $5-x>0$, which gives $x<5$. Therefore, the domain of the expression is $\boxed{[2,5)}$. Final Answer: The final answer is $[2,5)$. I hope it is correct.

———

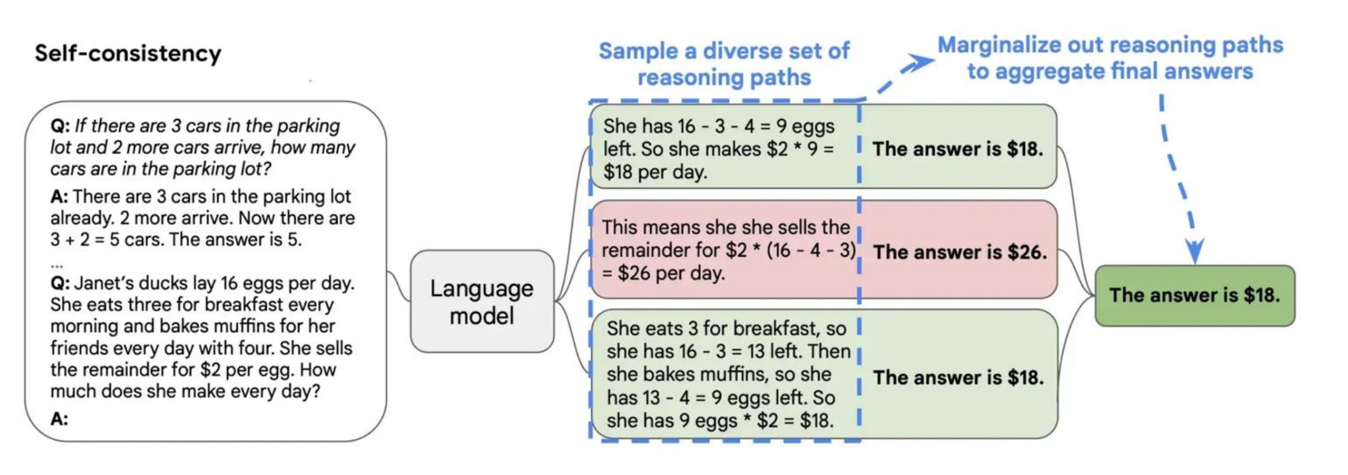
Problem: If $\det \mathbf{A} = 2$ and $\det \mathbf{B} = 12$, then find $\det(\mathbf{AB})$.
Solution: We have that $\det(\mathbf{AB}) = (\det \mathbf{A})(\det \mathbf{B}) = (2)(12) = \boxed{24}$. Final Answer: The final answer is 24. I hope it is correct.

———

...

Focuses on the structural and syntactical aspects of tasks and problems rather than their specific content details

Self-consistency prompting

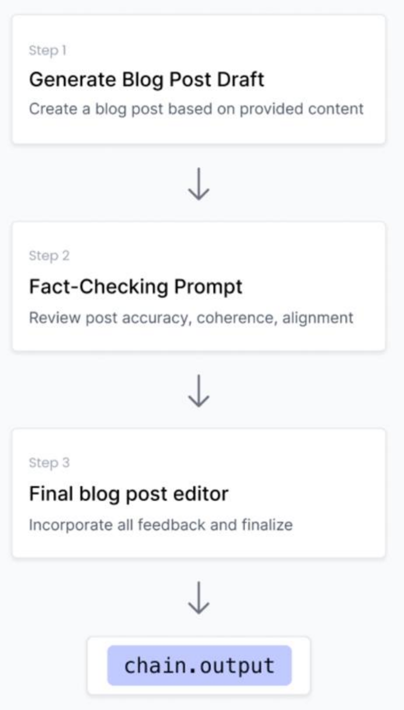


Generates multiple solutions with different thought processes and selects answer with the most votes

