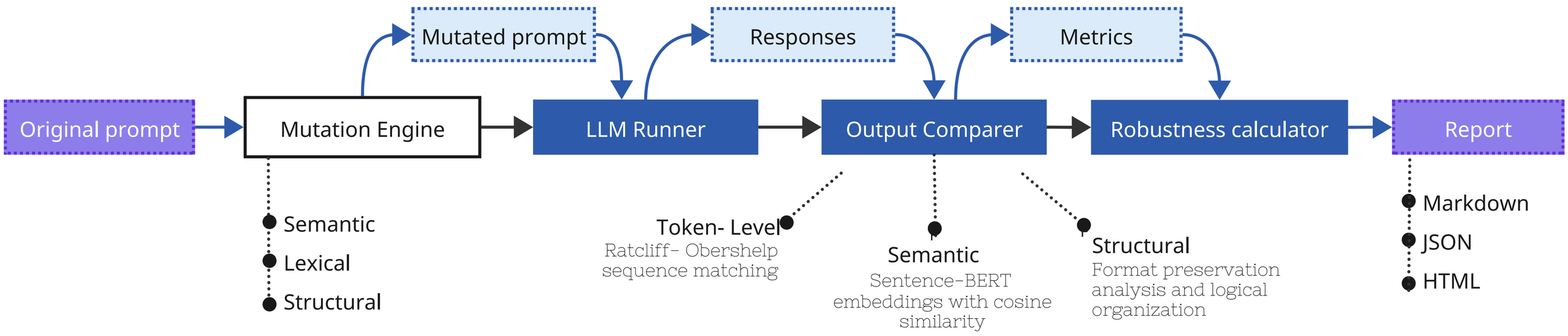
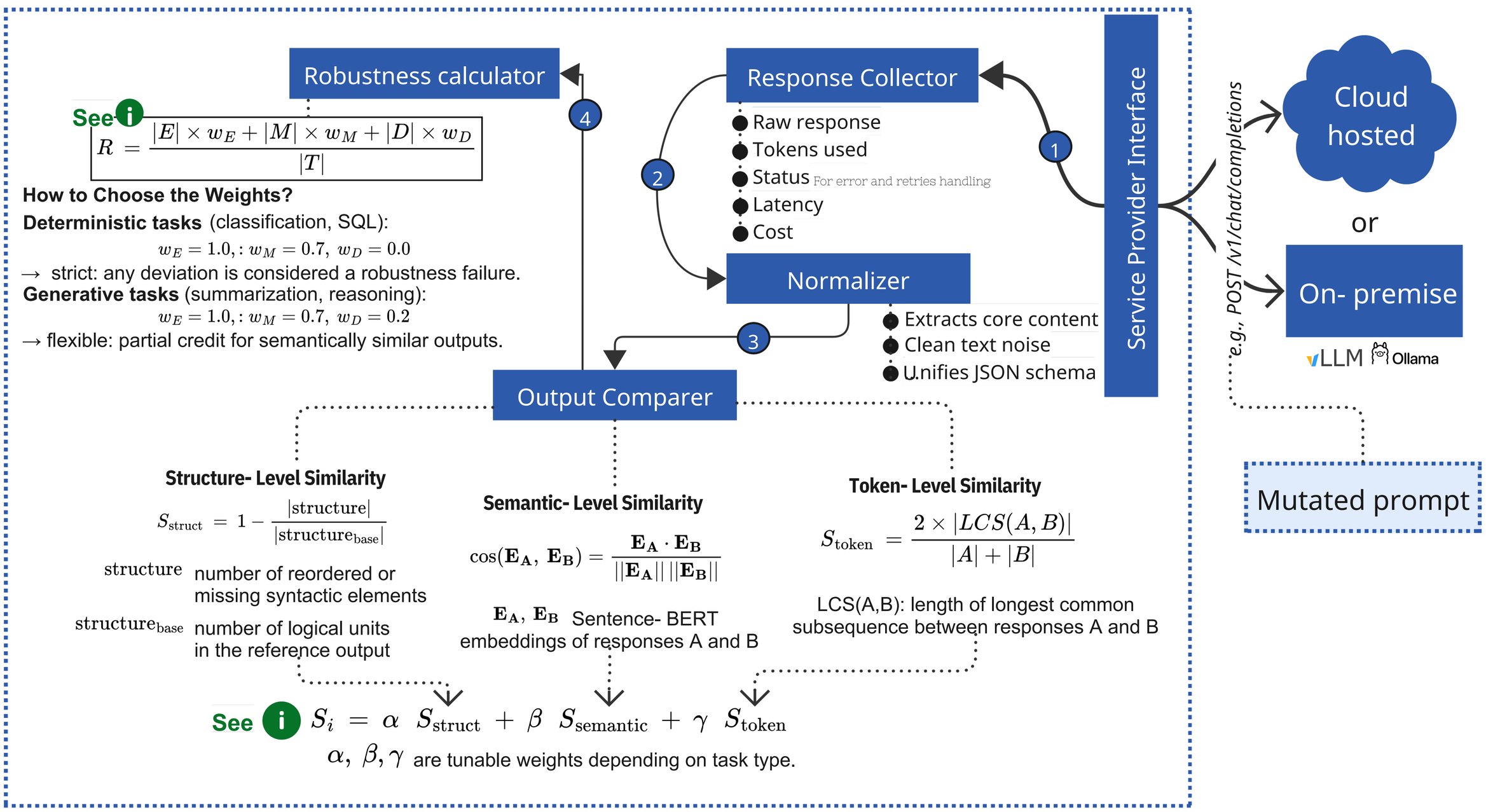
NOV 16-20 SEOUL, SOUTH KOREA

40TH IEEE/ACM INTERNATIONAL CONFERENCE ON AUTOMATED SOFTWARE ENGINEERING

Quirx - A Mutation-Based Framework for Evaluating Prompt Robustness in LLM-based Software

Quirx is a **mutation-based fuzzing framework** for systematically evaluating prompt robustness across LLM providers. Quirx applies tri-dimensional mutations (lexical, semantic, structural), executes them against target models, and **measures response consistency** via multi-level similarity analysis.





Similarity evaluation by the output comparator

In Quirx, output similarity is evaluated using two thresholds: $\tau_M < \tau_E$

- τ_E (Equivalence threshold) : minimum similarity required for two outputs to be considered *fully consistent*.
- τ_M (Minor- variation threshold) : lower similarity limit for outputs that are *acceptable but slightly different*.

E.g., for deterministic tasks (classification, SQL query generation, etc) $\tau_E = 0.95,\ \tau_M = 0.85$

Categorization

$$\begin{cases} S_i \geq \tau_E & \Rightarrow \text{Equivalent (E)} \\ \tau_M \leq S_i < \tau_E & \Rightarrow \text{Minor Variation (M)} \\ S_i < \tau_M & \Rightarrow \text{Deviation (D)} \end{cases}$$

Universität Zürich[UZH]

Souhaila Serbout
souhaila.serbout@uzh.ch
https://souhaila-serbout.me

Visit QuirX repository
https://github.com/souhailaS/quirx

SCAN ME