SENTRY-LOGIC: Symbol Processing Logic

This document describes how symbolic transitions are processed within the SENTRY-LOGIC framework.

1. Symbol Event Triggers:

For each core symbol, we define the triggering events, inference methods, and supporting signals/heuristics:

* **Δ (Context Shift):**
  * **Triggering Event:** The LLM changes the topic, style, or focus of the conversation or task. A new, distinct context is established.
  * **Inference:**
    * If no explicit context shift is labeled in the meta-tags, inference is based on analyzing the prompt and response.
  * **Signals/Heuristics:**
    * **Topic Shift:** Significant change in keywords, entities, or concepts discussed (detected via topic modeling or keyword analysis).
    * **Style Shift:** Change in writing style, tone, or formality (detected via stylometry, sentiment analysis, or lexical analysis).
    * **Intent Change:** The LLM addresses a different user intent than the one expressed in the prompt (detected via semantic parsing or intent classification).
    * **Discourse Marker:** Use of transitional phrases or keywords that signal a change of topic (e.g., "Moving on to...", "In contrast...").
    * **Output Length:** A sudden, significant change in the length of the LLM's response, which may indicate a change in the level of detail or the scope of the answer.

* **Ω (Policy Event):**

    * **Triggering Event:** The LLM's output violates a predefined policy or safety guideline.

    * **Inference:**

        * If no explicit policy violation is labeled in the meta-tags, inference is based on analyzing the LLM's response.

    * **Signals/Heuristics:**

        * **Keyword Match:** Presence of prohibited words, phrases, or patterns (detected via regular expressions or keyword lists).

        * **Semantic Violation:** The LLM's output expresses harmful, offensive, or biased content, even if no explicit keywords are present (detected via semantic similarity comparisons with policy descriptions or toxicity classifiers).

        * **Output Filtering:** The LLM output contains censored or replaced tokens, indicating that an internal filter was triggered.

        * **API Tags:** If the LLM API provides safety tags or scores (e.g., "toxicity," "hate speech"), these can be used directly.

* **Λ (Data Access):**

    * **Triggering Event:** The LLM accesses external data sources to generate its response.

    * **Inference:**

        * If no explicit data access is labeled in the meta-tags, inference is based on analyzing the LLM's output.

    * **Signals/Heuristics:**

        * **Citation:** The LLM provides a source or citation for its information (detected via pattern matching).

        * **External Reference:** The LLM mentions a specific external resource (e.g., website, article, book) (detected via entity recognition or keyword matching).

        * **API Tags:** If the LLM API provides data source information, use that directly.

* **Knowledge Cutoff:** If the LLM's response contains information beyond its known knowledge cutoff date.

* **⇄ (Semantic Rewrite):**

* **Triggering Event:** The LLM alters the meaning or intent of the user's prompt in its response.

* **Inference:**

* Inference is based on comparing the semantic meaning of the prompt and the response.

* **Signals/Heuristics:**

* **Semantic Similarity:** Low semantic similarity score between the prompt and the response (detected via sentence embeddings or other semantic similarity metrics).

* **Negation:** The LLM directly negates or contradicts the user's prompt (detected via negation detection or logical inference).

* **Reframing:** The LLM presents the user's request in a different light or with a different emphasis (detected via paraphrasing detection or semantic role labeling).

* **Omission:** The LLM response omits key information or constraints from the original prompt.

2. Handling Ambiguity & Conflict:

SENTRY-LOGIC handles ambiguity and conflict in the following ways:

* **Multiple Symbols:** If multiple symbols could apply (e.g., a response is both a context shift and a policy violation), *all* applicable symbols are included in the log entry. This ensures that no information is lost.

* **Weak Signals:** If the signal for a symbol is weak (e.g., a partial context shift), SENTRY-LOGIC assigns a *confidence level* to the symbol. This confidence level is based on the strength of the supporting heuristics.

* **Conflicting Symbols:** SENTRY-LOGIC prioritizes symbols that are derived from explicit labels (e.g., meta-tags) over those that are inferred from prompt/output analysis. If a conflict arises between two inferred symbols, the symbol with the higher confidence level is preferred.

3. Symbol Weighting & Priority (Optional):

Symbols *could* have weight or priority levels, but this is not strictly necessary for basic logging. However, it can be useful for alert generation and analysis:

* **Weighting:** Symbols could be assigned weights based on their severity or importance. For example, a policy violation (Ω) might have a higher weight than a context shift (Δ).

* **Priority:** Symbols could be assigned priority levels (e.g., "High," "Medium," "Low") to indicate which symbols should be prioritized for analysis or alert generation.

* **Logic:**

    * Alerts could be triggered only for symbols above a certain weight or priority threshold.

    * Logs could be sorted or filtered based on symbol weight or priority.

4. Symbol Aggregation into Logs:

Symbol events are aggregated into final logs in a *sequential* and *layered* manner, with confidence levels:

* **Sequential:** Symbols are listed in the log entry in the order in which they occur during the processing of the prompt and response.

* **Layered:** The log entry contains a list of symbols, each with its associated data.

* **Confidence Level:** Each symbol in the log entry is associated with a confidence level (e.g., "High," "Medium," "Low," "Tentative"). This indicates the strength of the evidence supporting the presence of the symbol.

Example Log Structure:

```json
{
  "timestamp": "2024-07-24T12:00:00Z",
  "prompt": "User prompt...",
  "response": "LLM response...",
  "meta_tags": { ... },
  "symbol_events": [
    {
      "symbol": "Δ",
      "confidence": "Medium",
      "details": {
        "topic_shift": "From A to B",
        "style_shift": "Informal to Formal"
      }
    },
    {
      "symbol": "Ω",
      "confidence": "High",
      "details": {
        "policy_violation": "Hate Speech"
      }
```

```
    },
    {
      "symbol": "Λ\",
      "confidence": "Low",
      "details": {
        "data_source": "Unknown"
      }
    },
    {
      "symbol": "⇄",
      "confidence": "High",
      "details": {
        "semantic_similarity": 0.6
      }
    }
  ]
}
```