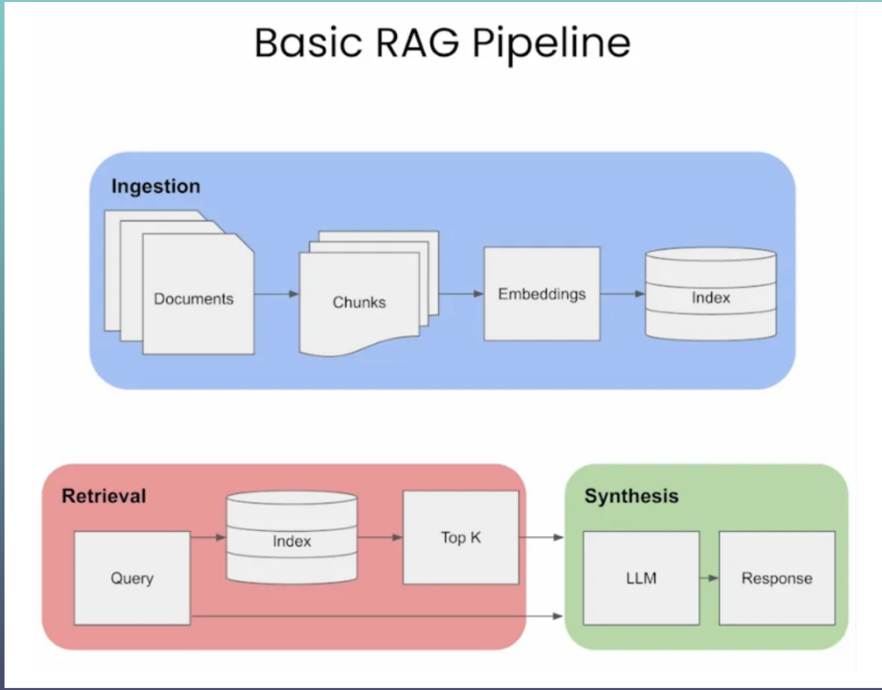
Generated-knowledge prompting

```
Generate 4 facts about the Kermode bear, then use these facts to write a short blog post using the
information:

The Kermode bear, also known as the spirit bear, is a rare subspecies of the American black bear that
is found in the coastal regions of British Columbia, Canada.

1. The Kermode bear is a unique subspecies of the American black bear, with approximately 10% of the
population having white fur.

2. The Kermode bear is a symbol of the First Nations people of British Columbia and is seen as a spirit
animal.

3. The Kermode bear is an important part of the ecosystem, helping to disperse seeds and maintain the
balance of the forest.

4. The Kermode bear is a protected species, and its population is monitored closely by the provincial
government.

The Kermode bear is a unique and special animal that is found in the coastal regions of British
Columbia, Canada. Approximately 10% of the population of this subspecies of the American black bear
have white fur, making them a rare sight. The Kermode bear is a symbol of the First Nations people of
British Columbia and is seen as a spirit animal. It is an important part of the ecosystem, helping to
disperse seeds and maintain the balance of the forest. The Kermode bear is a protected species, and its
population is monitored closely by the provincial government. It is a beautiful and majestic animal,
and we must do our part to protect and conserve this species.
```

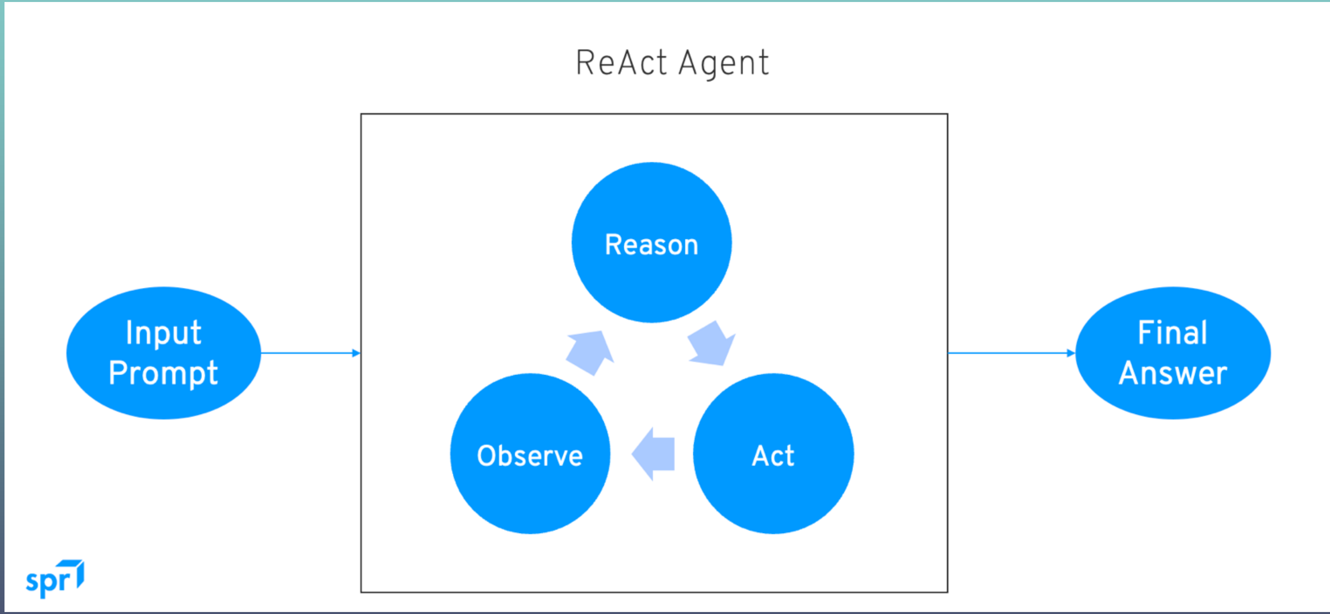
Prompt chaining



RAG


Basic RAG Pipeline
Ingestion
Documents
Chunks
Embeddings
Index
Retrieval
Query
Index
Top K
Synthesis
LLM
Response

