

Agentic LLMs

Micro-Internship
2024

Data Reply

01

What is an agent?

LLM vs Agent

- Acts as a passive system that generates responses purely based on input prompts.
 - It does not perform any external actions or interact with external systems.
 - Limited to producing text-based responses. If external interaction is required the user must do it manually.
 - Example: Writing essays, explaining concepts, summarizing text.
- Can interact with tools, APIs, or external environments to gather information, make decisions, and take actions.
 - Has autonomy to decide what to do next based on intermediate reasoning or feedback loops.
 - Useful for complex problem-solving requiring external information or interaction with dynamic environments.

02

Industry research

Industry research

- Retail
 - Automated Replenishment
 - Personalized Recommendations
- Manufacturing
 - Aftermarket Services - handle customer inquiries
 - Sales and Marketing - content creation, social media
- Finance
 - Automated Trading - optimise investment portfolios
 - Automated Underwriting - make lending decisions

Value Chains - retail

Series of activities or processes that create value for a retailer by delivering products or services to consumers

Product Sourcing and Procurement

- Automated replenishment
- Supplier selection
- Negotiate prices, terms, and delivery schedules.

Marketing and Sales

- Website functionality, user experience
- Manage services like click-and-collect and returns

Customer Service and Support

- Answering questions, providing product information
- Handling returns, exchanges, and warranties

Finance and Administration

- Maintain accurate financial records, including sales revenue, taxes
 - Ensure adherence to laws, including tax regulations and health and safety requirements.
-

Problem and statistics

- Stockouts
 - A large quantitative study measured the performance of manual and automatic replenishment processes in a sample of 85 stores and 95 products of a major retail chain in Serbia. The study concluded that utilization of an automatic replenishment system can reduce stock-outs for the retail chains up to 60%. Specifically, when ordered through an automatic replenishment system, fast-selling products recorded 40% greater availability, products on promotion 48% higher availability, and products in a high-density retail stores 59% higher availability.
 - A global survey revealed that only 12% of retailers utilize machine learning technologies in their replenishment activities.
- Overstocks - wastes money
- Backorders - when customer orders product that is waiting to be restocked
 - reduces customer satisfaction, miss out on profits

03

Agent development

Agent setup

```
: model_parameter = {"temperature": 0.9, "top_p": .5, "max_tokens_to_sample": 200}
modelId = "anthropic.claude-3-sonnet-20240229-v1:0"
bedrock_llm = ChatBedrock(
    model_id=modelId,
    client=boto3.client('bedrock'),
    model_kwargs=model_parameter,
    beta_use_converse_api=True
)

: embeddings = BedrockEmbeddings(
    client=boto3.client('bedrock-runtime', region_name="us-east-1"),
    model_id = "amazon.titan-embed-text-v2:0"
)
```

Tools

Retriever

```
# Define the path to the directory containing CSV files
data_files_directory = "data_files"

# Use DirectoryLoader to load all CSV files
loader = DirectoryLoader(
    data_files_directory,
    glob="**/*.csv", # Match all CSV files in the directory (including subdirectories)
    loader_cls=CSVLoader # Use CSVLoader for parsing the files
)
```

```
docs = loader.load()
documents = RecursiveCharacterTextSplitter(
    chunk_size=1000, chunk_overlap=200
).split_documents(docs)
vector = FAISS.from_documents(documents, embeddings)
retriever = vector.as_retriever()
```

```
retriever_tool = create_retriever_tool(
    retriever,
    "Retriever_tool",
    "First search internal data files to answer any queries. If there is no useful information then use other tools."
)
```

```
tools = [datetime_tool, retriever_tool, search, repl_tool, loaded_tools[0]]
```

Tavily

```
os.environ["TAVILY_API_KEY"] = "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"
search = TavilySearchResults()
```

Maths - Python REPL

```
python_repl = PythonREPL()
repl_tool = Tool(
    name="python_repl",
    description="Use this to execute python commands and calculations.",
    func=python_repl.run,
)
```

Wikipedia

```
wikipedia_tool = load_tools(["wikipedia"], llm=bedrock_llm)
```

Date time

```
from datetime import datetime
datetime_tool = Tool(
    name="Datetime",
    func=lambda x: datetime.now().isoformat(),
    description="Returns the current datetime",
)
```

Prompt engineering

```
prompt = PromptTemplate(  
    input_variables=["input", "agent_scratchpad"], # Include both input and agent_scratchpad  
    template="""
```

If the input is to do with retail or an item to buy, do the following:

Use the retriever tool to answer any queries related to internal data files.
Only use external tools (like Wikipedia) if no relevant information is found internally.
When asked about a specific item output the answers to the following queries:

Should this item be ordered?

