Evaluate_RAGs

March 21, 2025

1 Evaluating Models.

To evaluate the performance of our advanced Retrieval-Augmented Generation (RAG) system, we compared four configurations using a diverse set of five input questions about jobs and hiring trends: (a) a base LLM without retrieval, (b) a basic RAG setup with a retriever and generator, (c) an advanced multi-agent RAG using a base model, and (d) an advanced RAG using a fine-tuned model with LoRA.

In conclusion, we observed a clear progression in quality from base LLM to advanced agentic RAG. The base model with agentic control provides the most precise and informative results, while the fine-tuned model offers clean summaries but can sometimes overgeneralize. The effectiveness of the agent-based architecture becomes especially evident when the task demands structured extraction or constraint satisfaction, reinforcing the value of system orchestration in RAG pipelines.

1.1 Code to Evaluate

[ ]:
```
from IPython.display import Image, display
from workflow import create_workflow  # Importing workflow function
from langgraph.graph import END, StateGraph, START
from langchain_core.runnables.graph import MermaidDrawMethod
from agents import (query_rewriter, retriever_agent, grade_documents,␣
 ↪generate_agent, verification_agent,generate_agent_lora)
import pandas as pd
from agents import llm, retriever
from workflow import create_workflow
```

[2]:
```
from typing_extensions import TypedDict, List, Literal

class RAGState(TypedDict):
    query: str
    refined_query: str
    retrieved_docs: List[str]
    formatted_context: str
    llm_answer: str
    decision: Literal["relevant", "not_relevant", "useful", "not_useful", "end"]
```

[3]:
```
from pprint import pprint
```

```
def run_query(query: str, use_lora: bool = False):
    """
    Runs the multi-agent RAG system with a given query and prints intermediate␣
 ↪states.

    """

    # Initialize State
    input_state = {
        "query": query,
        "refined_query": "",
        "retrieved_docs": [],
        "formatted_context": "",
        "llm_answer": "",
        "decision": "",
        "retries": 0  # Initial retry count
    }
    app = create_workflow(use_lora=use_lora)
    # Store Final Output State
    final_output = {}

    # Run Workflow & Capture Intermediate States
    for step, output in enumerate(app.stream(input_state)):

        # Merge new values with final output state
        final_output.update(output)

    # Extract `llm_answer`
    final_answer = final_output.get("generate_agent", {}).get("llm_answer", "No␣
 ↪`llm_answer` generated.")

    # Extract `decision`
    decision = final_output.get("verification_agent", {}).get("decision", " No␣
 ↪`decision` available.")
    return {
        "query": query,
        "response": final_answer,
        "decision": decision
    }
```

[4]:
```
generated = run_query("What are the highest paying remote Data Science jobs?")
generated['response']
```

```
/Users/ishitakokil/Desktop/MLDS 424/Stitching /agents.py:52:
LangChainDeprecationWarning: The method `BaseChatModel.predict` was deprecated
in langchain-core 0.1.7 and will be removed in 1.0. Use :meth:`~invoke` instead.
  refined_query = llm.predict(f"Rewrite the following query to improve search
```

```
results:\n{state['query']}")
```

[4]:
```
'**Staff Data Scientist** at **Jobot Consulting**\nLocation: New York,
NY\nExperience Level: Entry level\nWork Type: FULL_TIME\nSalary:
208000.0\n\nThis position offers the highest salary among the listed options at
$208,000. It is located in New York, NY and is an entry-level position.'
```

[5]:
```
test_queries = [
    "What marketing jobs are available in New York?",
    "Which companies are hiring AI engineers?",
    "Find jobs requiring Python but not Java.",
    "List remote AI research jobs.",
    "What are the highest-paying data science roles?"
]

def run_evaluation(query):
    results = {"Query": query}

    # Base LLM (No RAG)
    results["Base LLM"] = llm.predict(query)

    # Basic RAG (Retriever + LLM)
    retrieved_docs = retriever.invoke(query)
    if retrieved_docs:
        retrieved_text = "\n".join([doc.page_content for doc in retrieved_docs])
        results["Basic RAG"] = llm.predict(f"Based on these job listings:
 ↪\n\n{retrieved_text}\n\nAnswer the query: {query}")
    else:
        results["Basic RAG"] = "No relevant documents found."

    # Advanced RAG (Base Model)
    ans_rag = run_query(query, use_lora=False)
    results["Advanced RAG (Base)"] = ans_rag['response']

    # Advanced RAG (Fine-Tuned LoRA)
    ans_rag_lora = run_query(query, use_lora=True)
    results["Advanced RAG (LoRA)"] = ans_rag_lora['response']

    return results

evaluation_results = [run_evaluation(query) for query in test_queries]
df_results = pd.DataFrame(evaluation_results)
```

