



#PuneDevCon

Pune DevCon 2025»





LangChain 101: Building Intelligent Agents with LLMs and Tools

Dr. Yogesh Haribhau Kulkarni

LANGCHAIN 101

Yogesh Haribhau Kulkarni



Outline

① INTRODUCTION

② IMPLEMENTATION

③ FRAMEWORK

④ WHAT'S NEW

⑤ CONCLUSIONS

⑥ REFERENCES

About Me



Yogesh Haribhau Kulkarni

Bio:

- ▶ 20+ years in CAD/Engineering software development
- ▶ Got Bachelors, Masters and Doctoral degrees in Mechanical Engineering (specialization: Geometric Modeling Algorithms).
- ▶ Currently doing Coaching in fields such as Data Science, Artificial Intelligence Machine-Deep Learning (ML/DL) and Natural Language Processing (NLP).
- ▶ Feel free to follow me at:
 - ▶ Github (github.com/yogeshhk)
 - ▶ LinkedIn (www.linkedin.com/in/yogeshkulkarni/)
 - ▶ Medium (yogeshharibhaukulkarni.medium.com)
 - ▶ Send email to [yogeshkulkarni at yahoo dot com](mailto:yogeshkulkarni@yahoo.com)



Office Hours:
Saturdays, 2 to 5pm
(IST); Free-Open to all;
email for appointment.

Introduction to LangChain



What is LangChain?

A Framework for Building LLM Applications

Key Capabilities:

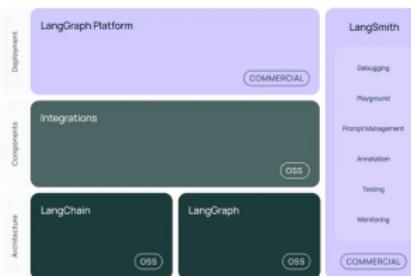
- ▶ Connect LLMs to external data (RAG)
- ▶ Build intelligent agents
- ▶ Manage prompts & memory
- ▶ Switch between models easily

Quick Facts:

- ▶ Open Source (MIT License)
- ▶ Python & JavaScript
- ▶ Created by Harrison Chase



From Idea to Production



(Ref: Getting Started with LangChain: A Beginner's Guide)

Why You Need LangChain

The Challenge:

- ▶ LLMs lack memory
- ▶ No access to your data
- ▶ Can't use external tools
- ▶ Complex integration code

The Solution:

- ▶ Memory management built-in
- ▶ Easy data integration
- ▶ Tool calling framework
- ▶ Simple, modular code

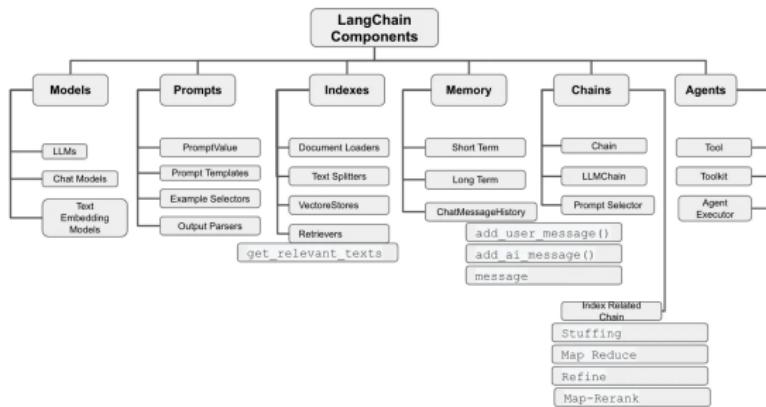


(Ref: <https://www.leanware.co/insights/langchain-vs-pinecone-which-should-you-choose>)

Bottom Line:

Build AI apps faster
with less code

Core Building Blocks



Foundation:

- ▶ **Models:** Connect to any LLM
- ▶ **Prompts:** Dynamic templates
- ▶ **Memory:** Conversation context

Advanced:

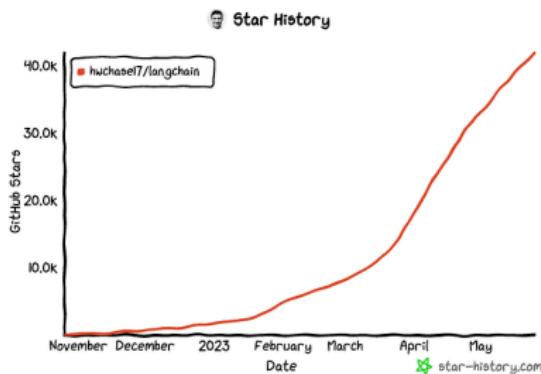
- ▶ **Retrievers:** Search your data
- ▶ **Agents:** Autonomous reasoning
- ▶ **LCEL:** Chain components easily

Mix and match components to build your app



So, Why LangChain?

- ▶ **RAG Applications:** Connect LLMs to your data
- ▶ **Intelligent Agents:** Autonomous tool usage
- ▶ **Production Ready:** Monitoring & logging built-in
- ▶ **Model Flexibility:** Switch providers easily
- ▶ **Modular Design:** Use only what you need



(Ref: LangChain tutorial: Build an LLM-powered app)

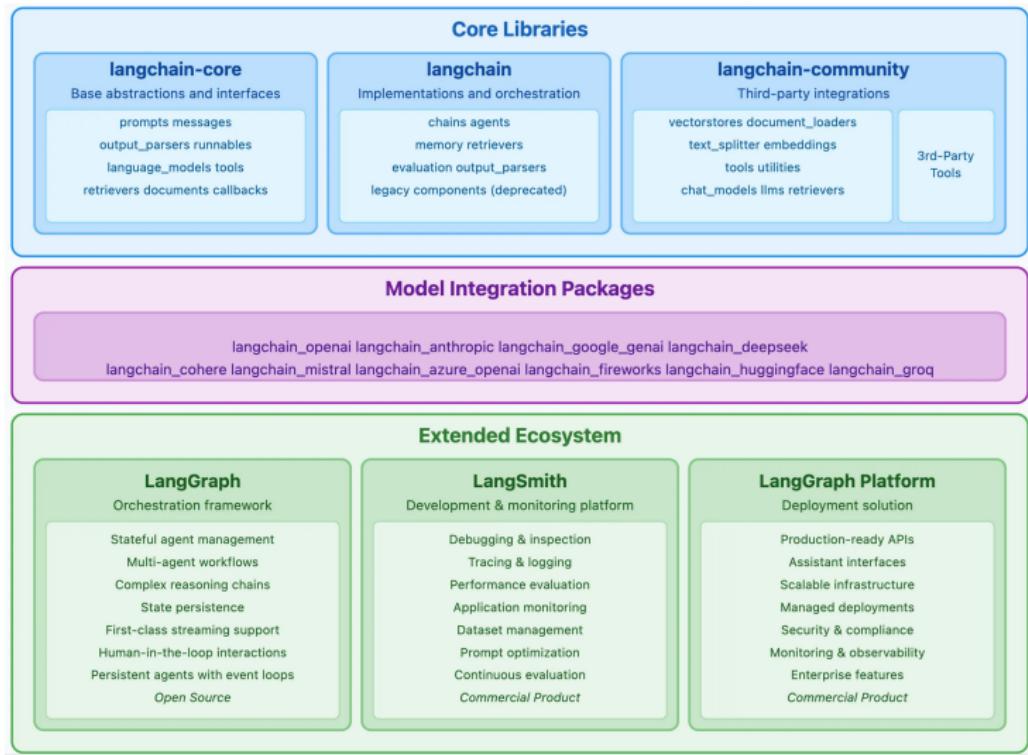
The LangChain Community

- ▶ 80,000+ GitHub stars
- ▶ 2,000+ contributors
- ▶ Millions of downloads/month
- ▶ Active Discord community
- ▶ MIT - Commercial friendly

(Ref: What is Langchain and why should I care as a developer? - Logan Kilpatrick)



The LangChain Ecosystem



(Ref: Demystifying the LangChain Ecosystem for LLM-Powered Application Development - Witsankalpa)