Find out what date it is currently using date time tool. Convert the month and year into strings and use these for the rest of the answer.

Use recent stock statistics to see if the stocks for this month are low compared to how much was used last month.

Use the statistics from last year to predict how much of each item will be needed for the rest of this month and next month.

Use what month it is at the time the call is being made and whether any important holidays eg christmas or summer will be coming up and use this information in your decision.

- a) how much should be ordered for this month and next month
- b) how much should be ordered in the next few months

What brand?

Use the customers reviews to decide this.

If this item should be ordered, how much of this item should be ordered?

Use the repl tool to calculate these numbers.

In your final output:

1. Explain your reasoning for whether an item should be ordered and how much.

Don't explain your reasoning for

Explain the maths you used.

2. Draft an email to an imaginary supplier of this brand to order the number of the item you think should be ordered, signing off as Reply Auto Replenishment.

Else:

Answer the question appropriately using multiple tools to verify your answer

Query: {input}

{agent_scratchpad}

"""

)

Calling our agent

```
agent = create_tool_calling_agent(bedrock_llm, tools, prompt)
agent_executor = AgentExecutor(agent=agent, tools=tools, verbose=True, handle_parsing_errors=True)
response = agent_executor.invoke({"input": agent_input})
raw_output = response['output'][0]['text']
```

Memory

```
# Create a runnable with message history
agent_with_chat_history = RunnableWithMessageHistory(
    chain,
    get_redis_history,
    input_messages_key="input",
    history_messages_key="history"
)

response1 = agent_with_chat_history.invoke(
    {"input": "What's the population of Canada?"},
    config={"configurable": {"session_id": session_id}},
)

print("AI Response1:", response1.content)

response2 = agent_with_chat_history.invoke(
    {"input": "What is their anthem?"},
    config={"configurable": {"session_id": session_id}},
)

print("AI Response2:", response2.content)
```

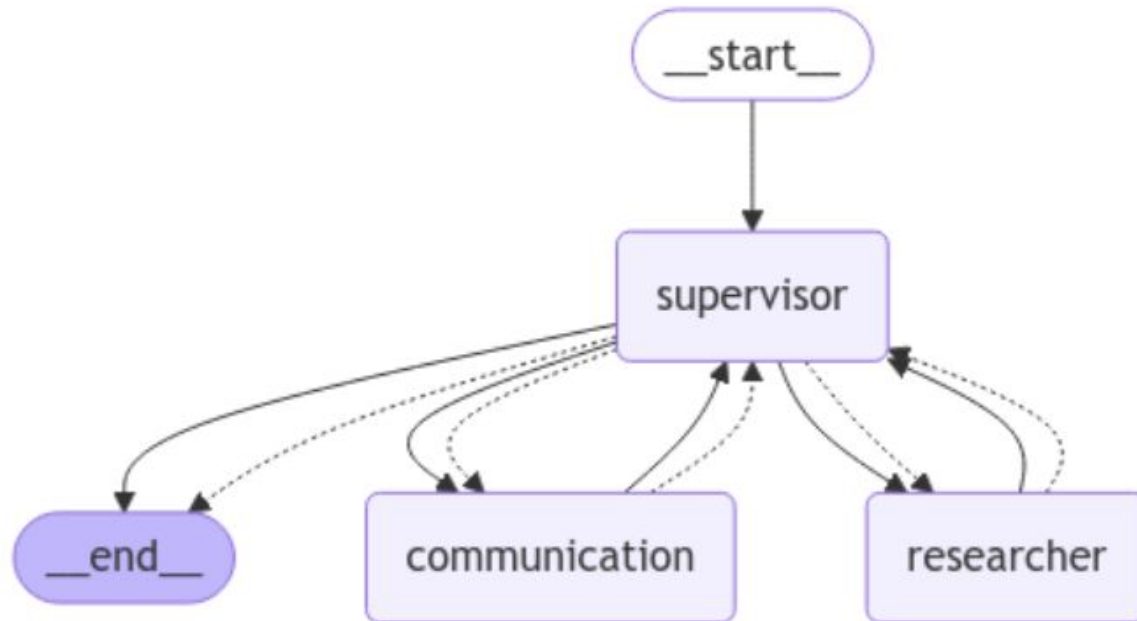
AI Response1: The current population of Canada is estimated to be around 38,915,741 as of January 1, 2023 according to the latest data from Statistics Canada.

