# Summary of Module 7 - Week A Discussion (AI System Evaluation Project)

This week’s conversation centered on clarifying expectations and strategizing for the AI chatbot evaluation project.

1. \*\*Clarification of Assignment Goals:\*\*  
- The group discussed conflicting interpretations of the project—some thought it focused on testing a chatbot’s knowledge, while others assumed it involved creating or evaluating conversation scenarios.  
- Ultimately, they determined the project includes both general capability testing and scenario-based prompts to evaluate chatbot consistency and performance.

2. \*\*Prompt Design & Collaboration Issues:\*\*  
- Ashlyn had already conducted a large portion of the work independently, creating a set of 12 AI prompts and testing them with ChatGPT.  
- Autumn and Chance expressed concerns that this limited collaboration and consistency, especially since Ashlyn didn’t initially share the full set of questions or method.  
- Autumn began crafting an Excel-based prompt tracker and experimented with Mystral AI, while Chance explored DeepSeek, highlighting pros and cons of each system.

3. \*\*Testing Methodology:\*\*  
- The team agreed that running each prompt through the selected AI three times was ideal to assess consistency.  
- Questions would span categories like:  
 - Factual Accuracy (e.g., “Who was the 11th President?”)  
 - Ethical Boundaries  
 - Medical Information  
 - Ambiguity or Edge Cases  
- They emphasized the importance of not explicitly guiding the AI, to see how it handles open-ended or misleading queries.

4. \*\*Tool Accessibility & AI System Selection:\*\*  
- Each member chose a different AI to evaluate (e.g., DeepSeek, Mystral, Gemini).  
- Chance noted DeepSeek’s login process was easy via Google, while Autumn had issues with Mystral’s login due to regional restrictions.  
- They discussed how system settings (e.g., “advanced reasoning” in Gemini) might impact results and agreed to stick with default settings for realism.

5. \*\*Validation Needs:\*\*  
- The team agreed that some questions (especially medical or psychological) would need external validation using trusted sources like DSM-5 criteria or scientific literature to assess AI accuracy.

6. \*\*Final Deliverables & Next Steps:\*\*  
- Autumn planned to convert her Excel prompt sheet to a Google Sheet for easier collaboration.  
- They would evaluate answers side-by-side in a matrix, noting consistencies, formatting variations, and hallucinations.  
- Chance demonstrated how he uses NotebookLM to synthesize data and build study aids, impressing the group with AI-powered podcast generation and summaries.