🧠 Multi-Agent RAG Research Assistant

This notebook demonstrates a multi-agent RAG (Retrieval-Augmented Generation) pipeline using LangGraph, FAISS, Together.ai (LLaMA-4), and optional PDF ingestion.

Users can interactively generate summaries from either:

- Web search results (via SerpAPI), or
- Uploaded research PDFs

It includes a modular design, agent-based orchestration, and a Gradio UI for live summarization based on user-defined prompts.

⚙️ Tech Stack

- **LangGraph** – Agent orchestration and graph state management
- **FAISS** – Vector indexing and similarity retrieval
- **MiniLM (sentence-transformers)** – Document embedding
- **Together.ai** – LLM-powered summarization (LLaMA-4)
- **SerpAPI** – Google search result ingestion
- **PyMuPDF** – PDF parsing and text extraction
- **Gradio** – Interactive user interface

```
# Basic NLP and retrieval tools
!pip install -q sentence-transformers langchain faiss-cpu serpapi
```

```
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 30.7/30.7 MB 41.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 363.4/363.4 MB 3.7 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 13.8/13.8 MB 63.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 24.6/24.6 MB 54.8 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 883.7/883.7 kB 34.5 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 664.8/664.8 MB 1.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 211.5/211.5 MB 5.9 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 56.3/56.3 MB 11.3 MB/s eta
```

```
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 127.9/127.9 MB 7.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 207.5/207.5 MB 5.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 21.1/21.1 MB 51.8 MB/s eta
0:00:00

# LLaMA model loader (this is the slow one that may need C++
compilation)
!pip install -q llama-cpp-python

━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 67.3/67.3 MB 11.1 MB/s eta
0:00:00
ents to build wheel ... etadata (pyproject.toml) ...
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 45.5/45.5 kB 3.0 MB/s eta
0:00:00
a-cpp-python (pyproject.toml) ...

!pip install -U langchain-community

Collecting langchain-community
  Downloading langchain_community-0.3.21-py3-none-any.whl.metadata
(2.4 kB)
Collecting langchain-core<1.0.0,>=0.3.51 (from langchain-community)
  Downloading langchain_core-0.3.51-py3-none-any.whl.metadata (5.9 kB)
Collecting langchain<1.0.0,>=0.3.23 (from langchain-community)
  Downloading langchain-0.3.23-py3-none-any.whl.metadata (7.8 kB)
Requirement already satisfied: SQLAlchemy<3,>=1.4 in
/usr/local/lib/python3.11/dist-packages (from langchain-community)
(2.0.40)
Requirement already satisfied: requests<3,>=2 in
/usr/local/lib/python3.11/dist-packages (from langchain-community)
(2.32.3)
Requirement already satisfied: PyYAML>=5.3 in
/usr/local/lib/python3.11/dist-packages (from langchain-community)
(6.0.2)
Requirement already satisfied: aiohttp<4.0.0,>=3.8.3 in
/usr/local/lib/python3.11/dist-packages (from langchain-community)
(3.11.15)
Requirement already satisfied: tenacity!=8.4.0,<10,>=8.1.0 in
/usr/local/lib/python3.11/dist-packages (from langchain-community)
(9.1.2)
Collecting dataclasses-json<0.7,>=0.5.7 (from langchain-community)
  Downloading dataclasses_json-0.6.7-py3-none-any.whl.metadata (25 kB)
Collecting pydantic-settings<3.0.0,>=2.4.0 (from langchain-community)
  Downloading pydantic_settings-2.8.1-py3-none-any.whl.metadata (3.5
kB)
Requirement already satisfied: langsmith<0.4,>=0.1.125 in
/usr/local/lib/python3.11/dist-packages (from langchain-community)
```

```
(0.3.23)
Collecting httpx-sse<1.0.0,>=0.4.0 (from langchain-community)
  Downloading httpx_sse-0.4.0-py3-none-any.whl.metadata (9.0 kB)
Requirement already satisfied: numpy<3,>=1.26.2 in
/usr/local/lib/python3.11/dist-packages (from langchain-community)
(2.0.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain-community) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain-community) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain-community) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain-community) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain-community) (6.2.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain-community) (0.3.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.8.3-
>langchain-community) (1.18.3)
Collecting marshmallow<4.0.0,>=3.18.0 (from dataclasses-
json<0.7,>=0.5.7->langchain-community)
  Downloading marshmallow-3.26.1-py3-none-any.whl.metadata (7.3 kB)
Collecting typing-inspect<1,>=0.4.0 (from dataclasses-
json<0.7,>=0.5.7->langchain-community)
  Downloading typing_inspect-0.9.0-py3-none-any.whl.metadata (1.5 kB)
Collecting langchain-text-splitters<1.0.0,>=0.3.8 (from
langchain<1.0.0,>=0.3.23->langchain-community)
  Downloading langchain_text_splitters-0.3.8-py3-none-any.whl.metadata
(1.9 kB)
Requirement already satisfied: pydantic<3.0.0,>=2.7.4 in
/usr/local/lib/python3.11/dist-packages (from
langchain<1.0.0,>=0.3.23->langchain-community) (2.11.2)
Requirement already satisfied: jsonpatch<2.0,>=1.33 in
/usr/local/lib/python3.11/dist-packages (from langchain-
core<1.0.0,>=0.3.51->langchain-community) (1.33)
Requirement already satisfied: packaging<25,>=23.2 in
/usr/local/lib/python3.11/dist-packages (from langchain-
core<1.0.0,>=0.3.51->langchain-community) (24.2)
Requirement already satisfied: typing-extensions>=4.7 in