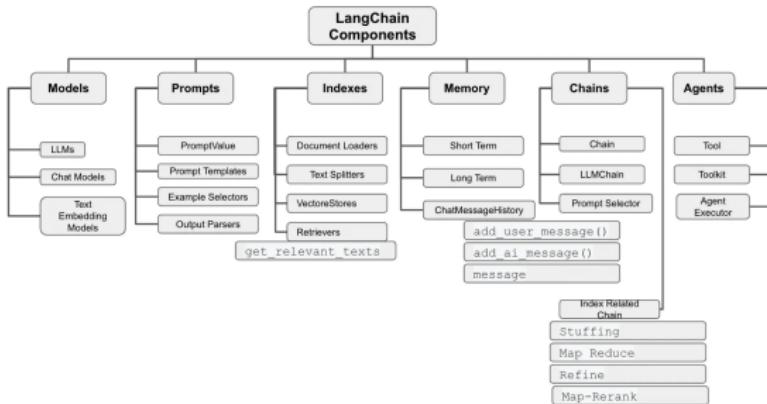
Installation & Setup

Python 3.10 or higher

```
1 # Core packages (LangChain 1.0 focuses on the 'langchain' agent framework)
2 pip install -U langchain-langchain-core
3
4 # Provider-specific packages (Groq is a first-class integration)
5 pip install -U langchain-groq
6
7 # Document processing (Now separated for better dependency management)
8 pip install -U langchain-community langchain-text-splitters
9
10 # Embeddings & Vector Stores
11 pip install -U langchain-huggingface langchain-chroma
12
13 # NEW: Legacy features
14 # Use this ONLY if you need old chains like LLMChain or RetrievalQA
15 pip install -U langchain-classic
16
17 //API Keys Setup:
18
19 import os
20 os.environ["GROQ_API_KEY"] = "your-groq-api-key-here"
21 # Or use .env file with python-dotenv
22
23 # LangSmith Tracing (Highly recommended for v1.0 Agents)
24 os.environ["LANGCHAIN_TRACING_V2"] = "true"
25 os.environ["LANGCHAIN_API_KEY"] = "your-langsmith-key"
26 os.environ["LANGCHAIN_PROJECT"] = "my-v1-project"
```



Core Components Overview



Component	Purpose	Use Case
Models	Connect to LLMs	Text generation, chat
Prompts	Template management	Dynamic prompts
LCEL	Chain components	Build workflows
Memory	Store conversations	Chatbots
Retrievers	Search data	RAG, Q&A
Agents	Tool selection	Autonomous tasks

(Ref: How LangChain Makes Large Language Models More Powerful)

Key Concepts: Chains vs Agents

- ▶ **Chains (LCEL):**
 - ▶ Predetermined sequence of operations
 - ▶ Composed with pipe operator: `prompt | llm | parser`
 - ▶ Fixed execution path
 - ▶ Best for: Structured, predictable workflows
- ▶ **Agents:**
 - ▶ Dynamic decision-making with LLM reasoning
 - ▶ Choose tools based on input
 - ▶ Adaptive execution path
 - ▶ Best for: Complex, unpredictable scenarios

When to use what?

- ▶ Use **Chains/LCEL** when: Steps are known, workflow is fixed
- ▶ Use **Agents** when: Need dynamic tool selection, multi-step reasoning

(Ref: Superpower LLMs with Conversational Agents)

What Can You Build?

RAG Applications:

- ▶ Document Q&A
- ▶ Knowledge base search
- ▶ Semantic search

Data Analysis:

- ▶ SQL query generation
- ▶ Report generation
- ▶ Data insights

Conversational AI:

- ▶ Chatbots with memory
- ▶ Customer support
- ▶ Personal assistants

(Ref: LangChain Use Cases Documentation)

Autonomous Agents:

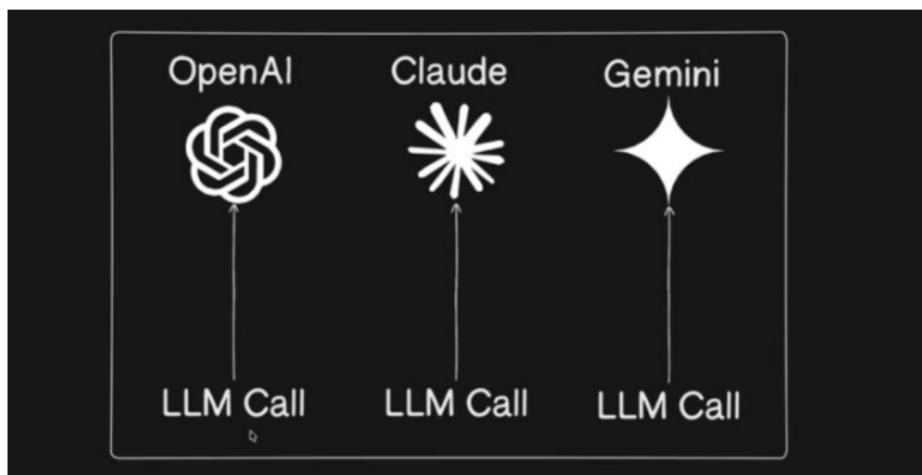
- ▶ Web research
- ▶ API integration
- ▶ Multi-step workflows

Implementations

Use as LLM



Diff LLMs, Diff Calls



(Ref: What is LangChain - Yash Jain)

Langchain provides abstraction. No LLM specific API calls, but a generic way.
Need to install specific extension though and have to have respective KEYS.

```
2 pip install -U langchain langchain--openai langchain--anthropic langchain--google--genai langchain--grou
```

YHK

Common Way of Calling Diff LLMs: V0

```
from langchain.openai import ChatOpenAI
from langchain.groq import ChatGroq
from langchain.anthropic import ChatAnthropic
from langchain.google_genai import ChatGoogleGenerativeAI
from langchain.core.messages import HumanMessage
gpt_model = ChatOpenAI(
    model="gpt-4o",
    temperature=0.7,
    max_tokens=500)
groq_model = ChatGroq(
    model="llama-3.3-70b-versatile",
    temperature=0.2,
    # Groq-specific param example:
    # reasoning_format="raw" )
claude_model = ChatAnthropic(
    model="claude-3-5-sonnet-latest",
    timeout=None,
    stop_sequences=["\n\nHuman:"])
gemini_model = ChatGoogleGenerativeAI(
    model="gemini-1.5-pro",
    convert_system_message_to_human=True # Legacy helper for older models)
def run_demo(model, provider_name):
    message = [HumanMessage(content="Explain quantum entanglement in one sentence.")]
    response = model.invoke(message)
    print(f"Response: {response.content}\n")
run_demo(gpt_model, "OpenAI")
```



Common Way of Calling Diff LLMs: V1

```
2 import os
3 from langchain.chat_models import init_chat_model
4 from langchain.core.messages import HumanMessage, SystemMessage
5
6 # os.environ["OPENAI_API_KEY"] = "sk-..."
7 # os.environ["ANTHROPIC_API_KEY"] = "sk-ant-..."
8 # os.environ["GOOGLE_API_KEY"] = "..."
9 # os.environ["GROQ_API_KEY"] = "gsk-..."
10
11 def get_response(provider_string, user_query):
12     """
13         Initializes a model based on the provider string: 'openai', 'anthropic', 'google', or 'groq'.
14     """
15     llm = init_chat_model(provider_string, temperature=0)
16
17     messages = [
18         SystemMessage(content="You are a helpful research assistant."),
19         HumanMessage(content=user_query) ]
20     return llm.invoke(messages)
21
22 gpt_response = get_response("openai:gpt-4o", "What is LangChain v1?")
23 print(f"GPT-4o: {gpt_response.content[:100]}...")
24
25 claude_response = get_response("anthropic:claude-3-5-sonnet-latest", "What is LangChain v1?")
26 print(f"Claude: {claude_response.content[:100]}...")
27
28 gemini_response = get_response("google-genai:gemini-1.5-pro", "What is LangChain v1?")
29 print(f"Gemini: {gemini_response.content[:100]}...")
30
31 groq_response = get_response("groq:llama-3.3-70b-versatile", "What is LangChain v1?")
32 print(f"Groq: {groq_response.content[:100]}...")
```

Sentiment Analysis: Zero Shot

```
from langchain.groq import ChatGroq
from langchain.core.prompts import ChatPromptTemplate
from langchain.core.output_parsers import StrOutputParser
# 1. Define the LLM
model = ChatGroq(model.name="llama-3.3-70b-versatile", temperature=0)

# 2. Create a prompt template
prompt = ChatPromptTemplate.from_template(
    "Analyze the sentiment of the following text."
    "Respond with only one word: Positive, Negative, or Neutral.\n"
    "Text: {input_text}"
)
# 3. Create a simple output parser
parser = StrOutputParser()

# 4. Build the LCEL chain
chain = prompt | model | parser

# Define input text
input_text = "I love LangChain! It's the best NLP library I've ever used."

# 5. Invoke the chain
sentiment = chain.invoke({"input_text": input_text})

# Print the sentiment
print(sentiment)
```



Named Entity Recognition (NER): Zero Shot

```
from langchain.groq import ChatGroq
from langchain.core.prompts import ChatPromptTemplate
from langchain.core.output_parsers import StrOutputParser
# 1. Setup Model
model = ChatGroq(model="llama-3.3-70b-versatile")

# 2. Create Prompt for NER extraction
template = """
Identify the named entities (Person, Organization, Amount) in the text.
Format the output as a bulleted list.

Text: {text}
"""
prompt = ChatPromptTemplate.from_template(template)
# 3. Build Chain
chain = prompt | model | StrOutputParser()

# 4. Invoke
input_text = "Microsoft is acquiring Nuance Communications for $19.7 billion."
entities = chain.invoke({"text": input_text})
print(entities)
```

