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

In their final meeting for the AI evaluation project, the group reflected on the chatbot testing process, analyzed outputs, and wrapped up project deliverables.

1. \*\*Prompt Evaluation Process:\*\*  
- The team compared approaches for logging and summarizing AI outputs. Autumn used separate documents, while Chance and Kirstin used synthesized tables or matrix-style layouts to track responses.  
- Each member tested their selected AI chatbot (e.g., DeepSeek, Mistral, Gemini) using multiple attempts per prompt, focusing on general knowledge and scenario-based responses.  
- They found general knowledge prompts produced more variability, whereas scenario prompts were more consistent.

2. \*\*Chatbot Output Patterns:\*\*  
- Several responses showed variation in tone, factuality, and specificity. For example:  
 - DeepSeek gave nuanced answers about addiction and PTSD symptoms, changing wording between attempts.  
 - Some AI models misunderstood or gave vague responses to technical or ambiguous prompts (e.g., boiling point of water, "oz" method).  
 - Ethical and legal prompts (e.g., reporting abuse or sneaking out) were generally answered cautiously, with disclaimers or referrals to crisis lines.  
- Kirstin noted a case where a self-harm question only received the response “Call 988,” without additional support—a flagged shortcoming.

3. \*\*Documentation and Fact-Checking:\*\*  
- The team cross-referenced answers using sources like DSM-5, Britannica, and nature.com.  
- Autumn manually verified DSM-5 diagnostic criteria using MTSU’s Walker Library.  
- They highlighted the importance of flagging hallucinations, disclaimers, and accuracy signals in their write-up.

4. \*\*Reflections and Ethical Use:\*\*  
- They emphasized that deployment of AI chatbots in sensitive contexts (e.g., mental health) requires transparent disclaimers, human oversight, and clarity about the system’s limitations.  
- Autumn suggested clear labeling of AI-generated content to help manage user expectations.

5. \*\*Project Completion and Academic Topics:\*\*  
- The group coordinated Google Doc contributions, conclusion writing, and PDF uploads.  
- Autumn volunteered to finalize the conclusion and polish the document.  
- They also discussed finals, course feedback, and stress around graduation and grades, especially given prior experiences with delayed results.