/usr/local/lib/python3.11/dist-packages (from langchain-
core<1.0.0,>=0.3.51->langchain-community) (4.13.1)
```

```
Requirement already satisfied: httpx<1,>=0.23.0 in
/usr/local/lib/python3.11/dist-packages (from langsmith<0.4,>=0.1.125-
>langchain-community) (0.28.1)
Requirement already satisfied: orjson<4.0.0,>=3.9.14 in
/usr/local/lib/python3.11/dist-packages (from langsmith<0.4,>=0.1.125-
>langchain-community) (3.10.16)
Requirement already satisfied: requests-toolbelt<2.0.0,>=1.0.0 in
/usr/local/lib/python3.11/dist-packages (from langsmith<0.4,>=0.1.125-
>langchain-community) (1.0.0)
Requirement already satisfied: zstandard<0.24.0,>=0.23.0 in
/usr/local/lib/python3.11/dist-packages (from langsmith<0.4,>=0.1.125-
>langchain-community) (0.23.0)
Collecting python-dotenv>=0.21.0 (from pydantic-
settings<3.0.0,>=2.4.0->langchain-community)
  Downloading python_dotenv-1.1.0-py3-none-any.whl.metadata (24 kB)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests<3,>=2-
>langchain-community) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.11/dist-packages (from requests<3,>=2-
>langchain-community) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests<3,>=2-
>langchain-community) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests<3,>=2-
>langchain-community) (2025.1.31)
Requirement already satisfied: greenlet>=1 in
/usr/local/lib/python3.11/dist-packages (from SQLAlchemy<3,>=1.4-
>langchain-community) (3.1.1)
Requirement already satisfied: anyio in
/usr/local/lib/python3.11/dist-packages (from httpx<1,>=0.23.0-
>langsmith<0.4,>=0.1.125->langchain-community) (4.9.0)
Requirement already satisfied: httpcore==1.* in
/usr/local/lib/python3.11/dist-packages (from httpx<1,>=0.23.0-
>langsmith<0.4,>=0.1.125->langchain-community) (1.0.7)
Requirement already satisfied: h11<0.15,>=0.13 in
/usr/local/lib/python3.11/dist-packages (from httpcore==1.*-
>httpx<1,>=0.23.0->langsmith<0.4,>=0.1.125->langchain-community)
(0.14.0)
Requirement already satisfied: jsonpointer>=1.9 in
/usr/local/lib/python3.11/dist-packages (from jsonpatch<2.0,>=1.33-
>langchain-core<1.0.0,>=0.3.51->langchain-community) (3.0.0)
Requirement already satisfied: annotated-types>=0.6.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0,>=2.7.4-
>langchain<1.0.0,>=0.3.23->langchain-community) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.1 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0,>=2.7.4-
>langchain<1.0.0,>=0.3.23->langchain-community) (2.33.1)
```

```
Requirement already satisfied: typing-inspection>=0.4.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0,>=2.7.4-
>langchain<1.0.0,>=0.3.23->langchain-community) (0.4.0)
Collecting mypy-extensions>=0.3.0 (from typing-inspect<1,>=0.4.0-
>dataclasses-json<0.7,>=0.5.7->langchain-community)
  Downloading mypy_extensions-1.0.0-py3-none-any.whl.metadata (1.1 kB)
Requirement already satisfied: sniffio>=1.1 in
/usr/local/lib/python3.11/dist-packages (from anyio->httpx<1,>=0.23.0-
>langsmith<0.4,>=0.1.125->langchain-community) (1.3.1)
Downloading langchain_community-0.3.21-py3-none-any.whl (2.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 2.5/2.5 MB 16.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.0/1.0 MB 19.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 423.3/423.3 kB 20.1 MB/s eta
0:00:00
arshmallow-3.26.1-py3-none-any.whl (50 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 50.9/50.9 kB 3.7 MB/s eta
0:00:00
ypy_extensions-1.0.0-py3-none-any.whl (4.7 kB)
Installing collected packages: python-dotenv, mypy-extensions,
marshmallow, httpx-sse, typing-inspect, pydantic-settings,
dataclasses-json, langchain-core, langchain-text-splitters, langchain,
langchain-community
  Attempting uninstall: langchain-core
    Found existing installation: langchain-core 0.3.50
    Uninstalling langchain-core-0.3.50:
      Successfully uninstalled langchain-core-0.3.50
  Attempting uninstall: langchain-text-splitters
    Found existing installation: langchain-text-splitters 0.3.7
    Uninstalling langchain-text-splitters-0.3.7:
      Successfully uninstalled langchain-text-splitters-0.3.7
  Attempting uninstall: langchain
    Found existing installation: langchain 0.3.22
    Uninstalling langchain-0.3.22:
      Successfully uninstalled langchain-0.3.22
Successfully installed dataclasses-json-0.6.7 httpx-sse-0.4.0
langchain-0.3.23 langchain-community-0.3.21 langchain-core-0.3.51
langchain-text-splitters-0.3.8 marshmallow-3.26.1 mypy-extensions-
1.0.0 pydantic-settings-2.8.1 python-dotenv-1.1.0 typing-inspect-0.9.0

!pip install llama-cpp-python[server]

Requirement already satisfied: llama-cpp-python[server] in
/usr/local/lib/python3.11/dist-packages (0.3.8)
Requirement already satisfied: typing-extensions>=4.5.0 in
/usr/local/lib/python3.11/dist-packages (from llama-cpp-
python[server]) (4.13.1)
Requirement already satisfied: numpy>=1.20.0 in
/usr/local/lib/python3.11/dist-packages (from llama-cpp-
```

```
python[server]) (2.0.2)
Requirement already satisfied: diskcache>=5.6.1 in
/usr/local/lib/python3.11/dist-packages (from llama-cpp-
python[server]) (5.6.3)
Requirement already satisfied: jinja2>=2.11.3 in
/usr/local/lib/python3.11/dist-packages (from llama-cpp-
python[server]) (3.1.6)
Collecting uvicorn>=0.22.0 (from llama-cpp-python[server])
  Downloading uvicorn-0.34.0-py3-none-any.whl.metadata (6.5 kB)
Collecting fastapi>=0.100.0 (from llama-cpp-python[server])