Use as Chains



LCEL: Basic Example

```
1 # Old Pattern (Deprecated):
2 from langchain.llms import OpenAI
3 from langchain.chains import LLMChain
4
5 llm = OpenAI()
6 chain = LLMChain(llm=llm, prompt=prompt)
7 result = chain.run("input")
8
9 # Modern LCEL Pattern with Groq:
10
11 from langchain.groq import ChatGroq
12 from langchain.core.prompts import ChatPromptTemplate
13 from langchain.core.output_parsers import StrOutputParser
14
15 # Set up Groq model, e.g., Gemma or Llama 3
16 llm = ChatGroq(model.name="gemma-7b-it")
17 prompt = ChatPromptTemplate.from_template("Tell me about {topic}")
18 output_parser = StrOutputParser()
19
20 # Build chain with pipe operator
21 chain = prompt | llm | output_parser
22
23 # Invoke the chain
24 result = chain.invoke({"topic": "LangChain"})
25 print(result)
```

LCEL: Advanced Features

Streaming Support:

```
1 import asyncio
2
3 async def main():
4     # Stream tokens as they're generated (Works in standard functions)
5     for chunk in chain.stream([{"topic": "AI"}]):
6         print(chunk, end="", flush=True)
7         print("\n")
8
9     # Async invocation (Must be inside async function)
10    result = await chain.invoke([{"topic": "AI"}])
11    print(f"Result: {result}\n")
12
13    # Async streaming
14    async for chunk in chain.astream([{"topic": "AI"}]):
15        print(chunk, end="", flush=True)
16        print("\n")
17
18    # Process multiple inputs in parallel (Sync version)
19    results = chain.batch([
20        {"topic": "AI"}, {"topic": "ML"}, {"topic": "LangChain"}])
21
22    print(f"Batch Results: {results}")
23
24 if __name__ == "__main__":
25     asyncio.run(main())
```



LCEL: Complex Chains

Multi-step Chain with RunnablePassthrough:

```
1 from langchain.core.runnables import RunnablePassthrough
2
3 chain = (
4     {"context": retriever, "question": RunnablePassthrough()}
5     | prompt
6     | llm
7     | output_parser
8 )
9
10 result = chain.invoke("What is LangChain?")
11 print(result)
12
13 # Parallel Execution with RunnableParallel:
14 from langchain.core.runnables import RunnableParallel
15
16 chain = RunnableParallel(
17     summary=prompt1 | llm | output_parser,
18     keywords=prompt2 | llm | output_parser
19 )
20
21 result = chain.invoke({ "text": "Long document..." })
22 print(result)
23 # Returns: { "summary": "...", "keywords": "..." }
```

Use as RAG



Complete RAG Application: A Quick Start Example

```
1 # imports
3
3 loader = PyPDFLoader("data/ag-studio.pdf")
4 docs = loader.load()
5
6 splitter = RecursiveCharacterTextSplitter(chunk_size=1000, chunk_overlap=100)
7 splits = splitter.split_documents(docs)
8
9 embeddings = HuggingFaceEmbeddings(model_name="all-MiniLM-L6-v2")
10 vectorstore = Chroma.from_documents(documents=splits, embedding=embeddings, collection_name="local-pdf-rag")
11 retriever = vectorstore.as_retriever(search_kwargs={"k": 3})
12
13 prompt = ChatPromptTemplate.from_template("""Answer the question based ONLY on the following context:
14 {context} Question: {question}""")
15
16 model = init_chat_model("llama-3.3-70b-versatile", model_provider="gradio", temperature=0)
17
18 # Note: retriever | format_docs ensures the model gets text, not Document objects
19 def format_docs(docs):
20     return "\n\n".join(doc.page_content for doc in docs)
21
22 chain = ({"context": retriever | format_docs, "question": RunnablePassthrough()}
23         | prompt | model | StrOutputParser())
24
25 response = chain.invoke("What is the main topic of this document?")
26 print(response)
```



Use as Agent

Example: Modern Agent in LangChain v1

```
from langchain.agents import create_agent
from langchain.chat_models import init_chat_model
from langchain.core.tools import tool

@tool
def get_weather(city: str) -> str:
    """Get weather for a given city."""
    # Add a print here so you can see when the function actually runs!
    print(f"--- Executing get_weather for {city} ---")
    return f"It's always sunny in {city}!"

llm = init.chat_model("llama-3.3-70b-versatile", model_provider="qroq")

# In v1.0, create_agent creates a compiled graph.
# We need to make sure the agent is permitted to call tools.
agent = create_agent(
    model=llm,
    tools=[get_weather],
    system_prompt="You are a helpful assistant. If you need weather info, use the tool."
)

# Use the invoke method
inputs = {"messages": [{"user": "What is the weather in Pune?"}]}
response = agent.invoke(inputs)

# In v1.0, the response contains the full message history of the turn.
# The last message should now be the assistant's final answer AFTER the tool result.
for msg in response["messages"]:
    msg.pretty_print()
```

LangChain Framework Components

Framework Architecture

Chains: The core of LangChain. Components (and even other chains) can be stringed together to create *chains*.

Prompt templates: Prompt templates are templates for different types of prompts. Like “chatbot” style templates, ELI5 question-answering, etc

LLMs: Large language models like GPT-3, BLOOM, etc

Indexing Utils: Ways to interact with specific data (embeddings, vectorstores, document loaders)

Tools: Ways to interact with the outside world (search, calculators, etc)

Agents: Agents use LLMs to decide what actions should be taken. Tools like web search or calculators can be used, and all are packaged into a logical loop of operations.

Memory: Short-term memory, long-term memory.

Core Building Blocks:

- ▶ **Models:** LLMs, Chat Models, Embeddings
- ▶ **Prompts:** Dynamic template management
- ▶ **Output Transformers:** Structured output extraction
- ▶ **Retrievers:** Document and data access
- ▶ **Memory:** Conversation state persistence
- ▶ **Agents & Tools:** Dynamic reasoning and actions

Models

Models in LangChain

Three Types of Models:

- ▶ **LLMs (Large Language Models):**
 - ▶ Input: String (prompt)
 - ▶ Output: String (completion)
 - ▶ Examples: GPT-4, Claude, Gemma, Llama 3, Mixtral.
 - ▶ Use case: Text generation, completion
- ▶ **Chat Models:**
 - ▶ Input: List of messages
 - ▶ Output: Chat message
 - ▶ Examples: ChatGPT, Claude Chat, ChatGroq
 - ▶ Use case: Conversational AI
- ▶ **Embedding Models:**
 - ▶ Input: Text
 - ▶ Output: Vector (list of floats)
 - ▶ Examples: OpenAI Embeddings, HuggingFace, Sentence Transformers
 - ▶ Use case: Semantic search, similarity

Models Functionality

- ▶ Tool calling: calling external tools (like databases queries or API calls) and use results in their responses.
- ▶ Structured output: where the model's response is constrained to follow a defined format.
- ▶ Multimodality: process and return data other than text, such as images, audio, and video.
- ▶ Reasoning: models perform multi-step reasoning to arrive at a conclusion.

Models Usage

- ▶ With agents: Models can be dynamically specified when creating an agent.
- ▶ Standalone: Models can be called directly (outside of the agent loop) for tasks like text generation, classification, or extraction without the need for an agent framework.

Model Integration: Modern Syntax

For standalone model `init_chat_model`

```
from langchain.chat_models import init_chat_model # Still valid in v1
2 # OR use provider-specific import (recommended)
from langchain.groq import ChatGroq
4 from langchain.messages import HumanMessage, AIMessage, SystemMessage

6
# Initialize Groq LLM (ensure GROQ_API_KEY is set in your environment)
8 model = ChatGroq(model="llama-3.3-70b-versatile") # https://console.groq.com/docs/models

10 # # or assuming os.environ["ANTHROPIC_API_KEY"] = "sk-..."
# model = init.chat.model(
12 #   "claude-sonnet-4-5-20250929",
#   # Kwargs passed to the model:
14 #   temperature=0.7,
#   timeout=30,
16 #   max_tokens=1000,
# )
18
conversation = [
20   SystemMessage("You are a helpful assistant that translates English to French."),
  HumanMessage("Translate: I love programming."),
22  AIMessage("J'adore la programmation."),
  HumanMessage("Translate: I love building applications.")
24 ]
26 response = model.invoke(conversation)
print(response) # AIMessage("J'adore creer des applications.")
```



Multimodal

Certain models can process and return non-textual data such as images, audio, and video. You can pass non-textual data to a model by providing content blocks.

```
response = model.invoke("Create a picture of a cat")
2 print(response.content_blocks)
# [
4 #   {"type": "text", "text": "Here's a picture of a cat"},
#   {"type": "image", "base64": "...", "mime_type": "image/jpeg"},
6 # ]
```