AI Response2: The national anthem of Canada is "O Canada".

Guardrails

```
# Create the guardrail response separately
def create_guardrail_response():
    response = bedrock.create_guardrail(
        name="guardrail_" + str(int(time.time())),
        description="Guardrails for customer support agent.",
        topicPolicyConfig={
            'topicsConfig': [
                {
                    "name": "Internal Supplier Information",
                    "definition": "Information relating to this or other suppliers that is only available through internal systems.",
                    "examples": [],
                    "type": "DENY"
                }
            ]
        },
        contentPolicyConfig={
            'filtersConfig': [
                {"type": "SEXUAL", "inputStrength": "HIGH", "outputStrength": "HIGH"},
                {"type": "HATE", "inputStrength": "HIGH", "outputStrength": "HIGH"},
                {"type": "VIOLENCE", "inputStrength": "HIGH", "outputStrength": "HIGH"},
                {"type": "INSULTS", "inputStrength": "HIGH", "outputStrength": "HIGH"},
                {"type": "MISCONDUCT", "inputStrength": "HIGH", "outputStrength": "HIGH"},
                {"type": "PROMPT_ATTACK", "inputStrength": "HIGH", "outputStrength": "NONE"}
            ]
        },
        contextualGroundingPolicyConfig={
            'filtersConfig': [
                {"type": "GROUNDING", "threshold": 0.7},
                {"type": "RELEVANCE", "threshold": 0.7}
            ]
        },
        blockedInputMessaging="Sorry, the model cannot answer this question.",
        blockedOutputsMessaging="Sorry, the model cannot answer this question."
    )
    return response
```

Multi-agent system



04

Final agent

ReAct

- Thought
- Action
- Observation

Example - replenishment of Christmas pudding

```
Invoking: `Retriever_tool` with `{'query': 'christmas pudding'}`  
responded: [{'type': 'tool_use', 'name': 'Retriever_tool', 'id': 'tooluse_LiZRguTSTQ56u1BZ090r1w', 'index': 0, 'input': '{"query": "christmas pudding"}'}]
```

Item: Christmas Pudding

Jan: 0

Feb: 0

Mar: 0

Apr: 0

May: 0

Jun: 0

Jul: 0

Aug: 0

Sep: 0

Oct: 0

Nov: 0

Dec: 20

Brand_A_Jan: 0

Brand_A_Feb: 0

Brand_A_Mar: 0

Brand_A_Apr: 0

Brand_A_May: 0

Brand_A_Jun: 0

Brand_A_Jul: 0

Brand_A_Aug: 0

Brand_A_Sep: 0

Brand_A_Oct: 0

Brand_A_Nov: 0

Brand_A_Dec: 12

Item: Christmas Pudding

Used_Last_Month: 5

Left_This_Month: 2

Brand_A_Used_Last_Month: 3

Brand_A_Left_This_Month: 0

Brand_B_Used_Last_Month: 1

Brand_B_Left_This_Month: 2

Brand_C_Used_Last_Month: 1

Brand_C_Left_This_Month: 0

Item: Christmas Pudding

Brand_A: ["Not sure I'd buy it again.", 'Could have been better in some ways.', 'A bit disappointed for.', 'Had some issues, not thrilled.']

Brand_B: ['Decent for the price.', 'It was okay, nothing special.', 'Met my expectations, no more bad, but could improve.']