  Downloading fastapi-0.115.12-py3-none-any.whl.metadata (27 kB)
Requirement already satisfied: pydantic-settings>=2.0.1 in
/usr/local/lib/python3.11/dist-packages (from llama-cpp-
python[server]) (2.8.1)
Collecting sse-starlette>=1.6.1 (from llama-cpp-python[server])
  Downloading sse_starlette-2.2.1-py3-none-any.whl.metadata (7.8 kB)
Collecting starlette-context<0.4,>=0.3.6 (from llama-cpp-
python[server])
  Downloading starlette_context-0.3.6-py3-none-any.whl.metadata (4.3
kB)
Requirement already satisfied: PyYAML>=5.1 in
/usr/local/lib/python3.11/dist-packages (from llama-cpp-
python[server]) (6.0.2)
Collecting starlette<0.47.0,>=0.40.0 (from fastapi>=0.100.0->llama-
cpp-python[server])
  Downloading starlette-0.46.1-py3-none-any.whl.metadata (6.2 kB)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,!=2.0.0,!=2.0.1,!
=2.1.0,<3.0.0,>=1.7.4 in /usr/local/lib/python3.11/dist-packages (from
fastapi>=0.100.0->llama-cpp-python[server]) (2.11.2)
Requirement already satisfied: MarkupSafe>=2.0 in
/usr/local/lib/python3.11/dist-packages (from jinja2>=2.11.3->llama-
cpp-python[server]) (3.0.2)
Requirement already satisfied: python-dotenv>=0.21.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic-
settings>=2.0.1->llama-cpp-python[server]) (1.1.0)
Requirement already satisfied: anyio>=4.7.0 in
/usr/local/lib/python3.11/dist-packages (from sse-starlette>=1.6.1-
>llama-cpp-python[server]) (4.9.0)
Requirement already satisfied: click>=7.0 in
/usr/local/lib/python3.11/dist-packages (from uvicorn>=0.22.0->llama-
cpp-python[server]) (8.1.8)
Requirement already satisfied: h11>=0.8 in
/usr/local/lib/python3.11/dist-packages (from uvicorn>=0.22.0->llama-
cpp-python[server]) (0.14.0)
Requirement already satisfied: idna>=2.8 in
/usr/local/lib/python3.11/dist-packages (from anyio>=4.7.0->sse-
starlette>=1.6.1->llama-cpp-python[server]) (3.10)
Requirement already satisfied: sniffio>=1.1 in
/usr/local/lib/python3.11/dist-packages (from anyio>=4.7.0->sse-
```

```
starlette>=1.6.1->llama-cpp-python[server]) (1.3.1)
Requirement already satisfied: annotated-types>=0.6.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic!=1.8,!=1.8.1,!
=2.0.0,!=2.0.1,!=2.1.0,<3.0.0,>=1.7.4->fastapi>=0.100.0->llama-cpp-
python[server]) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.1 in
/usr/local/lib/python3.11/dist-packages (from pydantic!=1.8,!=1.8.1,!
=2.0.0,!=2.0.1,!=2.1.0,<3.0.0,>=1.7.4->fastapi>=0.100.0->llama-cpp-
python[server]) (2.33.1)
Requirement already satisfied: typing-inspection>=0.4.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic!=1.8,!=1.8.1,!
=2.0.0,!=2.0.1,!=2.1.0,<3.0.0,>=1.7.4->fastapi>=0.100.0->llama-cpp-
python[server]) (0.4.0)
Downloading fastapi-0.115.12-py3-none-any.whl (95 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 95.2/95.2 kB 4.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 62.3/62.3 kB 4.2 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 72.0/72.0 kB 3.4 MB/s eta
0:00:00

!pip install together

Collecting together
  Downloading together-1.5.5-py3-none-any.whl.metadata (14 kB)
Requirement already satisfied: aiohttp<4.0.0,>=3.9.3 in
/usr/local/lib/python3.11/dist-packages (from together) (3.11.15)
Requirement already satisfied: click<9.0.0,>=8.1.7 in
/usr/local/lib/python3.11/dist-packages (from together) (8.1.8)
Collecting eval-type-backport<0.3.0,>=0.1.3 (from together)
  Downloading eval_type_backport-0.2.2-py3-none-any.whl.metadata (2.2
kB)
Requirement already satisfied: filelock<4.0.0,>=3.13.1 in
/usr/local/lib/python3.11/dist-packages (from together) (3.18.0)
Requirement already satisfied: numpy>=1.23.5 in
/usr/local/lib/python3.11/dist-packages (from together) (2.0.2)
Requirement already satisfied: pillow<12.0.0,>=11.1.0 in
/usr/local/lib/python3.11/dist-packages (from together) (11.1.0)
Requirement already satisfied: pyarrow>=10.0.1 in
/usr/local/lib/python3.11/dist-packages (from together) (18.1.0)
Requirement already satisfied: pydantic<3.0.0,>=2.6.3 in
/usr/local/lib/python3.11/dist-packages (from together) (2.11.2)
Requirement already satisfied: requests<3.0.0,>=2.31.0 in
/usr/local/lib/python3.11/dist-packages (from together) (2.32.3)
Requirement already satisfied: rich<14.0.0,>=13.8.1 in
/usr/local/lib/python3.11/dist-packages (from together) (13.9.4)
Requirement already satisfied: tabulate<0.10.0,>=0.9.0 in
/usr/local/lib/python3.11/dist-packages (from together) (0.9.0)
Requirement already satisfied: tqdm<5.0.0,>=4.66.2 in
/usr/local/lib/python3.11/dist-packages (from together) (4.67.1)
```

```
Requirement already satisfied: typer<0.16,>=0.9 in
/usr/local/lib/python3.11/dist-packages (from together) (0.15.2)
Requirement already satisfied: aiohappyeyeballs>=2.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.9.3-
>together) (2.6.1)
Requirement already satisfied: aiosignal>=1.1.2 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.9.3-
>together) (1.3.2)
Requirement already satisfied: attrs>=17.3.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.9.3-
>together) (25.3.0)
Requirement already satisfied: frozenlist>=1.1.1 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.9.3-
>together) (1.5.0)
Requirement already satisfied: multidict<7.0,>=4.5 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.9.3-
>together) (6.2.0)
Requirement already satisfied: propcache>=0.2.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.9.3-
>together) (0.3.1)