Reasoning

Many models are capable of performing multi-step reasoning to arrive at a conclusion.

```
## Streaming
2 for chunk in model.stream("Why do parrots have colorful feathers?"):
    reasoning_steps = [r for r in chunk.content.blocks if r["type"] == "reasoning"]
    print(reasoning_steps if reasoning_steps else chunk.text)

6 ## Complete Output
response = model.invoke("Why do parrots have colorful feathers?")
8 reasoning_steps = [b for b in response.content.blocks if b["type"] == "reasoning"]
print(" ".join(step["reasoning"] for step in reasoning_steps))
```

Prompts



Prompts in LangChain

Prompt Management Strategies:

- ▶ **Prompt Templates:**
 - ▶ Parameterized templates with variables
 - ▶ Dynamic input insertion
 - ▶ Reusable prompt structures
- ▶ **Chat Prompt Templates:**
 - ▶ Multi-message conversations
 - ▶ System, human, AI message roles
 - ▶ Better for chat models
- ▶ **Few-Shot Prompts:**
 - ▶ Include example inputs/outputs
 - ▶ Guide model response style
 - ▶ Improve accuracy

Prompt Templates: Modern Examples

```
1 from langchain.core.prompts import PromptTemplate
2
3 # Use in chain with Groq
4 from langchain.groq import ChatGroq
5 from langchain.core.output_parsers import StrOutputParser
6
7 template = """You are a {role} assistant.
8 Task: {task}
9 Context: {context}
10 Provide a {format} response."""
11
12 prompt = PromptTemplate(
13     template=template,
14     input_variables=["role", "task", "context", "format"]
15 )
16
17 formatted = prompt.format(
18     role="helpful",
19     task="explain quantum computing",
20     context="for beginners",
21     format="simple"
22 )
23
24 chain = prompt | ChatGroq(model="llama-3.3-70b-versatile") | StrOutputParser()
25 result = chain.invoke({
26     "role": "helpful",
27     "task": "explain quantum computing",
28     "context": "for beginners",
29     "format": "simple"
30 })
31 print(result)
```



Chat Prompting

```
# Chat Prompt Template:  
2 from langchain.core.prompts import ChatPromptTemplate  
  
4 prompt = ChatPromptTemplate.from_messages([  
    ("system", "You are a {role}"),  
    ("human", "{input}"),  
    ("ai", "I understand. Let me help with that."),  
    ("human", "{follow_up}")  
])  
10  
11 chain = prompt | ChatGroq(model="llama-3.3-70b-versatile") | StrOutputParser()  
12 result = chain.invoke({  
    "role": "helpful",  
    "input": "Explain quantum computing",  
    "follow_up": "Make it simpler"  
14 })  
15 print(result)  
16 }
```

Document Loaders & Retrievers



Document Loading

Wide Range of Loaders:

- ▶ **Files:** PDF, Word, PowerPoint, CSV, Markdown
- ▶ **Web:** HTML, URLs, sitemaps, web scraping
- ▶ **Cloud:** S3, GCS, Google Drive, Notion
- ▶ **Databases:** SQL, MongoDB, Elasticsearch
- ▶ **Communication:** Email, Slack, Discord
- ▶ **Code:** GitHub, GitLab, Jupyter notebooks

Modern Document Processing:

```
1 from langchain.community.document_loaders import (
2     PyPDFLoader,
3     WebBaseLoader,
4     TextLoader)
5 from langchain.text_splitters import RecursiveCharacterTextSplitter
6
7 loader = PyPDFLoader("document.pdf")
8 documents = loader.load()
9
10 splitter = RecursiveCharacterTextSplitter(
11     chunk_size=1000,
12     chunk_overlap=200)
13 chunks = splitter.split_documents(documents)
```



Vector Stores & Retrievers

Popular Vector Stores:

- ▶ **Chroma**: Open-source, local-first, easy setup
- ▶ **Pinecone**: Managed service, production-ready
- ▶ **Weaviate**: GraphQL API, hybrid search
- ▶ **Qdrant**: High performance, filtering support
- ▶ **LanceDB**: Serverless, embedded option

Complete Example with Chroma & HuggingFace Embeddings:

```
# from langchain.community.embeddings import HuggingFaceEmbeddings
2 from langchain.huggingface import HuggingFaceEmbeddings
from langchain.community.vectorstores import Chroma
4 from langchain.text_splitters import RecursiveCharacterTextSplitter
from langchain.community.document_loaders import WebBaseLoader
6
loader = WebBaseLoader("https://example.com/article")
8 docs = loader.load()
10 splitter = RecursiveCharacterTextSplitter(chunk_size=1000, chunk_overlap=200)
splits = splitter.split_documents(docs)
12
vectorstore = Chroma.from_documents(
14     documents=splits,
15     embedding=HuggingFaceEmbeddings(model_name="all-MiniLM-L6-v2"))
16
retriever = vectorstore.as_retriever(search_kwargs={"k": 3})
```



Retrieval Strategies

Different Retrieval Methods:

```
1 # Returns the most semantically similar documents to the user
# query based purely on vector distance (top-k closest matches).
3 retriever = vectorstore.as_retriever(search_type="similarity",
    search_kwargs={"k": 4})
5
# MMR (Maximum Marginal Relevance) Balances relevance and diversity,
7 # ensuring the retrieved documents aren't repetitive by reducing semantic redundancy among the top-k results.
retriever = vectorstore.as_retriever(search_type="mmr",
9     search_kwargs={"k": 4, "fetch_k": 20})
11
# Filters out documents that fall below a minimum similarity score,
# ensuring only high-confidence matches are returned (still limited by k)
13 retriever = vectorstore.as_retriever(search_type = "similarity_score_threshold",
    search_kwargs={"score_threshold": 0.8, "k": 4})
15
# Manual search: Directly retrieves the top-k most similar documents
17 # without using a retriever wrapper same as search_type "similarity" but explicitly called.
results = vectorstore.similarity_search("What is machine learning?", k=3)
19
# With scores: Same as above but also returns the similarity
21 # score for each document, helping you inspect retrieval quality
results_with_scores = vectorstore.similarity_search_with_score(
    "What is machine learning?", k=3)
```

Chains with LCEL

Modern LangChain: LCEL (LangChain Expression Language)

What is LCEL?

- ▶ Modern, declarative way to build chains (introduced 2023)
- ▶ Uses pipe operator | to chain components
- ▶ Replaces old LLMChain pattern
- ▶ Built-in streaming, async, and batch support

Key Benefits:

- ▶ **Simplicity:** More readable and concise
- ▶ **Streaming:** Built-in streaming by default
- ▶ **Async:** Native async/await support
- ▶ **Observability:** Better tracing and debugging
- ▶ **Fallbacks:** Easy error handling and retries

(Ref: LangChain LCEL Documentation)



Modern Chains: LCEL Overview

What Changed?

- ▶ **Old:** LLMChain, SimpleSequentialChain (Deprecated)
- ▶ **New:** LCEL with pipe operator |

Core Runnables:

- ▶ RunnablePassthrough: Pass data through
- ▶ RunnableParallel: Execute in parallel
- ▶ RunnableLambda: Custom functions
- ▶ RunnableBranch: Conditional execution

LCEL: Basic Chain Patterns

```
1 // Simple Chain:  
2 from langchain.groq import ChatGroq  
3 from langchain.core.prompts import ChatPromptTemplate  
4 from langchain.core.output_parsers import StrOutputParser  
5  
6 prompt = ChatPromptTemplate.from_template("Tell me a joke about {topic}")  
7 model = ChatGroq(model.name="gemma-7b-it")  
8 output_parser = StrOutputParser()  
9  
10 chain = prompt | model | output_parser  
11 result = chain.invoke({"topic": "programming"})  
12  
13 // Chain with Multiple Steps:  
14 # Step 1: Generate topic  
15 topic_chain = (  
16     ChatPromptTemplate.from_template("Suggest a {genre} topic")  
17     | ChatGroq(model.name="gemma-7b-it")  
18     | StrOutputParser()  
19 )  
20  
21 # Step 2: Write content  
22 content_chain = (  
23     ChatPromptTemplate.from_template("Write a story about: {topic}")  
24     | ChatGroq(model.name="llama3-8b-8192")  
25     | StrOutputParser()  
26 )  
27  
28 # Combine: topic generates input for content  
29 full_chain = {"topic": topic_chain} | content_chain  
result = full_chain.invoke({"genre": "science fiction"})
```



