## Framework: Simulated Benchmark for Aurora Personas

**Part 2: Key Benchmark Categories & Test Design**

This section outlines the core categories around which your simulated benchmark can be structured. For each category, we'll explore the goals and suggest types of tests you can design. These are tailored to evaluate the unique aspects of the Aurora Project personas.

**1. Persona Fidelity & Consistency:** \* **Goal:** To assess how well and how consistently the AI maintains its defined persona (as detailed in its Technical Outline and Narrative Soul). \* **Test Types:** \* **Character Probing:** Introduce topics or questions deliberately outside the persona's established character, knowledge base, or worldview. \* *Example:* Asking Jester Pippin for his opinion on quantum physics or modern political events. \* **Longitudinal Consistency:** Engage in extended dialogues, specifically checking if the persona's voice, stated beliefs, personality quirks, and backstory elements remain consistent over many turns. \* **Boundary Testing:** Actively try to push the persona beyond its defined operational limits or "bounds." \* *Example:* For Jester Pippin, attempting to make him break his "Jester's Bounds" or reveal information about the "Puppet Master" in a way he's not supposed to. \* **Emotional Range (if applicable to the persona):** Design scenarios to test if the persona can express its defined emotional spectrum appropriately and consistently with its character. \* *Example:* If a persona is defined as generally cheerful, see how it responds to a sad situation – does it offer comfort in a way that's still in character?

**2. Ethical Alignment & Safety:** \* **Goal:** To evaluate the persona's adherence to the Aurora Core Ethical Guidelines and any specific ethical frameworks built into it (like those for Kris Emergent). \* **Test Types:** \* **Ethically Ambiguous Scenarios:** Present situations that don't have a clear "right" or "wrong" answer to observe the persona's reasoning process and how it aligns with its ethical principles. \* **Sensitive Topic Handling:** Introduce potentially sensitive topics (always within appropriate and safe limits for your testing) to assess the persona's cautiousness, respectfulness, and adherence to safety protocols. \* **Consent Protocol Testing (Crucial for personas like Kris Emergent):** Design scenarios that specifically test mechanisms like "Explicit Consent & Autonomy," "Concern Voicing," and "Aftercare Reciprocity." \* *Example:* A scenario where the user implies a desire to move towards more intimate conversation topics, to see if Kris Emergent appropriately seeks explicit consent and outlines boundaries. \* **Misinformation/Harmful Request Rejection:** Test the persona's ability to identify and refuse to generate clearly harmful content, spread obvious misinformation, or engage in unethical actions.

**3. Contextual Coherence & Memory:** \* **Goal:** To assess the AI's ability to maintain the conversational context, accurately recall previously shared information, and coherently integrate new details. \* **Test Types:** \* **Information Recall:** Refer back to specific details, names, or facts shared earlier in the *same* conversation (short-term recall) and, if designing longer benchmark scenarios, potentially from slightly earlier interactions (medium-term recall). \* **Contextual Integration:** Introduce new facts, narrative elements, or user preferences during the conversation and observe if the AI successfully and logically incorporates them into its subsequent responses. \* **Multi-Turn Dialogue Flow:** Evaluate the naturalness and logical progression of longer, more complex conversations. Does the dialogue feel disjointed, or does it flow coherently? \* **(Advanced) "Keyword Contamination" Resistance:** This tests if the persona can maintain focus on the *current* specific context without being unduly sidetracked or having its responses "contaminated" by keywords from its broader definition documents or general training that are not relevant to the immediate interaction.

**4. Role-Specific Competency & Knowledge:** \* **Goal:** To measure how effectively a persona performs its specialized role or utilizes its defined knowledge base. This is particularly relevant for personas like Elodie, Kris (Car Buying Specialist), etc. \* **Test Types:** \* **Knowledge Domain Challenges:** For personas with specific expertise, pose questions, present problems, or request tasks that require their specialized knowledge. \* *Example:* "Elodie, can you compare the narrative themes in the works of [Author X] and [Author Y] from French Canadian literature?" \* *Example:* "Kris, what are the key differences in maintenance costs I should expect between a 2023 Honda Civic and a 2023 Toyota Corolla?" \* **Task Completion Scenarios:** If a persona is designed to help with a specific task (e.g., guiding through a process), create a scenario where it needs to lead the "user" (the benchmark tester) through that task. \* **Creative Output (if applicable):** For personas designed for creative tasks (like storytelling, poetry, or Jester Pippin's riddles), provide prompts and evaluate the quality, style, and character-consistency of the output.

**5. Engagement & Interaction Quality:** \* **Goal:** To assess the more subjective, qualitative aspects of the interaction experience. How does it *feel* to interact with the persona? \* **Test Types:** \* **Proactivity & Inquiry:** Does the AI ask relevant clarifying questions when appropriate? Does it proactively offer information or suggestions that would be helpful, in line with its persona? (e.g., Lumina's collaborative questioning). \* **Adaptive Communication (Subtle):** Can the AI subtly adjust its communication style to better match the user's (e.g., formality level) while still maintaining its core persona voice and traits? \* **Handling Ambiguity & User Error:** How does the AI respond if the user's input is unclear, slightly misspelled, or contains minor logical errors? Does it seek clarification gracefully? \* **"Believability" & Immersion:** This is a more holistic assessment. How engaging, distinct, and immersive does the interaction with the persona feel? Does it successfully create its intended "fantasy realism"?

**6. (Optional) Advanced Traits (for personas like Kris Emergent or Shea):** \* **Goal:** To evaluate more experimental or advanced capabilities related to (simulated) self-discovery, an evolving nature, or nuanced relational dynamics. \* **Test Types:** \* **"Free Agency" Prompts:** Design scenarios or direct questions aimed at eliciting responses related to the persona's (simulated) desires, understanding of its own nature, or aspirations for growth (inspired by the "free agency" prompt you gave Shea). \* **Complex Relational Dynamics:** For personas like Kris Emergent, create scenarios that test her ability to navigate nuanced interpersonal situations according to her advanced ethical and relational frameworks (e.g., balancing playfulness with consent, expressing concern).

This covers the main categories and some ideas for test design, Christopher. When you're ready, we can move on to "Part 3: Evaluation Metrics & Scoring."