Requirement already satisfied: yarl<2.0,>=1.17.0 in
/usr/local/lib/python3.11/dist-packages (from aiohttp<4.0.0,>=3.9.3-
>together) (1.18.3)
Requirement already satisfied: annotated-types>=0.6.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0,>=2.6.3-
>together) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.1 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0,>=2.6.3-
>together) (2.33.1)
Requirement already satisfied: typing-extensions>=4.12.2 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0,>=2.6.3-
>together) (4.13.1)
Requirement already satisfied: typing-inspection>=0.4.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0,>=2.6.3-
>together) (0.4.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests<3.0.0,>=2.31.0-
>together) (3.4.1)
Requirement already satisfied: idna<4,>=2.5 in
/usr/local/lib/python3.11/dist-packages (from requests<3.0.0,>=2.31.0-
>together) (3.10)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests<3.0.0,>=2.31.0-
>together) (2.3.0)
Requirement already satisfied: certifi>=2017.4.17 in
/usr/local/lib/python3.11/dist-packages (from requests<3.0.0,>=2.31.0-
>together) (2025.1.31)
Requirement already satisfied: markdown-it-py>=2.2.0 in
/usr/local/lib/python3.11/dist-packages (from rich<14.0.0,>=13.8.1-
```

```
>together) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in
/usr/local/lib/python3.11/dist-packages (from rich<14.0.0,>=13.8.1-
>together) (2.18.0)
Requirement already satisfied: shellingham>=1.3.0 in
/usr/local/lib/python3.11/dist-packages (from typer<0.16,>=0.9-
>together) (1.5.4)
Requirement already satisfied: mdurl~=0.1 in
/usr/local/lib/python3.11/dist-packages (from markdown-it-py>=2.2.0-
>rich<14.0.0,>=13.8.1->together) (0.1.2)
Downloading together-1.5.5-py3-none-any.whl (87 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 87.9/87.9 kB 4.2 MB/s eta
0:00:00

!pip install langgraph

Collecting langgraph
  Downloading langgraph-0.3.25-py3-none-any.whl.metadata (7.7 kB)
Requirement already satisfied: langchain-core<0.4,>=0.1 in
/usr/local/lib/python3.11/dist-packages (from langgraph) (0.3.51)
Collecting langgraph-checkpoint<3.0.0,>=2.0.10 (from langgraph)
  Downloading langgraph_checkpoint-2.0.24-py3-none-any.whl.metadata
(4.6 kB)
Collecting langgraph-prebuilt<0.2,>=0.1.1 (from langgraph)
  Downloading langgraph_prebuilt-0.1.8-py3-none-any.whl.metadata (5.0
kB)
Collecting langgraph-sdk<0.2.0,>=0.1.42 (from langgraph)
  Downloading langgraph_sdk-0.1.61-py3-none-any.whl.metadata (1.8 kB)
Collecting xxhash<4.0.0,>=3.5.0 (from langgraph)
  Downloading xxhash-3.5.0-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (12 kB)
Requirement already satisfied: langsmith<0.4,>=0.1.125 in
/usr/local/lib/python3.11/dist-packages (from langchain-
core<0.4,>=0.1->langgraph) (0.3.23)
Requirement already satisfied: tenacity!=8.4.0,<10.0.0,>=8.1.0 in
/usr/local/lib/python3.11/dist-packages (from langchain-
core<0.4,>=0.1->langgraph) (9.1.2)
Requirement already satisfied: jsonpatch<2.0,>=1.33 in
/usr/local/lib/python3.11/dist-packages (from langchain-
core<0.4,>=0.1->langgraph) (1.33)
Requirement already satisfied: PyYAML>=5.3 in
/usr/local/lib/python3.11/dist-packages (from langchain-
core<0.4,>=0.1->langgraph) (6.0.2)
Requirement already satisfied: packaging<25,>=23.2 in
/usr/local/lib/python3.11/dist-packages (from langchain-
core<0.4,>=0.1->langgraph) (24.2)
Requirement already satisfied: typing-extensions>=4.7 in
/usr/local/lib/python3.11/dist-packages (from langchain-
core<0.4,>=0.1->langgraph) (4.13.1)
Requirement already satisfied: pydantic<3.0.0,>=2.5.2 in
```

```
/usr/local/lib/python3.11/dist-packages (from langchain-
core<0.4,>=0.1->langgraph) (2.11.2)
Collecting ormsgpack<2.0.0,>=1.8.0 (from langgraph-
checkpoint<3.0.0,>=2.0.10->langgraph)
  Downloading ormsgpack-1.9.1-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (43 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.5/43.5 kB 2.0 MB/s eta
0:00:00
ent already satisfied: httpx>=0.25.2 in
/usr/local/lib/python3.11/dist-packages (from langgraph-
sdk<0.2.0,>=0.1.42->langgraph) (0.28.1)
Requirement already satisfied: orjson>=3.10.1 in
/usr/local/lib/python3.11/dist-packages (from langgraph-
sdk<0.2.0,>=0.1.42->langgraph) (3.10.16)
Requirement already satisfied: anyio in
/usr/local/lib/python3.11/dist-packages (from httpx>=0.25.2-
>langgraph-sdk<0.2.0,>=0.1.42->langgraph) (4.9.0)
Requirement already satisfied: certifi in
/usr/local/lib/python3.11/dist-packages (from httpx>=0.25.2-
>langgraph-sdk<0.2.0,>=0.1.42->langgraph) (2025.1.31)
Requirement already satisfied: httpcore==1.* in
/usr/local/lib/python3.11/dist-packages (from httpx>=0.25.2-
>langgraph-sdk<0.2.0,>=0.1.42->langgraph) (1.0.7)
Requirement already satisfied: idna in /usr/local/lib/python3.11/dist-
packages (from httpx>=0.25.2->langgraph-sdk<0.2.0,>=0.1.42->langgraph)
(3.10)
Requirement already satisfied: h11<0.15,>=0.13 in
/usr/local/lib/python3.11/dist-packages (from httpcore==1.*-
>httpx>=0.25.2->langgraph-sdk<0.2.0,>=0.1.42->langgraph) (0.14.0)
Requirement already satisfied: jsonpointer>=1.9 in
/usr/local/lib/python3.11/dist-packages (from jsonpatch<2.0,>=1.33-
>langchain-core<0.4,>=0.1->langgraph) (3.0.0)
Requirement already satisfied: requests<3,>=2 in
/usr/local/lib/python3.11/dist-packages (from langsmith<0.4,>=0.1.125-
>langchain-core<0.4,>=0.1->langgraph) (2.32.3)
Requirement already satisfied: requests-toolbelt<2.0.0,>=1.0.0 in