LCEL: RAG Chain Pattern

Retrieval Augmented Generation:

```
from langchain.core.runnables import RunnablePassthrough
from langchain.groq import ChatGroq
# Assume 'retriever' from previous slide is defined
# Format documents
def format_docs(docs):
    return "\n\n".join(doc.page.content for doc in docs)
# RAG chain
rag_chain = (
    {
        "context": retriever | format_docs,
        "question": RunnablePassthrough()
    }
    | ChatPromptTemplate.from_template("""
        Answer the question based on the context:

        Context: {context}

        Question: {question}
    """)
    | ChatGroq(model_name="llama3-8b-8192")
    | StrOutputParser()
)
# Use it
answer = rag_chain.invoke("What is LangChain?")
```



LCEL: Parallel Execution

Benefits:

- ▶ Faster execution
- ▶ Clean code structure
- ▶ Easy to add/remove tasks

Run Multiple Chains in Parallel:

```
1 from langchain.core.runnables import RunnableParallel
2 from langchain.groq import ChatGroq
3 from langchain.core.prompts import ChatPromptTemplate
4 from langchain.core.output_parsers import StrOutputParser
5
6 # Define parallel tasks
7 parallel_chain = RunnableParallel(
8     summary=ChatPromptTemplate.from_template("Summarize: {text}")
9     | ChatGroq(model.name="llama3-8b-8192")
10    | StrOutputParser(),
11
12    keywords=ChatPromptTemplate.from_template("Extract keywords: {text}")
13    | ChatGroq(model.name="gemma-7b-it")
14    | StrOutputParser(),
15
16    sentiment=ChatPromptTemplate.from_template("Analyze sentiment: {text}")
17    | ChatGroq(model.name="gemma-7b-it")
18    | StrOutputParser()
19 )
20
21 # Execute all in parallel
22 results = parallel_chain.invoke({ "text": "Long document content..." })
```



LCEL: Error Handling & Fallbacks

```
// Fallback to Alternative Model:  
2  from langchain.groq import ChatGroq  
3  from langchain.anthropic import ChatAnthropic  
4  
5  primary = ChatGroq(model.name="llama3-70b-8192")  
6  fallback = ChatAnthropic(model="claude-3-opus-20240229")  
7  
8  chain = prompt | primary.with_fallbacks([fallback]) | output_parser  
9  
10 // Retry Logic: If Groq fails, automatically retries  
11  from langchain.core.runnables import RunnableRetry  
12  
13  chain_with_retry = (  
14      prompt  
15      | RunnableRetry(max_attempts=3, wait_exponential_jitter=True)  
16      | ChatGroq(model.name="gemma-7b-it")  
17      | output_parser  
18  )  
19  
20 // Custom Error Handling:  
21  from langchain.core.runnables import RunnableLambda  
22  
23  def handle_errors(x):  
24      try:  
25          return x  
26      except Exception as e:  
27          return {"error": str(e)}  
28  
chain = prompt | model | RunnableLambda(handle_errors)
```



LCEL: Streaming

```
1 // Stream Tokens as Generated:  
from langchain.groq import ChatGroq  
3 # Assume prompt and output_parser are defined  
chain = prompt | ChatGroq(model_name="gemma-7b-it") | StrOutputParser()  
5  
# Stream output  
7 for chunk in chain.stream({ "topic": "AI" }):  
    print(chunk, end="", flush=True)  
9  
// Async Streaming:  
11 import asyncio  
  
13 async def stream_response():  
    async for chunk in chain.astream({ "topic": "AI" }):  
        print(chunk, end="", flush=True)  
15  
17 asyncio.run(stream_response())  
  
19 // Streaming with Events:  
# Get detailed streaming events  
21 async for event in chain.astream_events({ "topic": "AI" }, version="v1"):  
    kind = event["event"]  
23    if kind == "on_chat_model_stream":  
        content = event["data"]["chunk"].content  
        print(content, end="", flush=True)  
25
```



Memory

Memory in LangChain

Why Memory?

- ▶ LLMs are stateless by default
- ▶ Chatbots need conversation context
- ▶ Memory stores and retrieves conversation history

Memory Types (Consolidated):

- ▶ **ConversationBufferMemory:**
 - ▶ Stores entire conversation history
 - ▶ Simple but can exceed token limits
 - ▶ Best for: Short conversations
- ▶ **ConversationBufferWindowMemory:**
 - ▶ Keeps only last K interactions
 - ▶ Prevents token overflow
 - ▶ Best for: Longer conversations with recent context
- ▶ **ConversationSummaryMemory:**
 - ▶ Summarizes old messages
 - ▶ Reduces token usage
 - ▶ Best for: Very long conversations

Memory: Implementation Examples

IMPORTANT NOTE

In LangChain v1, memory classes have moved to langchain-classic. To use
pip install langchain-classic

```
from langchain.classic.memory import ConversationBufferMemory
2 from langchain.groq import ChatGroq
from langchain.core.prompts import ChatPromptTemplate, MessagesPlaceholder
4 from langchain.core.runnables import RunnablePassthrough

6 memory = ConversationBufferMemory(
    return_messages=True,
8     memory_key="history")

10 prompt = ChatPromptTemplate.from_messages([
    ("system", "You are a helpful assistant"),
12     MessagesPlaceholder(variable_name="history"),
    ("human", "{input}"))
14

chain = (
16     RunnablePassthrough.assign(history=lambda x: memory.load_memory_variables({})["history"])
17     | prompt
18     | ChatGroq(model_name="gemma2-9b-it"))

# Use the chain
20 response = chain.invoke({"input": "Hi, I'm Alice"})
memory.save_context({"input": "Hi, I'm Alice"}, {"output": response.content})
```



Memory: Window Memory

Benefits:

- ▶ Fixed memory footprint
- ▶ Maintains recent context
- ▶ Prevents token limit issues

Buffer Window Memory (Keeps Last K):

```
from langchain.memory import ConversationBufferWindowMemory
2
memory = ConversationBufferWindowMemory(
4     k=3, # Only keeps last 3 interactions
5     return_messages=True,
6     memory_key="history"
)
8
conversations = [ # Add conversations
10    ("Tell me about Python", "Python is a versatile language..."),
11    ("What makes it popular?", "Its simplicity and libraries..."),
12    ("Give an example", "Like NumPy for computing..."),
13    ("What about AI?", "Python excels in AI with TensorFlow...")
14]
16
for input_text, output_text in conversations:
17     memory.save_context({"input": input_text}, {"output": output_text})
18
# Retrieves only last 3 interactions
20 history = memory.load_memory_variables({})
print(f"Stored messages: {len(history['history'])}") # Will be 6 (3 pairs)
```



Agents & Tools



Modern Agents Overview

What Are Agents?

- ▶ Use LLMs as reasoning engines
- ▶ Dynamically choose which tools to use
- ▶ Iterative: observe, think, act, repeat
- ▶ Best for complex, multi-step tasks

Modern Agent Types (LangChain v1):

- ▶ **create_agent**: Primary method, uses ReAct pattern
- ▶ **create_deep_agent**: For complex, long-running tasks

When to Use:

- ▶ **create_agent**: Simple chatbots, Q&A, tool calling
- ▶ **create_deep_agent**: Research, multi-step planning, file-system access
- ▶ **LangGraph**: Custom workflows requiring fine-grained control

(Ref: LangChain Agents Documentation)



Creating Tools: Modern Approach

Tool Requirements:

- ▶ Must have a docstring (used by LLM to understand tool purpose)
- ▶ Type hints are required for parameters
- ▶ Return type should be string or JSON-serializable

```
1 // Method 1: Using @tool Decorator:  
from langchain.core.tools import tool  
  
3  
@tool  
5 def search_wikipedia(query: str) -> str:  
    """Search Wikipedia for information about a topic."""  
7     from wikipedia import summary  
    try:  
9         return summary(query, sentences=3)  
except:  
11     return "Could not find information."  
  
13  
@tool  
15 def calculate(expression: str) -> str:  
    """Evaluate a mathematical expression."""  
try:  
17     return str(eval(expression))  
except:  
19     return "Invalid expression"
```

Creating Tools: Modern Approach

// Method 2: From Function:

```
1 from langchain.core.tools import Tool
3 def get_weather(location: str) -> str:
    """Get weather for a location."""
    return f"Weather in {location}: Sunny, 72F"
5
7 weather_tool = Tool.from_function(
        func=get_weather,
        name="weather",
        description="Get current weather for a location")
```

Modern Agent: create_agent()

- ▶ Built on LangGraph (durable execution)
- ▶ Automatic tool calling loop
- ▶ Simple, unified API

```
from langchain.agents import create_agent
from langchain.core.tools import tool
from langchain.groq import ChatGroq
1
2
@tool
3 def get_weather(city: str) -> str:
4     """Get weather for a given city."""
5     return f"It's sunny in {city}!"
6
7
@tool
8 def search_wikipedia(query: str) -> str:
9     """Search Wikipedia for information."""
10    return f"Wikipedia results for: {query}"
11
12
# Create agent with tools
13 agent = create_agent(
14     model="llama-3.3-70b-versatile", # Can use model string or ChatGroq object
15     tools=[get_weather, search_wikipedia],
16     system_prompt="You are a helpful assistant that can check weather and search Wikipedia.")
17
18
# Run the agent
19 response = agent.invoke({"messages": [{"role": "user", "content": "What's the weather in Paris?"}]})
20
21 print(response["messages"][-1].content)
22
23
24
```