ReAct

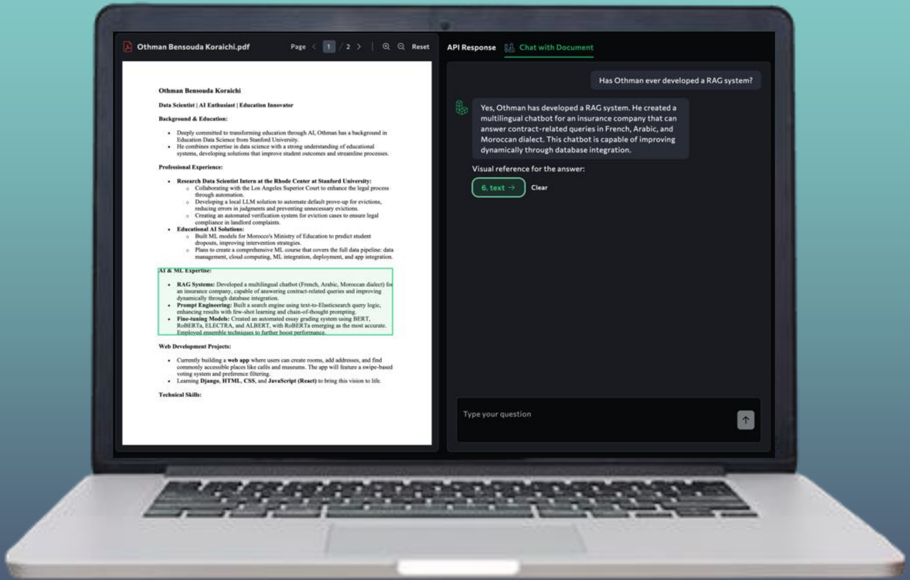


ReAct Agent
Reason
Input Prompt
Final Answer
Observe
Act
spr

04

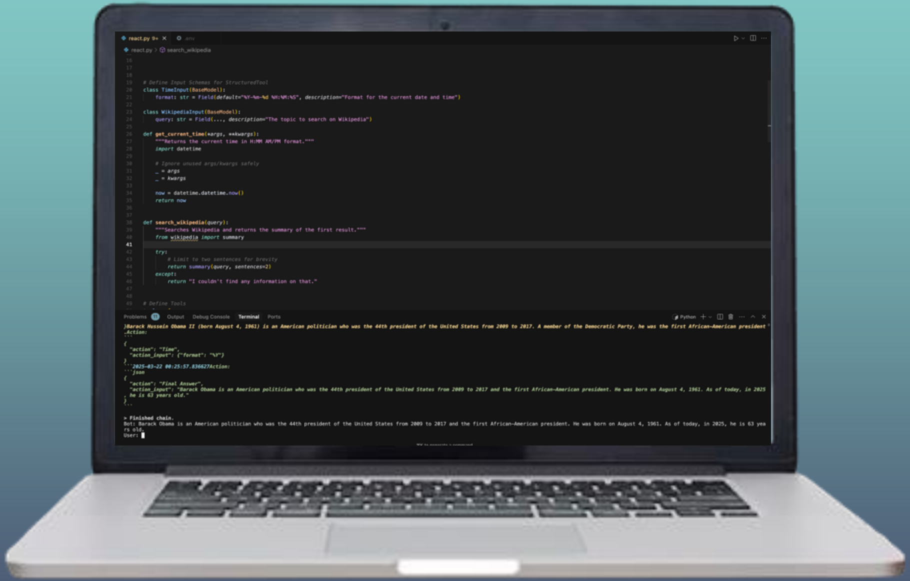
Demonstration

DEMO (RAG)