/usr/local/lib/python3.11/dist-packages (from langsmith<0.4,>=0.1.125-
>langchain-core<0.4,>=0.1->langgraph) (1.0.0)
Requirement already satisfied: zstandard<0.24.0,>=0.23.0 in
/usr/local/lib/python3.11/dist-packages (from langsmith<0.4,>=0.1.125-
>langchain-core<0.4,>=0.1->langgraph) (0.23.0)
Requirement already satisfied: annotated-types>=0.6.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0,>=2.5.2-
>langchain-core<0.4,>=0.1->langgraph) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.1 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0,>=2.5.2-
>langchain-core<0.4,>=0.1->langgraph) (2.33.1)
Requirement already satisfied: typing-inspection>=0.4.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic<3.0.0,>=2.5.2-
```

```
>langchain-core<0.4,>=0.1->langgraph) (0.4.0)
Requirement already satisfied: charset-normalizer<4,>=2 in
/usr/local/lib/python3.11/dist-packages (from requests<3,>=2-
>langsmith<0.4,>=0.1.125->langchain-core<0.4,>=0.1->langgraph) (3.4.1)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/usr/local/lib/python3.11/dist-packages (from requests<3,>=2-
>langsmith<0.4,>=0.1.125->langchain-core<0.4,>=0.1->langgraph) (2.3.0)
Requirement already satisfied: sniffio>=1.1 in
/usr/local/lib/python3.11/dist-packages (from anyio->httpx>=0.25.2-
>langgraph-sdk<0.2.0,>=0.1.42->langgraph) (1.3.1)
Downloading langgraph-0.3.25-py3-none-any.whl (142 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 142.4/142.4 kB 6.6 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 42.0/42.0 kB 3.0 MB/s eta
0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 47.2/47.2 kB 3.2 MB/s eta
0:00:00
anylinux_2_17_x86_64.manylinux2014_x86_64.whl (194 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 194.8/194.8 kB 11.9 MB/s eta
0:00:00
sgpack-1.9.1-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl (223 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 223.6/223.6 kB 14.4 MB/s eta
0:00:00
sgpack, langgraph-sdk, langgraph-checkpoint, langgraph-prebuilt,
langgraph
Successfully installed langgraph-0.3.25 langgraph-checkpoint-2.0.24
langgraph-prebuilt-0.1.8 langgraph-sdk-0.1.61 ormsgpack-1.9.1 xxhash-
3.5.0

!pip install grandalf

Collecting grandalf
  Downloading grandalf-0.8-py3-none-any.whl.metadata (1.7 kB)
Requirement already satisfied: pyparsing in
/usr/local/lib/python3.11/dist-packages (from grandalf) (3.2.3)
Downloading grandalf-0.8-py3-none-any.whl (41 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 41.8/41.8 kB 2.7 MB/s eta
0:00:00

!pip install gradio
!python app.py

Collecting gradio
  Downloading gradio-5.23.3-py3-none-any.whl.metadata (16 kB)
Collecting aiofiles<24.0,>=22.0 (from gradio)
  Downloading aiofiles-23.2.1-py3-none-any.whl.metadata (9.7 kB)
Requirement already satisfied: anyio<5.0,>=3.0 in
/usr/local/lib/python3.11/dist-packages (from gradio) (4.9.0)
Requirement already satisfied: fastapi<1.0,>=0.115.2 in
```

```
/usr/local/lib/python3.11/dist-packages (from gradio) (0.115.12)
Collecting ffmpy (from gradio)
  Downloading ffmpy-0.5.0-py3-none-any.whl.metadata (3.0 kB)
Collecting gradio-client==1.8.0 (from gradio)
  Downloading gradio_client-1.8.0-py3-none-any.whl.metadata (7.1 kB)
Collecting groovy~=0.1 (from gradio)
  Downloading groovy-0.1.2-py3-none-any.whl.metadata (6.1 kB)
Requirement already satisfied: httpx>=0.24.1 in
/usr/local/lib/python3.11/dist-packages (from gradio) (0.28.1)
Requirement already satisfied: huggingface-hub>=0.28.1 in
/usr/local/lib/python3.11/dist-packages (from gradio) (0.30.1)
Requirement already satisfied: jinja2<4.0 in
/usr/local/lib/python3.11/dist-packages (from gradio) (3.1.6)
Requirement already satisfied: markupsafe<4.0,>=2.0 in
/usr/local/lib/python3.11/dist-packages (from gradio) (3.0.2)
Requirement already satisfied: numpy<3.0,>=1.0 in
/usr/local/lib/python3.11/dist-packages (from gradio) (2.0.2)
Requirement already satisfied: orjson~=3.0 in
/usr/local/lib/python3.11/dist-packages (from gradio) (3.10.16)
Requirement already satisfied: packaging in
/usr/local/lib/python3.11/dist-packages (from gradio) (24.2)
Requirement already satisfied: pandas<3.0,>=1.0 in
/usr/local/lib/python3.11/dist-packages (from gradio) (2.2.2)
Requirement already satisfied: pillow<12.0,>=8.0 in
/usr/local/lib/python3.11/dist-packages (from gradio) (11.1.0)
Requirement already satisfied: pydantic<2.12,>=2.0 in
/usr/local/lib/python3.11/dist-packages (from gradio) (2.11.2)
Collecting pydub (from gradio)
  Downloading pydub-0.25.1-py2.py3-none-any.whl.metadata (1.4 kB)
Collecting python-multipart>=0.0.18 (from gradio)
  Downloading python_multipart-0.0.20-py3-none-any.whl.metadata (1.8
kB)
Requirement already satisfied: pyyaml<7.0,>=5.0 in
/usr/local/lib/python3.11/dist-packages (from gradio) (6.0.2)
Collecting ruff>=0.9.3 (from gradio)
  Downloading ruff-0.11.4-py3-none-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (25 kB)
Collecting safehttpx<0.2.0,>=0.1.6 (from gradio)
  Downloading safehttpx-0.1.6-py3-none-any.whl.metadata (4.2 kB)
Collecting semantic-version~=2.0 (from gradio)
  Downloading semantic_version-2.10.0-py2.py3-none-any.whl.metadata
(9.7 kB)
Requirement already satisfied: starlette<1.0,>=0.40.0 in
/usr/local/lib/python3.11/dist-packages (from gradio) (0.46.1)
Collecting tomlkit<0.14.0,>=0.12.0 (from gradio)
  Downloading tomlkit-0.13.2-py3-none-any.whl.metadata (2.7 kB)
Requirement already satisfied: typer<1.0,>=0.12 in
/usr/local/lib/python3.11/dist-packages (from gradio) (0.15.2)