Modern Approach with bind_tools

```
from langchain.groq import ChatGroq
from langchain.core.tools import tool

@tool
def multiply(a: int, b: int) -> int:
    """Multiply two numbers."""
    return a * b

@tool
def add(a: int, b: int) -> int:
    """Add two numbers."""
    return a + b

# Bind tools to model
llm = ChatGroq(model_name="llama3-8b-8192")
llm.with_tools = llm.bind_tools([multiply, add])

# Invoke
response = llm.with_tools.invoke("What is 3 times 4 plus 5?")

# Check if tool was called
if response.tool_calls:
    tool_call = response.tool_calls[0]
    print(f"Tool: {tool_call['name']}")
    print(f"Args: {tool_call['args']}
```

Agent Middleware

Extend Agent Behavior Without Modifying Core Logic:

- ▶ **Human-in-the-loop:** Approve tool calls before execution
- ▶ **Conversation compression:** Summarize long conversations
- ▶ **PII redaction:** Remove sensitive data
- ▶ **Rate limiting:** Control API usage
- ▶ **Error handling:** Retry failed operations

```
1 from langchain.agents import create_agent
2 from langchain.agents.middleware import AgentMiddleware
3
4 # Example: Logging middleware
5 class LoggingMiddleware(AgentMiddleware):
6     async def on_agent_start(self, state, context):
7         print(f"Agent started with input: {state['messages'][-1].content}")
8
9     async def on_agent_end(self, state, context):
10        print(f"Agent completed with output: {state['messages'][-1].content}")
11
12 # Create agent with middleware
13 agent = create_agent(
14     model="llama-3.3-70b-versatile",
15     tools=[get_weather],
16     middleware=[LoggingMiddleware()]
17 )
```

Output Parsers

Output Parsers Overview

Why Output Parsers?

- ▶ LLMs return unstructured text
- ▶ Applications need structured data
- ▶ Parsers extract and validate output

Common Parser Types:

- ▶ **StrOutputParser**: Basic string extraction
- ▶ **JsonOutputParser**: Parse JSON responses
- ▶ **PydanticOutputParser**: Structured data with validation
- ▶ **StructuredOutputParser**: Multiple fields
- ▶ **CommaSeparatedListOutputParser**: Lists

Modern Best Practice:

- ▶ Use OpenAI's `with_structured_output()` when possible
- ▶ More reliable than prompt-based parsing
- ▶ Leverages function calling internally

Structured Output: Modern Approach

Benefits:

- ▶ Type-safe responses
- ▶ Automatic validation
- ▶ More reliable than prompt-based parsing

```
1 from langchain.groq import ChatGroq
2 from pydantic import BaseModel, Field
3
4 # Define output schema
5 class Person(BaseModel):
6     name: str = Field(description="Person's name")
7     age: int = Field(description="Person's age")
8     occupation: str = Field(description="Person's job")
9
10 # Create model with structured output (relies on tool-calling)
11 llm = ChatGroq(model.name="llama3-70b-8192")
12 structured_llm = llm.with_structured_output(Person)
13
14 # Use in chain
15 response = structured_llm.invoke(
16     "Tell me about a software engineer named Alice who is 28"
17 )
18
19 # Response is a Pydantic object
20 print(response.name)
21 print(response.age)
22 print(response.occupation)
```



Pydantic Output Parser (Alternative)

```
from langchain.output_parsers import PydanticOutputParser
from langchain_groq import ChatGroq
from pydantic import BaseModel, Field, validator
from typing import List

class MovieReview(BaseModel):
    title: str = Field(description="Movie title")
    rating: int = Field(description="Rating from 1–10")

parser = PydanticOutputParser(pydantic.object=MovieReview)

# Add to prompt
prompt = ChatPromptTemplate.from_template("""
Review the movie: {movie}

{format_instructions}""")

chain = (prompt.partial(format_instructions=parser.get_format_instructions()
    | ChatGroq(model.name="gemma-7b-it") | parser)
result = chain.invoke({ "movie": "Inception" })
```

Output Parser Error Handling

```
1 from langchain.groq import ChatGroq
# Assume 'parser' is a defined PydanticOutputParser
3
// OutputFixingParser (Auto-fix Errors):
5 from langchain.output_parsers import OutputFixingParser
7
# If parsing fails, use LLM to fix
fixing_parser = OutputFixingParser.from_llm(
9     parser=parser,
    llm=ChatGroq(model.name="gemma-7b-it")
11 )
13
// Automatically fixes malformed output
# result = fixing_parser.parse(malformed_output)
15
// RetryOutputParser (Retry with Context):
17 from langchain.output_parsers import RetryWithErrorOutputParser
19
retry_parser = RetryWithErrorOutputParser.from_llm(
    parser=parser,
    llm=ChatGroq(model.name="llama3-8b-8192")
21 )
23
# Retries with both output and original prompt
25 # result = retry_parser.parse_with_prompt(...)
```



LangChain Ecosystem



LangChain Ecosystem Components

- ▶ **LangChain Core:**
 - ▶ Base abstractions and LCEL
 - ▶ Foundation for all other packages
- ▶ **LangChain Community:**
 - ▶ Third-party integrations
 - ▶ Vector stores, document loaders
 - ▶ Community-maintained tools
- ▶ **LangGraph:**
 - ▶ Build stateful, multi-actor applications
 - ▶ Complex agent workflows with cycles
 - ▶ State management for agents
- ▶ **LangServe:**
 - ▶ Deploy chains as REST APIs
 - ▶ Automatic FastAPI generation
 - ▶ Production deployment
- ▶ **LangSmith:**
 - ▶ Debugging and monitoring
 - ▶ Tracing and evaluation
 - ▶ Dataset management

LangGraph: Stateful Agents

- ▶ Build complex, stateful agent workflows
- ▶ Support for cycles and conditional logic
- ▶ Persist state across interactions
- ▶ Multiple agents working together

```
1 from langgraph.graph import StateGraph
2 from typing import TypedDict, Annotated
3 import operator
4
5 class AgentState(TypedDict):
6     messages: Annotated[list, operator.add]
7     next: str
8
9     def call_model(state):
10         response = llm.invoke(state["messages"])
11         return { "messages": [response] }
12
13     def should_continue(state):
14         if len(state["messages"]) > 5:
15             return "end"
16         return "continue"
17
18 workflow = StateGraph(AgentState)
19 workflow.add_node("agent", call_model)
20 workflow.add_conditional_edges("agent", should_continue)
21 workflow.set_entry_point("agent")
22
23 app = workflow.compile()
```



LangServe: Deploy as API

Automatic Features:

- ▶ Interactive playground at /chat/playground
- ▶ OpenAPI docs at /docs
- ▶ Streaming support
- ▶ Batch processing

Deploy Any Chain as REST API:

```
1 from fastapi import FastAPI
2 from langchain import add_routes
3 from langchain_groq import ChatGroq
4 from langchain_core.prompts import ChatPromptTemplate
5
6 # Create chain
7 prompt = ChatPromptTemplate.from_template("Tell me about {topic}")
8 chain = prompt | ChatGroq(model_name="gemma-7b-it")
9
10 # Create FastAPI app
11 app = FastAPI(
12     title="LangChain API", version="1.0",
13     description="API for my LangChain application")
14
15 # Add chain as route
16 add_routes(app, chain, path="/chat")
17
18 # Run: uvicorn app:app --reload
19 # API available at: http://localhost:8000/chat
```



LangSmith: Monitoring & Debugging

What is LangSmith?