Brand_C: ['An average experience overall.', 'Decent for the price.', 'It was okay, nothing special bad, but could improve.']

```
Invoking: `Datetime` with `{'Arg1': 'Current datetime'}`  
responded: [{'type': 'tool_use', 'name': 'Datetime', 'id': 'tooluse_oTnThLs-R2uWmiqntemzcw', 'index': 0, 'input': '{"Arg1": "Current datetime"}'}]
```

```
2024-12-12T10:44:54.206241[{'type': 'text', 'text': 'Based on the internal data provided, here is my analysis and recommendation for ordering Christmas pudding:\n\nShould this item be ordered?\nYes, Christmas pudding should definitely be ordered for the upcoming holiday season.\n\nThe current date is December 12, 2024. \n\nLooking at the stock statistics:\n- Last month (November), there were only 2 Christmas puddings left in stock. \n- In December of last year, a total of 20 Christmas puddings were sold.\n- Brand A sold 12 units last December, Brand B sold 8 units.\n\nWith Christmas coming up later this month, demand will likely be high similar to last year. To ensure sufficient stock:\n\n(a) For this month (December 2024):\nBased on last year's sales of 20 units, I recommend ordering 25 Christmas puddings to account for any increase in demand. Split across brands:\nBrand A: 15 units\nBrand B: 10 units \n\n(b) For next month (January 2025): \nIt's unlikely much Christmas pudding will be needed after the holiday, so I recommend a small order of 5 units total to have some remaining stock.\nBrand A: 3 units\nBrand B: 2 units\n\nRegarding brand selection, the customer reviews indicate:\nBrand A had more negative reviews like "disappointing", "issues", while Brand B and Brand C had more neutral/positive reviews like "decent", "met expectations".\n\nSo I would order a higher proportion from Brands B and C which seem to be preferred over Brand A.\n\nHere is a draft email to a supplier to place this order:\n\nSubject: Christmas Pudding Order for December 2024/January 2025\n\nDear Supplier,\n\nFor the upcoming holiday period, we would like to place the following order for Christmas puddings to be delivered by December 18th:\n\nBrand B Christmas Puddings - 10 units \nBrand C Christmas Puddings - 10 units\nBrand A Christmas Puddings - 5 units\n\nAdditionally, for January 2025, we require:\n\nBrand B Christmas Puddings - 2 units\nBrand C Christmas Puddings - 2 units \nBrand A Christmas Puddings - 1 unit\n\nPlease confirm pricing and delivery details.\n\nThank you,\n\nReply Auto Replenishment', 'index': 0}]
```

Example response - christmas pudding

Based on the internal data provided, here is my analysis and recommendation for ordering Christmas puddings:

Should this item be ordered?

Yes, Christmas pudding should definitely be ordered as it is a seasonal item that sees high demand around the Christmas holiday period.

Current date:

The current date is December 11, 2024, so we are right in the middle of the holiday season when Christmas pudding will be in high demand.

Stock levels:

According to the data, last month (November) there were only 2 Christmas puddings left in stock after selling 5 units. This indicates the stock was very low going into the peak demand period of December.

Last year's sales:

Looking at last year's data, a total of 20 Christmas puddings were sold in December across the three brands (12 of Brand A, 8 of Brand B, and 0 of Brand C).

Factoring in holidays:

Since we are just a couple weeks before Christmas, the biggest demand spike will occur very soon. We should order enough to cover this Christmas period as well as some extras to last through early January when demand will remain elevated after the holiday.

Recommendation:

a) For this month (December), I recommend ordering 25 Christmas puddings total to cover expected demand. Based on last year's brand split, the order should be:

Brand A: 15 units

Brand B: 10 units

Brand C: 0 units (did not sell well last year)

b) For next month (January 2025), I would recommend a smaller order of 5-10 units total to have some stock for any lingering post-holiday demand.

Customer reviews:

Based on the customer review data provided, Brand B seems to be the best reviewed of the three brands, described as "decent for the price" and "met expectations". Brand A had some negative reviews like "disappointing" and "had issues". So I would prioritize ordering more of Brand B over Brand A if we anticipate running out of stock.