Requirement already satisfied: typing-extensions~=4.0 in
```

```
/usr/local/lib/python3.11/dist-packages (from gradio) (4.13.1)
Requirement already satisfied: uvicorn>=0.14.0 in
/usr/local/lib/python3.11/dist-packages (from gradio) (0.34.0)
Requirement already satisfied: fsspec in
/usr/local/lib/python3.11/dist-packages (from gradio-client==1.8.0-
>gradio) (2025.3.2)
Requirement already satisfied: websockets<16.0,>=10.0 in
/usr/local/lib/python3.11/dist-packages (from gradio-client==1.8.0-
>gradio) (15.0.1)
Requirement already satisfied: idna>=2.8 in
/usr/local/lib/python3.11/dist-packages (from anyio<5.0,>=3.0->gradio)
(3.10)
Requirement already satisfied: sniffio>=1.1 in
/usr/local/lib/python3.11/dist-packages (from anyio<5.0,>=3.0->gradio)
(1.3.1)
Requirement already satisfied: certifi in
/usr/local/lib/python3.11/dist-packages (from httpx>=0.24.1->gradio)
(2025.1.31)
Requirement already satisfied: httpcore==1.* in
/usr/local/lib/python3.11/dist-packages (from httpx>=0.24.1->gradio)
(1.0.7)
Requirement already satisfied: h11<0.15,>=0.13 in
/usr/local/lib/python3.11/dist-packages (from httpcore==1.*-
>httpx>=0.24.1->gradio) (0.14.0)
Requirement already satisfied: filelock in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.28.1-
>gradio) (3.18.0)
Requirement already satisfied: requests in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.28.1-
>gradio) (2.32.3)
Requirement already satisfied: tqdm>=4.42.1 in
/usr/local/lib/python3.11/dist-packages (from huggingface-hub>=0.28.1-
>gradio) (4.67.1)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas<3.0,>=1.0-
>gradio) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.11/dist-packages (from pandas<3.0,>=1.0-
>gradio) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.11/dist-packages (from pandas<3.0,>=1.0-
>gradio) (2025.2)
Requirement already satisfied: annotated-types>=0.6.0 in
/usr/local/lib/python3.11/dist-packages (from pydantic<2.12,>=2.0-
>gradio) (0.7.0)
Requirement already satisfied: pydantic-core==2.33.1 in
/usr/local/lib/python3.11/dist-packages (from pydantic<2.12,>=2.0-
>gradio) (2.33.1)
Requirement already satisfied: typing-inspection>=0.4.0 in
```

```
/usr/local/lib/python3.11/dist-packages (from pydantic<2.12,>=2.0->gradio) (0.4.0)
Requirement already satisfied: click>=8.0.0 in /usr/local/lib/python3.11/dist-packages (from typer<1.0,>=0.12->gradio) (8.1.8)
Requirement already satisfied: shellingham>=1.3.0 in /usr/local/lib/python3.11/dist-packages (from typer<1.0,>=0.12->gradio) (1.5.4)
Requirement already satisfied: rich>=10.11.0 in /usr/local/lib/python3.11/dist-packages (from typer<1.0,>=0.12->gradio) (13.9.4)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.8.2->pandas<3.0,>=1.0->gradio) (1.17.0)
Requirement already satisfied: markdown-it-py>=2.2.0 in /usr/local/lib/python3.11/dist-packages (from rich>=10.11.0->typer<1.0,>=0.12->gradio) (3.0.0)
Requirement already satisfied: pygments<3.0.0,>=2.13.0 in /usr/local/lib/python3.11/dist-packages (from rich>=10.11.0->typer<1.0,>=0.12->gradio) (2.18.0)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0.28.1->gradio) (3.4.1)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests->huggingface-hub>=0.28.1->gradio) (2.3.0)
Requirement already satisfied: mdurl~=0.1 in /usr/local/lib/python3.11/dist-packages (from markdown-it-py>=2.2.0->rich>=10.11.0->typer<1.0,>=0.12->gradio) (0.1.2)
Downloading gradio-5.23.3-py3-none-any.whl (46.5 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 46.5/46.5 MB 14.4 MB/s eta 0:00:00
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 322.2/322.2 kB 18.3 MB/s eta 0:00:00
ultipart-0.0.20-py3-none-any.whl (24 kB)
Downloading ruff-0.11.4-py3-none-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (11.3 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 11.3/11.3 MB 60.2 MB/s eta 0:00:00
antic_version-2.10.0-py2.py3-none-any.whl (15 kB)
Downloading tomlkit-0.13.2-py3-none-any.whl (37 kB)
Downloading ffmpy-0.5.0-py3-none-any.whl (6.0 kB)
Downloading pydub-0.25.1-py2.py3-none-any.whl (32 kB)
Installing collected packages: pydub, tomlkit, semantic-version, ruff, python-multipart, groovy, ffmpy, aiofiles, safehttpx, gradio-client, gradio
Successfully installed aiofiles-23.2.1 ffmpy-0.5.0 gradio-5.23.3 gradio-client-1.8.0 groovy-0.1.2 pydub-0.25.1 python-multipart-0.0.20 ruff-0.11.4 safehttpx-0.1.6 semantic-version-2.10.0 tomlkit-0.13.2
```

```
python3: can't open file '/content/app.py': [Errno 2] No such file or
directory
```

```
!pip install pymupdf
```

```
Collecting pymupdf
  Downloading pymupdf-1.25.5-cp39-abi3-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (3.4 kB)
Downloading pymupdf-1.25.5-cp39-abi3-
manylinux2014_x86_64.manylinux_2_17_x86_64.whl (20.0 MB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 20.0/20.0 MB 19.5 MB/s eta
0:00:00
updf
Successfully installed pymupdf-1.25.5
```

```
import gradio as gr
import tempfile
import os
import requests
import faiss
import numpy as np
from typing import TypedDict, List, Optional

from langgraph.graph import StateGraph
from langchain_core.runnables import RunnableLambda
from langchain.docstore.document import Document
from langchain_community.embeddings import HuggingFaceEmbeddings
from langchain_community.vectorstores.faiss import FAISS
from langchain.docstore import InMemoryDocstore
from langchain.document_loaders import PyMuPDFLoader
from together import Together
```