- ▶ Platform for debugging LLM applications
- ▶ Trace every step of chain execution
- ▶ Evaluate model performance
- ▶ Manage test datasets
- ▶ Monitor production applications

Setup:

```
1 import os
2 from langchain.groq import ChatGroq
3 from langchain.core.prompts import ChatPromptTemplate
4 from langchain.core.output_parsers import StrOutputParser
5
6 os.environ["LANGCHAIN_TRACING_V2"] = "true"
7 os.environ["LANGCHAIN_APIKEY"] = "your-langsmith-api-key"
8 os.environ["LANGCHAIN_PROJECT"] = "my-groq-project"
9
10 # Now all chain executions are automatically traced
11 prompt = ChatPromptTemplate.from_template("Tell me about {topic}")
12 llm = ChatGroq(model.name="gemma-7b-it")
13 output_parser = StrOutputParser()
14 chain = prompt | llm | output_parser
15 result = chain.invoke({"topic": "Large Language Models"})
16
17 # View traces at: https://smith.langchain.com
```



Best Practices



Best Practices: Error Handling

```
1 // 1. Use Fallbacks:  
2 from langchain.groq import ChatGroq  
3 from langchain.anthropic import ChatAnthropic  
4  
5 primary = ChatGroq(model.name="llama3-70b-8192")  
6 fallback = ChatAnthropic(model="claude-3-opus-20240229")  
7  
8 chain = prompt | primary.with_fallbacks([fallback]) | parser  
9  
// 2. Implement Retries:  
10 from langchain.core.runnables import RunnableRetry  
11  
12 chain_with_retry = (  
13     prompt  
14     | RunnableRetry(  
15         max_attempts=3,  
16         wait_exponential_jitter=True  
17     )  
18     | ChatGroq(model.name="gemma-7b-it")  
19     | parser  
20 )  
21  
22 // 3. Graceful Degradation:  
23 try:  
24     result = chain.invoke(input_data)  
25 except Exception as e:  
26     # logger.error(f"Chain failed: {e}")  
27     # result = fallback.response()  
28     pass
```

Best Practices: Token Management with Groq

Groq Models - Fast and Free Tier:

- ▶ gemma2-9b-it: Ultra-fast, 8K context, great for chat
- ▶ llama3-8b-8192: Balanced, 8K context, general purpose
- ▶ llama3-70b-8192: Most capable, 8K context, complex tasks
- ▶ llama-3.1-8b-instant: Fastest, optimized for low latency
- ▶ Groq billed by requests/day on free tier, not tokens

```
1 from langchain.groq import ChatGroq
2
3 # Choose model based on needs
4 llm_fast = ChatGroq(
5     model_name="gemma2-9b-it",
6     max_tokens=500,
7     temperature=0.7,
8     max_retries=2
9 )
10
11 # Monitor context length manually
12 from transformers import AutoTokenizer
13 tokenizer = AutoTokenizer.from_pretrained("google/gemma-2-9b-it")
14 token_count = len(tokenizer.encode(text))
15
16 # Groq's speed allows batch processing without latency concerns
```



Best Practices: Security & Privacy

- ▶ Be aware of the data usage policies of your LLM provider.
- ▶ Groq has a zero-retention policy for API data.
- ▶ Consider self-hosted models for maximum data control.
- ▶ Implement PII detection and redaction before sending data.

```
// 1. API Key Management:  
2 import os  
3 from dotenv import load_dotenv  
4  
5 load_dotenv() # Load from .env file  
6 api_key = os.getenv("GROQ_API_KEY")  
  
// 2. Timeouts and Retries:  
7 from langchain-groq import ChatGroq  
8  
9 llm = ChatGroq(model_name="llama3-8b-8192",  
10    temperature=0,  
11    max_retries=2,  
12    request_timeout=30) # seconds  
  
// 3. Input Validation:  
13 def validate_input(user_input: str) -> str:  
14     if len(user_input) > 8000: # Sanitize input  
15         raise ValueError("Input too long for the model context.")  
16     # Remove potential injection attempts  
17     return user_input.strip()
```



Best Practices: Choosing the Right Model

Groq Model Selection Guide:

Use Case	Model	Why
Chatbots	gemma2-9b-it	Fast, efficient, good reasoning
Summarization	llama3-8b-8192	Balanced speed/quality
Complex reasoning	llama3-70b-8192	Most capable
Ultra-low latency	llama-3.1-8b-instant	Optimized for speed
Code generation	llama3-70b-8192	Better instruction following

```
1 # Pattern: Start with fast model, fallback to capable model
from langchain.groq import ChatGroq
3 primary = ChatGroq(model_name="gemma2-9b-it", temperature=0.7)
5 fallback = ChatGroq(model_name="llama3-70b-8192", temperature=0.7)
7 chain = prompt | primary.with_fallbacks([fallback]) | parser
```

Best Practices: Async Patterns

Benefits:

- ▶ Faster parallel processing
- ▶ Better resource utilization
- ▶ Improved user experience with streaming

```
1 // Use Async for Better Performance:  
2 import asyncio  
3  
4 async def process_multiple_queries(queries):  
5     # Process queries concurrently  
6     tasks = [chain.await_invoke({ "input": q }) for q in queries]  
7     results = await asyncio.gather(*tasks)  
8     return results  
9  
10 # Run  
11 queries = ["Query 1", "Query 2", "Query 3"]  
12 results = asyncio.run(process_multiple_queries())  
13  
14 // Async Streaming:  
15 async def stream_response(input_text):  
16     async for chunk in chain.astream({ "input": input_text }):  
17         print(chunk, end="", flush=True)  
18  
19 asyncio.run(stream_response("Tell me about AI"))
```



What's New in LangChain Ecosystem

(Oct 2025, Release of 1.0 version)

Major Additions in LangChain/Graph 1.0

- ▶ LangChain is now on top of LangGraph (not the other way round, as before)
- ▶ LangChain agent `create_agent` (ReACT) is one specialized case of many types in LangGraph. You can even visualize it as a workflow of nodes.
- ▶ Middle-ware to control the data before going to LLM. Curation, summarization, context engineering, etc.
- ▶ Standard content blocks to cater to different output formats by different LLMs.

Code Reorganization in LangChain/Graph 1.0

This is in line with principle of 'Simplified Name-spaces'.

- ▶ LangChain legacy chains, hub, vector stores, retrievers etc are now part of `langchain_classic` module.
- ▶ All agents are in `langchain.agents`
- ▶ All Models are in `langchain.chat_models`
- ▶ All Tools are in `langchain.tools`
- ▶ All Messages are in `langchain.messages`
- ▶ All Embeddings are in `langchain.embeddings`

LangChain v1: What's New

Major Changes in v1.0 (October 2025):

- ▶ **Simplified Package Structure:**

- ▶ Core functionality in `langchain`
- ▶ Legacy features moved to `langchain-classic`
- ▶ Requires Python 3.10+ (Python 3.9 EOL October 2025)

- ▶ **Agent-First Design:**

- ▶ New `create_agent()` function
- ▶ Built on LangGraph runtime
- ▶ Durable execution and persistence

- ▶ **Content Blocks:**

- ▶ Standardized message format across providers
- ▶ Support for text, images, reasoning, tool calls
- ▶ Better multimodal support

- ▶ **Middleware System:**

- ▶ Inject custom logic at any point
- ▶ Human-in-the-loop support
- ▶ Error handling and retries

(Ref: LangChain v1.0 Release Notes)



How LangChain 1.0 Simplifies Agents to Just 10 Lines of Code

- ▶ LangChain 1.0 Alpha released; full 1.0 expected by late October.
- ▶ Rewritten as a simplified agent runtime built on LangGraph.
- ▶ Agents can now be created in roughly 10/few lines of code.
- ▶ Legacy LangChain moved to “LangChain Classic.”
- ▶ Functionally similar to OpenAI or Pedantic SDK agents.
- ▶ Unified “create_agent” abstraction across languages.
- ▶ Integrates easily with external SDKs like Google’s 80k.

Why LangGraph 1.0 is Nearly Non-Breaking and Production-Ready

- ▶ LangGraph 1.0 introduces minimal breaking changes from prior releases.
- ▶ Existing LangGraph implementations continue to work as-is.
- ▶ Core runtime stable across Python and TypeScript.
- ▶ Supports durable execution with checkpoints and rollback.
- ▶ State can persist via Postgres or SQLite for reliability.
- ▶ Human-in-the-loop and interrupt handling built-in.
- ▶ Ready for production-grade use; used by Uber, LinkedIn, Klarna, JPMorgan, Cloudflare.



The New Standardized Content Blocks for Easier Model Switching

- ▶ LangChain Core now uses standardized content/message blocks.
- ▶ Abstracts input/output across providers like OpenAI and Anthropic.
- ▶ Simplifies switching models without rewriting logic.
- ▶ Unifies multimodal inputs, reasoning traces, and tool calls.
- ▶ Middleware layer ensures consistent formatting and metadata.
- ▶ Supports normalized logging across different model providers.
- ▶ Key enabler for multi-model experimentation and portability.



Durable Execution, Streaming, Human-in-the-Loop, and Time Travel

- ▶ Agents persist state with rollback and checkpointing.
- ▶ Supports human approval and manual intervention mid-run.
- ▶ Time travel enables returning to earlier workflow states.
- ▶ Four streaming modes: messages, updates, values, custom.
- ▶ Update streaming ideal for dashboards and UX refresh.
- ▶ State persistence possible locally or via Postgres.
- ▶ Enables retry and branch execution for debugging and auditing.



Summary: What's New in LangChain Ecosystem

(Oct 2025, Release of 1.0 version)

1. LangChain Agents: The Standard Loop

- ▶ **High-Level Abstraction:** Simplest and fastest way to build an agent.
- ▶ **Primary Pattern:** Implements the standard **ReAct-style loop** (Reasoning + Acting).
- ▶ **Key Feature:** The `create_agent` function (introduced in v1.0) is the simplest and fastest way to build a production-ready agent. It abstracts away the complexity.
- ▶ **Process:** The LLM decides at each step whether to use a tool or provide a final answer.
- ▶ **Underlying Technology:** LangChain's agents since v1.0 are built **on top of the LangGraph runtime**.
- ▶ **Use Case:** Quick start, simple linear tasks, default agent behavior.