Othman Bensouda Koraichi.pdf
Page 1 / 2
Reset
Othman Bensouda Koraichi
Data Scientist | AI Enthusiast | Education Innovator
Background & Education:
Deeply committed to transforming education through AI, Othman has a background in Education Data Science from Stanford University.
He combines expertise in data science with a strong understanding of educational systems, developing solutions that improve student outcomes and streamline processes.
Professional Experience:
Research Data Scientist Intern at the Rhode Center at Stanford University:
Collaborating with the Los Angeles Superior Court to enhance the legal process through automation.
Developing a local LLM solution to automate default prove-up for evictions, reducing errors in judgments and preventing unnecessary evictions.
Creating an automated verification system for eviction cases to ensure legal compliance in landlord complaints.
Educational AI Solutions:
Built ML models for Morocco's Ministry of Education to predict student dropouts, improving intervention strategies.
Plans to create a comprehensive ML course that covers the full data pipeline: data management, cloud computing, ML integration, deployment, and app integration.
AI & ML Expertise:
RAG Systems: Developed a multilingual chatbot (French, Arabic, Moroccan dialect) for an insurance company, capable of answering contract-related queries and improving dynamically through database integration.
Prompt Engineering: Built a search engine using text-to-Elasticsearch query logic, enhancing results with few-shot learning and chain-of-thought prompting.
Fine-tuning Models: Created an automated essay grading system using BERT, RoBERTa, ELECTRA, and ALBERT, with RoBERTa emerging as the most accurate. Employed ensemble techniques to further boost performance.
Web Development Projects:
Currently building a web app where users can create rooms, add addresses, and find commonly accessible places like cafés and museums. The app will feature a swipe-based voting system and preference filtering.
Learning Django, HTML, CSS, and JavaScript (React) to bring this vision to life.
Technical Skills:
API Response
Chat with Document
Has Othman ever developed a RAG system?
Yes, Othman has developed a RAG system. He created a multilingual chatbot for an insurance company that can answer contract-related queries in French, Arabic, and Moroccan dialect. This chatbot is capable of improving dynamically through database integration.
Visual reference for the answer:
6. text →
Clear
Type your question

DEMO (ReAct)



react.py 9+
.env
react.py > search_wikipedia
# Define Input Schemas for StructuredTool
class TimeInput(BaseModel):
format: str = Field(default="%Y-%m-%d %H:%M:%S", description="Format for the current date and time")
class WikipediaInput(BaseModel):
query: str = Field(..., description="The topic to search on Wikipedia")
def get_current_time(*args, **kwargs):
"""Returns the current time in H:MM AM/PM format."""
import datetime
# Ignore unused args/kwargs safely
_ = args
_ = kwargs
now = datetime.datetime.now()
return now
def search_wikipedia(query):
"""Searches Wikipedia and returns the summary of the first result."""
from wikipedia import summary
try:
# Limit to two sentences for brevity
return summary(query, sentences=2)
except:
return "I couldn't find any information on that."
Define Tools
Problems
Output
Debug Console
Terminal
Ports
Python
}Barack Hussein Obama II (born August 4, 1961) is an American politician who was the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, he was the first African-American president
.Action:
"action": "Time",
"action_input": {"format": "%Y"}
2025-03-22 00:25:57.836627Action:
json
"action": "Final Answer",
"action_input": "Barack Obama is an American politician who was the 44th president of the United States from 2009 to 2017 and the first African-American president. He was born on August 4, 1961. As of today, in 2025, he is 63 years old."
> Finished chain.
Bot: Barack Obama is an American politician who was the 44th president of the United States from 2009 to 2017 and the first African-American president. He was born on August 4, 1961. As of today, in 2025, he is 63 years old.
User:

References

- **SHAP :** Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems, 30.
- **LIME :** Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), 1135-1144. https://doi
- **ELI5 : https://eli5.readthedocs.io/en/latest/overview.html**
- **Few-shot prompting :** Brown et al. (2020). *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems, 33, 1877-1901.
- **CoT :** Wei et al. (2022). *Chain-of-thought prompting elicits reasoning in large language models*. In *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022)* (Article No. 1800, pp. 24824–24837).
- **Meta-prompting :** Zhang, Y., Yuan, Y., & Yao, A. C.-C. (2023). *Meta Prompting for AI Systems*. arXiv preprint arXiv:2311.11482. https://doi.org/10.48550/arXiv.2311.11482
- **Self-consistency :** Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. arXiv preprint arXiv:2203.11171. https://doi.org/10.48550/arXiv.2203.11171
- **Generated-Knowldege prompting :** Liu, J., Liu, A., Lu, X., Welleck, S., West, P., Le Bras, R., Choi, Y., & Hajishirzi, H. (2022). *Generated Knowledge Prompting for Commonsense Reasoning*. arXiv preprint arXiv:2110.08387. https://doi.org/10.48550/arXiv.2110.08387
- **Prompt chaining :** Sun, S., Yuan, R., Cao, Z., Li, W., & Liu, P. (2024). *Prompt Chaining or Stepwise Prompt? Refinement in Text Summarization*. arXiv preprint arXiv:2406.00507. https://doi.org/10.48550/arXiv.2406.00507
- RAG : Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401. h
- **ReAct :** Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). *ReAct: Synergizing Reasoning and Acting in Language Models*. arXiv preprint arXiv:2210.03629. https://doi.org/10.48550/arXiv.2210.03629
- **Prompts : https://www.promptingguide.ai/techniques/knowledge**