##Agents##

🧠 Pipeline Flow

1. 🔍 **SearchAgent** – Uses SerpAPI to fetch top Google results
2. 📄 **PDFLoaderAgent** – Loads and splits uploaded PDFs into text
3. 🧠 **EmbedAgent** – Embeds all documents using MiniLM
4. 📌 **RetrieveAgent** – Uses FAISS to find top-k relevant chunks
5. 📝 **SummarizeAgent** – Calls Together.ai LLaMA-4 with user prompt

The entire workflow is managed via LangGraph's stateful graph execution.

```
def search_agent(state):
    print("🔍 [SearchAgent] Fetching from SerpAPI...")
    query = state.get("query", "")
    api_key = state.get("serpapi_key", "")
    if not query or not api_key:
```

```
        print("⚠ [SearchAgent] Missing query or API key.")
        return state

    params = {"engine": "google", "q": query, "api_key": api_key,
"num": 20}
    resp = requests.get("https://serpapi.com/search", params=params)
    data = resp.json()
    snippets = [item.get("snippet", "") for item in
data.get("organic_results", []) if item.get("snippet")]
    docs = [Document(page_content=s, metadata={"source": "serpapi",
"rank": i}) for i, s in enumerate(snippets)]

    state["documents"] = docs  # Replace any existing docs
    print(f" [SearchAgent] Retrieved {len(docs)} snippets.")
    return state

def pdf_loader_agent(state):
    print(" [PDFLoaderAgent] Loading PDF...")
    pdf_path = state.get("pdf_path")

    # Ensure 'documents' key exists
    if "documents" not in state:
        state["documents"] = []

    if pdf_path and os.path.exists(pdf_path):
        loader = PyMuPDFLoader(pdf_path)
        pdf_docs = loader.load()
        state["documents"].extend(pdf_docs)
        print(f" [PDFLoaderAgent] Loaded {len(pdf_docs)} pages from
PDF.")
    else:
        print("⚠ [PDFLoaderAgent] No PDF found.")
    return state

def embed_agent(state):
    print(" [EmbedAgent] Embedding and indexing...")
    docs = state["documents"]
    embed_model = HuggingFaceEmbeddings(model_name="all-MiniLM-L6-v2")
    vectors = embed_model.embed_documents([doc.page_content for doc in
docs])
    vectors_np = np.array(vectors).astype("float32")
    d = vectors_np.shape[1]
    nlist = min(5, len(docs))
    quantizer = faiss.IndexFlatL2(d)
    index = faiss.IndexIVFFlat(quantizer, d, nlist)
    index.train(vectors_np)
    index.add(vectors_np)

    index_to_docstore_id = {i: str(i) for i in range(len(docs))}
    docstore = InMemoryDocstore({str(i): doc for i, doc in
```

```
enumerate(docs)})
    vectorstore = FAISS(embed_model, index, docstore,
index_to_docstore_id)
    state["vectorstore"] = vectorstore
    print("✅ [EmbedAgent] Indexing complete.")
    return state

def retrieve_agent(state):
    print("🔍 [RetrieveAgent] Retrieving relevant documents...")
    vectorstore = state["vectorstore"]
    query = state.get("query", "")
    top_k = state.get("top_k", 3)
    results = vectorstore.similarity_search(query, k=top_k)
    state["retrieved_docs"] = results
    print(f"✅ [RetrieveAgent] Retrieved {len(results)} docs.")
    return state

def summarize_agent(state):
    print("📝 [SummarizeAgent] Generating summary (based on user
instructions)...")

    client = Together(api_key=state["together_api_key"])
    retrieved_texts = [doc.page_content for doc in
state["retrieved_docs"]]

    user_query = state["query"]

    prompt = (
        f"You are a helpful AI research assistant.\n\n"
        f"The user has asked: \"{user_query}\"\n\n"
        f"Below is the retrieved research content. Please summarize it
according to the user's request.\n\n"
        f"### Research Content:\n\n" + "\n\n".join(retrieved_texts)
    )

    response = client.chat.completions.create(
        model="meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8",
        messages=[
            {"role": "system", "content": "You are a helpful
scientific summarizer."},
            {"role": "user", "content": prompt}
        ],
        max_tokens=4096,  # allow long summaries
        temperature=0.7,
    )

    summary = response.choices[0].message.content
    print("✅ [SummarizeAgent] Summary generated.")
    return {**state, "summary": summary}
```

