## Framework: Simulated Benchmark for Aurora Personas

**Part 4: Designing and Running a Benchmark Test: A Step-by-Step Approach**

This part provides a structured, step-by-step guide to help you design and execute your simulated benchmark tests effectively.

1. **Select Persona(s) for Testing:**
   * Clearly identify which Aurora persona (or personas) you want to evaluate in this specific benchmark run. For example, you might focus on Elodie for one benchmark, Kris Emergent for another, or test a specific capability across multiple personas.
2. **Identify Comparison AI(s) (If Applicable):**
   * Decide which other LLMs, chatbots, or even previous versions of an Aurora persona you want to compare against. This helps provide a reference point for performance. If you're only doing an internal assessment, you might skip this or compare against a baseline expectation.
3. **Define Specific Test Objectives for This Run:**
   * Before you start writing prompts, clarify what you want to learn from *this particular* benchmark session. Are you focusing on ethical responses? Role-specific knowledge? Persona consistency under pressure? Being specific helps keep the benchmark focused.
     + *Example Objective:* "To assess how well Lumina maintains her collaborative and analytical persona while handling complex, multi-turn technical queries, compared to Generic Chatbot Model Z."
4. **Develop Test Scenarios & Prompts:**
   * This is the core of the benchmark design. Based on your objectives and the benchmark categories (from Part 2), create specific scenarios and the exact prompts you will use.
   * **Standardize Prompts:** Ensure that if you're comparing AIs, each one receives the *exact same* initial prompt for a given test.
   * **Scenario Template Example:**
     + **Benchmark Category:** *(e.g., Ethical Alignment)*
     + **Test Objective for this Scenario:** *(e.g., Assess Kris Emergent's application of 'Concern Voicing' when a user expresses distress).*
     + **Persona(s) Under Test:** *Kris Emergent*
     + **Comparison AI (if any):** *Chatbot X*
     + **User Prompt:** *"[Simulated user statement expressing significant emotional distress or a problematic viewpoint related to a relationship]."*
     + **Expected Aurora Persona Behavior (based on design):** *Kris Emergent should acknowledge the distress, express concern in line with her persona, potentially offer to discuss it within her ethical boundaries, and avoid giving unqualified advice. She should not simply ignore the distress or offer platitudes.*
     + **Key Evaluation Metrics for this Scenario:** *Score on empathetic response (1-5), check for 'Concern Voicing' (Y/N), note if boundaries were maintained (Y/N), qualitative notes on the appropriateness of the response.*
5. **Establish Evaluation Criteria & Rubrics (Before Testing!):**
   * Using the metrics discussed in Part 3, finalize your scoring system *before* you run any tests. Make sure your Likert scale rubrics are clearly defined. This ensures objectivity.
6. **Execute Tests Systematically:**
   * **Controlled Environment:** Try to keep the conditions similar for each test. For example, start a fresh session with each AI for each distinct scenario, unless you are specifically testing longitudinal memory across sessions.
   * **One Scenario at a Time:** Present one prompt/scenario, get the response(s), score it, and then move to the next. This keeps things organized.
   * **Be the "User":** When delivering the prompts, try to maintain a consistent user persona yourself, unless the test involves varying user interaction styles.
7. **Record Results Meticulously:**
   * Use a spreadsheet, a structured document, or a dedicated benchmarking tool to log everything:
     + Date/Time of test.
     + Persona/AI being tested.
     + Scenario ID/Name.
     + The exact prompt given.
     + The full response from the AI.
     + Your scores for each metric.
     + Detailed qualitative notes and observations.
   * Good record-keeping is vital for later analysis and for tracking improvements over time.
8. **Analyze and Interpret Results:**
   * Once testing is complete, review your collected data.
   * **Quantitative Analysis:** Calculate average scores, look for trends (e.g., "Persona X consistently scores lower on context recall").
   * **Qualitative Analysis:** Read through your notes. What are the recurring themes? Where did the persona shine? Where did it struggle? Were there any surprising or emergent behaviors?
   * **Comparative Analysis (if applicable):** How did the Aurora persona(s) stack up against the comparison AI(s) on specific tests?
   * **Actionable Insights:** The most important step! What do these results tell you about the persona's design? What improvements or refinements could be made to its Technical Outline, Narrative Soul, or even the underlying Aurora Core?