Assignment 5 (50 points)

Agents are an emerging field thats use reflection, tools, planning, and multi agent collaboration

In this assignment, we will build a research agent. We will use serverless LLM endpoints. To get started, you create an account with Together AI or Anthropic. They should provide you with a few dollars worth of credits that should be enough to complete the assignment. You are free to choose any other provider such as OpenAI, Mistral, Fireworks, or Groq. I encourage you to play around with different models to get a feel for how they work. For this assignment, the API usage cost should be around a couple dollars. Depending on the model you choose and how many attempts you use, it may be a couple cents. For OpenAI, Anthropic, and Mistral, double check what model you are using. The flagship models are significantly more expensive than the smaller models (pricing between models varies by 50x). For the purposes of this assignment, it is sufficient to use the smallest/cheapest models.

TogetherAI, Fireworks, and Groq run open source models. For these, it's better to run the mid-large tier models. Mixtral is a good place to start. It's good to play around with different models.

Many providers are able to use OpenAI's client library, but some do not (like Anthropic). Use whatever makes sense.

You can run this on Colab with a CPU, or locally and submit the Jupyter notebook as your submission. Since we are using third party providers for the LLMs, we will not load the model locally. If you run on Colab, take special care to not leak your API key. Here's an example of how to properly use secrets in Colab.

Research Agent

Build an LLM-based research agent that can take a research topic, find relevant information, and generate a short summary (~1 paragraph) on the given topic.

Tools to Implement (20 points, 4 points each):

1. Topic Breakdown Tool: Create a tool that takes a broad research topic and breaks it down into smaller, more focused subtopics or subqueries. You can use an LLM to generate these subtopics based on the main topic.

2. Query Expansion Tool: Develop a tool to expand the subqueries generated by the Topic Breakdown Tool. The tool should generate related keywords, synonyms, and phrases to enhance the search results.

3. Search Tool: Create a wrapper around the You API or Brave Search API, Serper.dev. Please note that the free tier is 1000 queries/month. Consider creating a mock while developing, and switch to actually call the You API once the agent is more stable. Additionally, consider caching the search results.

4. Critique Tool: Create a tool that critiques the summary, and offers suggestions of how to improve and potentially other relevant topics to search for.

5. Summarizer Tool *(optional)*: Create a tool that takes some input and summarizes its content using an LLM.

Workflow (30 points)

Implement an agent workflow that uses all of these tools. In the agent workflow, the agent should be provided with all the tools and it should decide which tool to use. For the individual tool implementations, if you use a call to an LLM you do not need to provide any tools.

Sample Agent Workflow:

1. The agent receives a research topic from the user.
2. It uses the Topic Breakdown Tool to generate subtopics or subqueries.
3. The Query Expansion Tool expands the subqueries with related keywords and phrases.
4. The Search Tool uses the expanded queries and subqueries to gather relevant information from various sources.
5. The agent generates the summary incorporating the search results. *(optional)*
6. The agent critiques the summary, and improves the results. *(optional)*
7. The agent presents the final summary to the user.

The sample workflow is the minimum implementation requirement. Feel free to add more tools, add loops in the workflow, etc.