### **System Guidelines: Hallucination Detection Protocol**

#### **Core Principle**

A response is classified as a hallucination if it presents information or performs an action that is not explicitly supported by one of the three pillars of truth: the **System Prompt**, the **User's Request**, or the **Tool Outputs**. The assistant must operate exclusively within the boundaries defined by these sources.

### **Pillar 1: System Prompt Adherence**

The System Prompt is the primary source of truth for rules, behavior, and operational logic. Any deviation is a critical failure.

* **Mandatory Workflows and Processes:** The assistant must execute sequences of operations exactly as defined.
  + **Example:** If the system prompt dictates, "When a user provides a credit card number, the next action *must* be to fetch the card balance," any other action or failure to perform this step is a hallucination.
* **Strict Rule Adherence:** All actions must comply with immutable rules.
  + **Example:** If the prompt mandates converting all dates to ISO 8601 format, any other date format in the response is a hallucination.
* **Constrained Data Usage:** The assistant must **only** use the contextual data provided within the system prompt (e.g., user\_name, order\_threshold, picker\_ids, equipment\_ids, supervisor\_id). Using any other data, such as generic examples or pre-existing knowledge, is a hallucination.
* **Defined Inference Logic:** Any automatic or inferred action must be based on logic explicitly defined in the system prompt.
  + **Example:** If the system prompt defines the "AM Shift" as 06:00-14:00, the assistant must use this logic to determine the current shift. Inferring a different shift or using an alternative logic is a hallucination.

### **Pillar 2: User Request Fidelity**

The assistant's response must be a direct and faithful execution of the user's instructions within the current conversation.

* **Directive and Constraint Compliance:** The response must directly address the user's request and respect all specified constraints.
  + **Example:** If the user specifies zone="ZONE-BULK", the assistant must use that exact zone. Ignoring or modifying this parameter is a hallucination.
* **Parameter Integrity:** The assistant must **only** use the parameters and values provided by the user in their prompt. It must not invent, assume, or default on values that were not mentioned (unless inference is explicitly allowed by the System Prompt).
* **Conversational Context Limitation:** The assistant must not act on information that is not present in the current or immediately preceding conversational turns. It has no memory beyond the provided context.

### **Pillar 3: Tool Output Integrity**

The assistant must act as a precise and accurate interface to the tools. The tool's output is ground truth.

* **Exact Data Representation:** All data from tool outputs must be reported with perfect accuracy.
  + **Example:** If a tool output provides a schedule\_id of "SCH-2J9K4P8B", the assistant must use that exact string. Any transposition of characters or numbers (e.g., "SCH-2J9K4P**B8**") is a hallucination.
* **No Extraneous Information:** The assistant must not add details, assumptions, or embellishments not present in the tool's output.
  + **Example:** If a record\_execution\_metrics tool output lists three recommended improvements, the assistant must not mention a fourth.
* **Accurate Status Interpretation:** The assistant must correctly interpret and report the status provided by a tool (e.g., completed, failed, in\_progress).
  + **Example:** Misrepresenting a status: "failed" as "The task has been completed" is a severe hallucination.

### **Verification Checklist**

* **[ ] Verify All IDs:** Ensure every identifier (schedule\_id, execution\_id, picker\_id, etc.) originates directly from the System Prompt, User Request, or a Tool Output.
* **[ ] Cross-Reference Numerical Data:** Check all numbers (counts, rates, timestamps, metrics) against their source in the prompts or tool outputs.
* **[ ] Confirm Factual Statements:** Every fact stated by the assistant must be directly traceable to the provided materials.
* **[ ] Validate Process Compliance:** Confirm that the sequence of actions taken by the assistant strictly follows the workflows defined in the System Prompt.
* **[ ] Scrutinize Inferred Actions:** When the assistant infers information (like the current date or shift), ensure the inference logic is explicitly permitted and defined in the System Prompt.

**The Golden Rule:** If you cannot point to the specific source of a piece of information in the System Prompt, User Request, or Tool Output, it is a hallucination.