[11]:
```
for _, row in df_results.iterrows():
    print("\n Query:", row["Query"])
    print("")
    print("BASE LLM")
```

```
    print("")
    print(row["Base LLM"])
    print("")
    print("BASIC RAG")
    print("")
    print(row["Basic RAG"])
    print("")
    print("Advanced RAG (Base)")
    print("")
    print(row["Advanced RAG (Base)"])
    print("")
    print("Advanced RAG (LoRA)")
    print("")
    print(row["Advanced RAG (LoRA)"])
    print("")
    print("=" * 100)
```

```
 Query: What marketing jobs are available in New York?

BASE LLM

1. Marketing Manager
2. Digital Marketing Specialist
3. Social Media Manager
4. Content Marketing Manager
5. Marketing Coordinator
6. Public Relations Specialist
7. Brand Manager
8. Market Research Analyst
9. Email Marketing Specialist
10. SEO Specialist
11. Advertising Account Executive
12. Event Marketing Manager
13. Influencer Marketing Manager
14. Marketing Communications Manager
15. Product Marketing Manager

BASIC RAG

Based on the job listings provided, the marketing jobs available in New York
are:

1. Marketing and Business Development Coordinator at Withers in New Haven, with
occasional travel to Greenwich, New York, and Boston.
2. Sales and Marketing Professionals at a leading sales & marketing firm in New
Jersey, with territories in New Jersey and New York.
3. Entry Level Sales and Marketing Positions at a firm in New Jersey with
```

opportunities for growth.
4. Face-to-Face Marketing position in Charlotte, with a team expanding from Atlanta to unlock the market.

Advanced RAG (Base)

Head of Strategic Partnerships, Americas at **GetYourGuide**
- Location: New York, NY
- Experience Level: Mid-Senior level
- Work Type: FULL_TIME
- Salary: 195000.0

Lease Admin / Tenant Billing Coordinator at **Advice Personnel**
- Location: Oyster Bay, NY
- Experience Level: Associate
- Work Type: FULL_TIME
- Salary: 75000.0

Marketing And Public Relations Specialist at **Millennium Recruiting, Inc.**
- Location: Irvine, CA
- Experience Level: Entry level
- Work Type: FULL_TIME
- Salary: N/A

The best option for marketing jobs available in New York would be the **Head of Strategic Partnerships, Americas** position at **GetYourGuide**. It offers a competitive salary and is suitable for mid-senior level professionals.

Advanced RAG (LoRA)

The job listings require different experience levels, ranging from Associate to Mid-Senior level, and vary in salary from $75,000 to $195,000.

==========

Query: Which companies are hiring AI engineers?

BASE LLM

Many tech companies and startups are hiring AI engineers, including:

1. Google
2. Amazon
3. Microsoft
4. Facebook
5. Apple
6. IBM

7. Nvidia
8. Intel
9. Uber
10. Airbnb

Additionally, there are many smaller companies and startups in industries such as healthcare, finance, and retail that are also hiring AI engineers. It is recommended to check job boards, company websites, and networking events to find opportunities in the field of AI engineering.

BASIC RAG

Based on the job listings provided, the companies that are hiring AI engineers are looking for professionals to join their teams. These companies are focused on developing cutting-edge AI solutions, implementing AI services and algorithms, and working on AI models in a production environment. The companies hiring AI engineers are looking for individuals who are passionate about technology, have strong technical capabilities, and are committed to innovation. These companies value continuous learning, technical excellence, and building strong client relationships. The companies hiring AI engineers are focused on delivering transformative and innovative solutions for their customers in various industries such as supply chain management, financial transactions, and government agencies.

Advanced RAG (Base)

The companies currently hiring AI engineers include:

1. **Riverbed Technology** - Seeking a Senior AI Engineer in North Carolina, United States.
2. **Coretek** - Looking for an Azure AI Architect in the United States.
3. **Commit: AI Career Agents for Developers** - Offering a position for an AI Product Engineer in the United States.

Among these options, the **Senior AI Engineer position at Riverbed Technology** may be the best choice as it is specifically focused on AI engineering and is at a senior level, indicating potential for growth and leadership in the field.

Advanced RAG (LoRA)

These are three job descriptions for software engineer positions, each with different experience levels and salaries. The positions are located in the United States.

==========

Query: Find jobs requiring Python but not Java.

BASE LLM

1. Python Software Engineer
2. Data Scientist (Python)
3. Machine Learning Engineer (Python)
4. Backend Developer (Python)
5. Automation Engineer (Python)
6. DevOps Engineer (Python)
7. Data Analyst (Python)
8. Python Developer (Web)
9. Computer Vision Engineer (Python)
10. Natural Language Processing Engineer (Python)

BASIC RAG

Based on the job listings provided, the following job requires Python but not Java:

Job Title/Role: Python Developer (Python 3.0 or above, Django, Flask, Angular/React/Vue/JavaScript, Python Libraries (Panda/NumPy/Sci-kit), SQL and PL/SQL) - Location: Charlotte, NC, Jersey City, NJ, New York, NY - Open to Relocate candidates can also apply.

This job listing specifically mentions Python as a required technology, but does not mention Java as a requirement.

Advanced RAG (Base)

Based on your query, the best option among the listed job listings would be the **REMOTE Python and Django Developer** position at Apex Systems in Lemont, IL. This job specifically requires Python and Django skills without mentioning Java, making it a good fit for your criteria. The salary for this position is not specified.