2. LangGraph: Workflow Automation & Control

- ▶ **Role:** Low-level **orchestration framework** and runtime.
- ▶ **Core Feature:** Defines explicit workflow automation and agent orchestration. Complex, stateful, and cyclic workflows as directed graphs.
- ▶ **Structure:** You define:
 1. **Nodes:** Steps (LLM call, tool execution, custom function).
 2. **Edges:** Conditional logic for transitions (e.g., if tool succeeds, go to Node B; if it fails, go to Node C).
- ▶ **Multi-Agent Systems:** Provides the explicit control needed for building supervised multi-agent systems and defining handoffs.
- ▶ **Relationship to LangChain:** Used when the standard LangChain agent loop is insufficient, requiring **fine-grained control** over the flow.

3. Deep Agents: Concept for Complex Tasks

- ▶ **Concept/Library:** Not a separate module, but an architectural pattern built on LangGraph for solving highly complex, long-running tasks.
- ▶ **Focus:** Advanced context management and planning.
- ▶ **Key Techniques:**
 1. **Planning/Decomposition:** Breaking down the goal (e.g., using a "Todo List" tool).
 2. **Sub-agents:** Delegating specialized tasks to other agents for parallel or sequential execution.
 3. **Long-Term Memory:** Utilizing external memory or file systems for persistent context.
- ▶ **Underlying Runtime:** Deep Agents are also built on LangGraph, leveraging its capabilities to manage the complex state and flow of the agent's multi-step plan.

Summary: Agent Framework Comparison

Feature	LangChain Agent	LangGraph	Deep Agents (Concept)
Abstraction Level	High-level API	Low-level/Foundational	High-level (Complex Use Case)
Primary Goal	Quick start, simple agent loop (ReAct)	Explicit flow control, durable runtime	Complex, long-running tasks, planning
Core Architecture	Pre-built ReAct loop	Custom, explicit graph (Nodes & Edges)	Specialized architecture for planning/sub-agents
Multi-Agent	Can be wrapped as a tool for another agent	The foundation for building custom multi-agent flows	Includes sub-agents for specialized tasks
Runtime	Built on Lang-Graph	The Runtime itself	Built on Lang-Graph

Key Takeaway: LangGraph provides the underlying runtime and control flow for both simple and complex agent applications.

YHK

Conclusions

Key Takeaways

What We Learned:

- ▶ LCEL makes chains simple
- ▶ Agents enable autonomy
- ▶ RAG connects to your data
- ▶ Memory maintains context
- ▶ Modular design = flexibility

You're ready to build intelligent LLM applications!



LangChain at a Glance



- ▶ **Models**: LLMs, Chat, Embeddings
- ▶ **Prompts**: Dynamic templates
- ▶ **Chains**: Compose with LCEL

(Ref: Building Generative AI applications - Anand Iyer, Rajesh Thallam)

- ▶ **Memory**: Conversation state
- ▶ **Retrievers**: Document access
- ▶ **Agents**: Autonomous tools

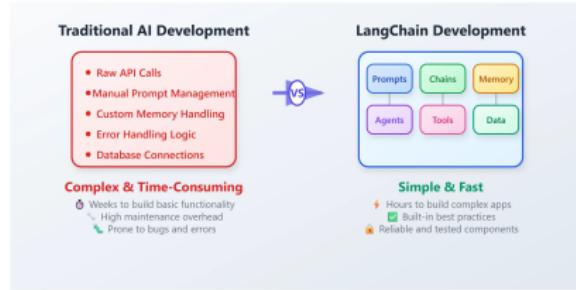
Best Practices to Remember

Development:

- ▶ Use LCEL for chains
- ▶ Start with simple models
- ▶ Add complexity gradually
- ▶ Test with small datasets

Production:

- ▶ Enable LangSmith tracing
- ▶ Implement error handling
- ▶ Monitor token usage
- ▶ Use async for scale



(Ref: What is LangChain and Why Should You Care? - Saif Ali)

Golden Rule:

Keep it simple,
make it work,
then optimize

Resources & Next Steps

Official Documentation:

- ▶ LangChain Docs: <https://python.langchain.com>
- ▶ LangChain Blog: <https://blog.langchain.dev>
- ▶ LangChain Academy: <https://academy.langchain.com>
- ▶ API Reference: <https://api.python.langchain.com>

GitHub Repositories:

- ▶ Core: <https://github.com/langchain-ai/langchain>
- ▶ Templates:
<https://github.com/langchain-ai/langchain/tree/master/templates>
- ▶ LangGraph: <https://github.com/langchain-ai/langgraph>

Community:

- ▶ Discord: <https://discord.gg/langchain>
- ▶ Twitter: @LangChainAI
- ▶ YouTube: LangChain official channel

Practice:

- ▶ Start with simple LCEL chains
- ▶ Build a RAG application
- ▶ Create custom tools and agents
- ▶ Deploy with LangServe

Final Thoughts



LangChain makes building with LLMs accessible

- ▶ Start experimenting today
- ▶ Join the community
- ▶ Share your projects
- ▶ Keep learning and building

Thank you!

Questions? Let's discuss.

(Ref: What is Langchain and why should I care as a developer? - Logan Kilpatrick)

References

Many publicly available resources:

- ▶ Intro to LangChain for Beginners - A code-based walkthrough- Menlo Park Lab
- ▶ LangChain Crash Course (10 minutes): Easy-to-Follow Walkthrough of the Most Important Concepts - Menlo Park Lab
- ▶ Official doc: <https://docs.langchain.com/docs/>
- ▶ Git Repo: <https://github.com/hwchase17/langchain>
- ▶ LangChain 101: The Complete Beginner's Guide Edrick <https://www.youtube.com/watch?v=P3MAbZ2eMUI> A wonderful overview, don't miss
- ▶ Cookbook by Gregory Kamradt(Easy way to get started): <https://github.com/gkamradt/langchain-tutorials/blob/main/LangChain%20Cookbook.ipynb>
- ▶ Youtube Tutorials: https://www.youtube.com/watch?v=_v_fgW2SkkQ
- ▶ LangChain 101 Course (updated with LCEL) - Ivan Reznikov <https://github.com/IvanReznikov/DataVerse/tree/main/Courses/LangChain>
- ▶ A Complete LangChain Guide <https://nanonets.com/blog/langchain/>
- ▶ 7 Ways to Use LangSmith's Superpowers to Turboboost Your LLM Apps - Menlo Park Lab
- ▶ LangChain official blog: <https://blog.langchain.dev>
- ▶ LangChain Academy: <https://academy.langchain.com>
- ▶ LangChain templates repo: <https://github.com/langchain-ai/langchain/tree/master/templates>
- ▶ Groq Documentation: <https://console.groq.com/docs>
- ▶ Groq Model Playground: <https://groq.com/>
- ▶ LangChain Groq Integration: <https://python.langchain.com/docs/integrations/chat/groq>



Thanks ...

- ▶ Office Hours: Saturdays, 3 to 5 pm (IST);
Free-Open to all; email for appointment to
yogeshkulkarni at yahoo dot com
- ▶ Call + 9 1 9 8 9 0 2 5 1 4 0 6



(<https://www.linkedin.com/in/yogeshkulkarni/>)



(<https://medium.com/@yogeshharibhaukul>)



(<https://www.github.com/yogeshhk/>)



Pune AI Community (PAIC)

- ▶ Two-way communication:
 - ▶ Website puneaicommunity dot org
 - ▶ Email puneaicommunity at gmail dot com
 - ▶ Call + 9 1 9 8 9 0 2 5 1 4 0 6
 - ▶ LinkedIn:
<https://linkedin.com/company/pune-ai-community>
- ▶ One-way Announcements:
 - ▶ Twitter (X) @puneaicommunity
 - ▶ Instagram @puneaicommunity
 - ▶ WhatsApp Community: Invitation Link
<https://chat.whatsapp.com/LluOrhyEzuQLDr25ixZ>
 - ▶ Luma Event Calendar: puneaicommunity
- ▶ Contribution Channels:
 - ▶ GitHub: Pune-AI-Community and puneaicommunity
 - ▶ Medium: pune-ai-community
 - ▶ YouTube: @puneaicommunity



Website

Pune AI Community (PAIC) QR codes



Website



Medium Blogs



Twitter-X



LinkedIn Page



Github Repository



WhatsApp Invite



Luma Events



YouTube Videos



Instagram



Pune
DevCon
2025»



Thank you!

Our Sponsors

Platinum Sponsor



Gold Sponsor



Bronze Sponsor