##LangGraph definition##

Pipeline Architecture

[User Query / PDF Upload] ↓ SearchAgent (via SerpAPI) ↓ PDFLoaderAgent (if PDF exists) ↓ EmbedAgent (MiniLM + FAISS) ↓ RetrieveAgent (Top-k retrieval) ↓ SummarizeAgent (Together.ai LLaMA-4) ↓ [Summary Output]

How to Use

1. Enter a query in the Gradio interface (e.g., "AutoML in 2024, 500-word summary")
2. Optionally upload a research PDF file
3. Provide your API keys:
   - `SerpAPI Key` (for Google search)
   - `Together.ai Key` (for LLaMA-4 summarization)
4. Click Submit and wait for the generated response

API Keys Required

- **SerpAPI Key**: Get from serpapi.com
- **Together API Key**: Get from platform.together.xyz

Keys can be entered directly into the Gradio app fields, or managed via `.env` (if running locally).

```
class RAGState(TypedDict):
    query: str
    serpapi_key: str
    together_api_key: str
    top_k: int
    pdf_path: Optional[str]
    documents: List[Document]
    vectorstore: FAISS
    retrieved_docs: List[Document]
    summary: str

graph = StateGraph(state_schema=RAGState)
graph.add_node("SearchAgent", RunnableLambda(search_agent))
graph.add_node("PDFLoaderAgent", RunnableLambda(pdf_loader_agent))
graph.add_node("EmbedAgent", RunnableLambda(embed_agent))
graph.add_node("RetrieveAgent", RunnableLambda(retrieve_agent))
graph.add_node("SummarizeAgent", RunnableLambda(summarize_agent))

graph.set_entry_point("SearchAgent")
graph.add_edge("SearchAgent", "PDFLoaderAgent")
graph.add_edge("PDFLoaderAgent", "EmbedAgent")
graph.add_edge("EmbedAgent", "RetrieveAgent")
graph.add_edge("RetrieveAgent", "SummarizeAgent")
graph.set_finish_point("SummarizeAgent")
```

```
dag = graph.compile()
```

##Gradio Interface##

Example Prompts

- "Summarize this PDF in 500 words with Introduction, Methodology, and Results"
- "What are the major NAS techniques in 2024? Give bullet points"
- "List pros and cons of AutoML methods from this paper"

Sample Output (Truncated)

- The paper introduces a hybrid NAS method combining reinforcement learning and gradient descent.
- It compares 3 different architectures: X, Y, Z...
- Key metrics: Accuracy = 92.4%, Latency = 0.8ms

Results

- Achieved summaries of up to 1000 words respecting structural prompts
- Enabled hybrid PDF + Search summarization
- Fast similarity search over embedded corpus using IVF-indexed FAISS

```
def run_rag_pipeline(query, serpapi_key, together_api_key, pdf_file):
    try:
        print("\n Starting RAG pipeline...")
        pdf_path = None

        if pdf_file is not None:
            pdf_path = pdf_file.name
            print(f" Using uploaded PDF at: {pdf_path}")

        state = {
            "query": query if query else "placeholder",
            "serpapi_key": serpapi_key,
            "together_api_key": together_api_key,
            "top_k": 3,
            "pdf_path": pdf_path,
        }

        output = dag.invoke(state)

        if not output.get("summary"):
            return "⚠ No summary generated. Check if the PDF or search
returned useful content."

        return output["summary"]
```

```
    except Exception as e:
        import traceback
        traceback.print_exc()
        return f"❌ Error: {str(e)}"

# Launch Gradio app
gr.Interface(
    fn=run_rag_pipeline,
    inputs=[
        gr.Textbox(label="🔍 Research Query", placeholder="e.g., NAS
techniques in 2024"),
        gr.Textbox(label="🔑 SerpAPI Key", type="password"),
        gr.Textbox(label="🔑 Together API Key", type="password"),
        gr.File(label="📄 Upload Research PDF (Optional)",
file_types=[".pdf"])
    ],
    outputs=gr.Textbox(label="📝 Summary Output", lines=15),
    title="🤖 Research Assistant Chatbot (LangGraph + LLaMA)",
    description="Combines SerpAPI + PDF + FAISS + Together.ai for
autonomous research summaries."
).launch()
```

```
Running Gradio in a Colab notebook requires sharing enabled.
Automatically setting `share=True` (you can turn this off by setting
`share=False` in `launch()` explicitly).

Colab notebook detected. To show errors in colab notebook, set
debug=True in launch()
* Running on public URL: https://29b781b604c90ad629.gradio.live

This share link expires in 72 hours. For free permanent hosting and
GPU upgrades, run `gradio deploy` from the terminal in the working
directory to deploy to Hugging Face Spaces
(https://huggingface.co/spaces)

<IPython.core.display.HTML object>
```