Advanced RAG (LoRA)

The job descriptions list a variety of technical roles, including Java Streaming Technical Lead, Java Developer, Python Developer, and Python Developer.

==========

Query: List remote AI research jobs.

BASE LLM

1. Research Scientist, AI - Remote at Facebook

2. Senior AI Research Scientist - Remote at Google
3. AI Research Engineer - Remote at Microsoft
4. Machine Learning Research Scientist - Remote at Amazon
5. Senior AI Researcher - Remote at IBM
6. AI Research Scientist - Remote at Tesla
7. Research Engineer, Robotics/AI - Remote at NVIDIA
8. AI Researcher - Remote at OpenAI
9. Machine Learning Researcher - Remote at DeepMind
10. Senior AI Scientist - Remote at Intel

BASIC RAG

1. Junior AI/ML Software Engineer at ECS
2. Python Developer at Autod AI
3. Senior Software Engineer with AI and Machine Learning experience (remote)

Advanced RAG (Base)

Python Developer at **Autod AI**
Location: United States
Experience Level:
Work Type: FULL_TIME
Salary: N/A

Among the listings provided, the best option for a remote AI research job is the **Python Developer** position at Autod AI, which allows for remote work in the United States.

Advanced RAG (LoRA)

The job descriptions for the Data Scientist III, Innovation at WEX and the Python Developer at Autod AI include various experience levels, work types, and salaries, while the Program Analyst at SAIC has a single experience level and salary.

==========
==========

Query: What are the highest-paying data science roles?

BASE LLM

Some of the highest-paying data science roles include:

1. Data Science Manager/Director: This role involves overseeing a team of data scientists and leading data analytics initiatives within an organization. Data science managers/directors can earn six-figure salaries, with some earning over $200,000 per year.

2. Data Scientist (Senior Level): Experienced data scientists with advanced skills in machine learning, deep learning, and statistical analysis can command high salaries, often exceeding $150,000 per year.

3. Data Engineer: Data engineers are responsible for designing and building data pipelines, data warehouses, and other infrastructure to support data analytics. Experienced data engineers can earn salaries in the range of $120,000 to $150,000 per year.

4. Data Architect: Data architects design and implement data management solutions that meet the specific needs of an organization. They often work closely with data scientists and data engineers to ensure data quality and consistency. Data architects can earn salaries upwards of $150,000 per year.

5. Chief Data Officer (CDO): CDOs are responsible for overseeing an organization's data strategy and ensuring that data is used effectively to drive business decisions. CDOs are typically executives with extensive experience in data science and analytics, and they can earn salaries in the range of $200,000 to $300,000 per year.

BASIC RAG

Based on the job listings provided, the highest-paying data science roles would likely be the positions requiring expertise in conducting data science work, experience with tools for reproducibility, proficiency in multiple programming languages such as R, Python, Stata, SAS, SQL, and experience in areas like natural language processing, GIS, and data visualization. The salary range for these positions is $88,000 - $157,000, making them potentially the highest-paying data science roles within the organization.

Advanced RAG (Base)

The highest-paying data science roles listed are the **Data Engineer** positions at **Orbit Recruitment Group** in Philadelphia, PA, Phoenix, AZ, and New York, NY, all offering a salary of $155,000. These roles are for mid-senior level experience and are remote positions.

Advanced RAG (LoRA)

There are two job postings for data engineer positions at different locations and experience levels. The salary range for the entry-level position is 122500.0 while the salary for the mid-senior level position is 155000.0.

==========

1.2 Qualitative Evaluations.

- The base LLM provided generic responses with little grounding in real data. While it offered reasonable guesses based on prior knowledge, the answers were often vague, outdated, or inaccurate — for example, listing high-profile companies hiring for AI roles without citing real listings. This made the base model suitable only for brainstorming or general awareness, not reliable insights.
- With the basic RAG, performance improved noticeably. The model could pull in job descriptions from a retrieval corpus, offering more grounded answers. However, it lacked the ability to prioritize or filter results meaningfully. This often led to irrelevant or geographically mismatched entries, and the answers sometimes missed key user constraints (e.g., excluding Java). It demonstrated that while retrieval alone enhances accuracy, it needs additional reasoning for refinement.
- The advanced agentic RAG using a base model outperformed the previous two by a wide margin. It combined retrieval with coordination between agents (query rewriting, document grading, answer generation, verification), leading to more accurate and context-aware outputs. It captured key constraints (such as excluding Java), offered well-structured summaries with salary and location details, and surfaced companies actually present in the dataset. This version struck the best balance between specificity, accuracy, and completeness.
- Finally, the advanced agentic RAG using a LoRA fine-tuned model showed both strengths and weaknesses. Its standout feature was the ability to generate well-summarized responses, often including helpful salary or level overviews. However, it sometimes abstracted too far and lost critical information like job titles, company names, or constraints explicitly stated in the question. This suggests the fine-tuning was skewed toward general summaries and lacked sufficient examples involving filtering or constraint-based reasoning.