CS 435 GenAI: Lecture 11 - Advanced Topics Class Assignment Continuation

As discussed and scaffolded during our previous class session, you are required to complete the following tasks in preparation for the next class. Please come ready to present your work and participate actively in the discussion (technology stack could vary, suggestions below).

Tasks to Complete:

1. Well Grounded Semantic Search and Retrieval (RAG Pipeline)

Complete the implementation of the semantic search and retrieval assignment using PDF documents as outlined. Your solution should be demonstrated clearly in a Google Colab notebook and should address the following:

- Text extraction from PDFs (minimum 3 documents)
- Embedding generation with SentenceTransformer (`all-MiniLM-L6-v2`)
- FAISS GPU-based vector database creation
- Semantic search functionality with top-5 retrieved results
- Reranking based on cosine similarity
- Answer generation using GPT-2 (Hugging Face)
- Clearly displayed results including query, generated answer, and metadata for top-3 retrieved results (PDF name, page number, similarity scores, excerpts)

Prepare at least 3 example queries for class demonstration.

2. Question Validation with NeMo Guardrails and CoLang

Finish the implementation of your NeMo Guardrails assignment. The solution should clearly demonstrate:

- NeMo Guardrails setup
- Defined validation criteria document with 5 specific brand-damaging question examples
- At least 10 CoLang rules preventing brand-damaging questions (negative assumptions, defamatory language, misinformation)
- Integration with a Generative AI model (GPT-3/GPT-4)
- Comprehensive testing with at least 5 challenging test questions
- Clear documentation of results (before and after guardrails, rule effectiveness explanation)

3. AWS Guardrails Reading

Review and familiarize yourself with AWS Guardrails for Amazon Bedrock at the following link:
`https://aws.amazon.com/bedrock/guardrails/`
Come prepared to discuss the key features and your understanding of the AWS Guardrails framework in the next class.

4. HuggingFace LLaMA Implementation

Implement a basic inference demonstration using Hugging Face's LLaMA model. You should:

- Set up Hugging Face transformers in Colab
- Load the LLaMA model
- Execute a simple prompt-based text generation
- Document clearly the steps and your observations regarding the model's performance

Be ready to demonstrate and discuss your findings.

Ensure your work is well-organized, documented, and ready for live class demonstration and interactive discussion.