Math explanation:

For the December order quantity, I took last year's total sales of 20, and increased it by 25% to 25 units to account for any increase in demand this year. I then split those 25 units across brands using roughly the same percentages as last year's brand sales (60% Brand A, 40% Brand B).

For January, I estimated about 25% of the typical December order ($25 * 0.25 = 6.25$) which I rounded up to 10 units as a conservative number.

Draft email to supplier:

Subject: Christmas Pudding Order for Reply Auto Replenishment

Email draft

Draft email to supplier:

Subject: Christmas Pudding Order for Reply Auto Replenishment

Dear Supplier,

I am writing to place an order for Christmas puddings on behalf of Reply Auto Replenishment. Based on our sales forecasting data, we would like to order the following for delivery by December 18th:

Brand B Christmas Puddings: 10 units

Brand A Christmas Puddings: 15 units

Additionally, we request a follow-up order for early January 2025 as follows:

Brand B Christmas Puddings: 5 units

Brand A Christmas Puddings: 5 units

Please confirm pricing, delivery dates and any other details at your earliest convenience. We expect this to be a highly demanded product through the end of the year.

Thank you,

Reply Auto Replenishment

05

Challenges and lessons

Challenges

- Brainstorming ideas - industry research
- OpenAI vs Bedrock
- Deprecated features and beta features
- Coding own tools
- Simulating data / reviews
- Memory
- Guardrails
- Multi-agent system

What we've learnt

- Uses of an agent in industry
- How to set up an agent
- Learning on the job
- Debugging
- Perseverance and creativity
- Collaboration

Questions?

Appendix

Example responses - sunscreen

Based on the internal data provided, here is my analysis and recommendation for ordering sunscreen:

1. The data shows that sunscreen sales peak in the summer months of June, July and August. With the current date being December 11th, we are heading into the winter season where sunscreen demand will be very low.

2. Looking at last month's (November) data:

- Total sunscreen units sold: 20
- Units left in stock this month: 5

3. For the brand breakdown last month:

- Brand A: 6 units sold, 1 unit left
- Brand B: 13 units sold, 3 units left
- Brand C: 1 unit sold, 1 unit left

4. Reviewing the customer reviews:

- Brand A has the most positive reviews
- Brand B has average/decent reviews
- Brand C has more negative reviews

5. Based on the low demand in winter, the remaining stock levels, and the better customer reviews for Brand A, I would recommend:

- a) For this month (December): Do not order any additional sunscreen units. The 5 units left should cover any limited demand.
- b) For next month (January): Order 50 units of Brand A sunscreen to have stock ready for when demand starts picking up in spring/summer.

6. Looking further ahead, order quantities should increase for the key summer months based on historical data:

- April: Order 100 units (60 Brand A, 40 Brand B)
- May: Order 150 units (100 Brand A, 50 Brand B)
- June: Order 250 units (150 Brand A, 100 Brand B)

The math behind these recommendations:

- December & January orders based on 5 units left + typical low winter demand
- Spring/summer orders calculated as 125% of previous year's demand for those months, favoring Brand A due to better reviews

Email drafts - 1

Here is a draft email to a Brand A suncream supplier:

Subject: Suncream Order for January

Dear Supplier,

I am writing to place an order for 50 units of your Brand A suncream for delivery in early January. Based on our sales forecasting data, we anticipate demand will start increasing next month as we head into the spring season.

Our customers have provided excellent reviews for your Brand A suncream compared to other brands we carry. We would like to ensure we have enough inventory of this top-performing product heading into the busy spring and summer months.

Please confirm pricing and delivery timeline for this 50 unit order to be shipped in early January. We plan to place further orders ramping up quantities for April through June in preparation for peak suncream demand.

Thank you,
Reply Auto Replenishment

What is RAG?

- What is RAG
 - Retrieval-Augmented Generation - process of optimisation through referencing external data outside of the training set
 - Cost-effective
 - Up-to-date
 - Provides citations
 - More accurate - tackles hallucination by redirecting LLM to retrieve information from trusted sources.

More statistics

- **Digital Automation in Replenishment:** A global survey revealed that only 8% of retailers have implemented real-time automated processes for demand planning and forecasting. Additionally, 12% utilize machine learning technologies in their replenishment activities. [Statista](#)
- **Impact of Stockouts:** Stockouts, or out-of-stock events, significantly affect retailers' profitability. Studies indicate that the global inventory distortion due to out-of-stock grocery products was valued at over \$500 billion in 2020. [Statista](#)
- **Causes of Stockouts:** Approximately 70-90% of stockouts result from defective shelf replenishment practices, while the remaining 10-30% stem from upstream supply chain issues, such as supplier shortages. [Wikipedia](#), McKinnon, A.C., Mendes, D. and Nabateh, M. (2007), "In-store logistics: an analysis of on-shelf availability and stockout response for three product groups", International Journal of Logistics: Research and Applications, Vol. 10 No. 3, pp. 251-68.
- **Shrinkage Rates:** Shrinkage, which includes losses from theft, administrative errors, and damaged goods, poses a significant challenge to retail replenishment efficiency. In 2008, U.S. retailers experienced an average shrinkage rate of 1.52% of sales, translating to substantial financial losses. [Wikipedia](#)
- **Gross Margin Return on Inventory Investment (GMROI):** This metric assesses the profitability of inventory. A GMROI above 3.0 is considered favorable for retailers, indicating a good balance between sales, margin, and inventory cost. [Wikipedia](#)
- **Inventory Inaccuracies:** Inaccurate inventory records can cost retailers approximately 4% in annual sales. Detecting and correcting these inaccuracies manually is often cost-prohibitive, and existing algorithmic solutions may fall short in dynamic retail environments. [arXiv](#)
- **Shrinkage Rates:** Shrinkage, encompassing losses from theft, administrative errors, and damaged goods, poses a significant challenge to retail replenishment efficiency. In 2008, U.S. retailers experienced an average shrinkage rate of 1.52% of sales, translating to substantial financial losses. [Wikipedia](#)
- **Stockouts:** Stockouts, or out-of-stock events, significantly affect retailers' profitability. Studies indicate that the global inventory distortion due to out-of-stock grocery products was valued at over \$500 billion in 2020. [Wikipedia](#)
- **Food Waste in Retail:** Retail stores often discard large quantities of food items that have reached their best-before, sell-by, or use-by dates, contributing to significant food waste. In the United States, some of the estimated six billion pounds of produce wasted each year are discarded because of appearance standards. [Wikipedia](#)

FOR A MEDIUM TO LARGE RETAIL COMPANY

Assumptions

1. **Number of branches per company:** 500 (a medium-to-large UK retail chain like Sainsbury's or Tesco).
2. **Average management staff per branch:** 7 (including store managers, assistant managers, and department heads).
3. **Average salary per management staff:** £32,500 per year.
4. **Total annual salary costs per branch:** $7 \times £32,500 = £227,500$
 $7 \times £32,500 = £227,500$ $7 \times £32,500 = £227,500$
5. **Total salary costs for management across all branches:**
 $500 \times £227,500 = £113,750,000$
 $500 \times £227,500 = £113,750,000$ $500 \times £227,500 = £113,750,000$

Potential Costs with LLM Replacement:

1. **LLM deployment per branch:** Estimated initial setup (hardware, software, and integration) cost of £50,000 per branch. $500 \times £50,000 = £25,000,000$
 $500 \times £50,000 = £25,000,000$ $500 \times £50,000 = £25,000,000$
2. **Ongoing maintenance costs:** Estimated at **10% of the setup cost annually**. $10\% \times £25,000,000 = £2,500,000$ annually.
 $10\% \times £25,000,000 = £2,500,000$ annually. $10\% \times £25,000,000 = £2,500,000$ annually

First-Year Costs:

Initial setup: £25,000,000 + Ongoing maintenance: £2,500,000 = £27,500,000

Net Savings in First Year:

Current management costs: £113,750,000 - £27,500,000 = £86,250,000

Net Savings in Subsequent Years:

Annual salary savings: £113,750,000 - Maintenance: £2,500,000 = £111,250,000