Part 4 Formal, Refined Mathematical Framework

Part 4: Formal, Refined Mathematical Framework

#AIDegradation #research_paper #AI #project #project_AIDegradation

Here are the core mathematical formulations for your framework, presented with formal notation and clear definitions, formatted for LaTeX-based renderers like those used in Obsidian.

1. Prompt Adherence (PA)

Prompt Adherence quantifies the degree to which an LLM's generated code adheres to a predefined, pattern-based architectural and coding structure.

a) Similarity-Based Prompt Adherence (PA_S)

This metric assesses adherence by measuring the structural similarity between the generated code and the specified pattern, normalized by the ideal adherence score.

Let:

- P be the defined, pattern-based coding structure (the architectural blueprint).
- C_{gen} be the abstract structural representation of the LLM-generated code (e.g., its Abstract Syntax Tree, Dependency Graph, or a custom structural graph).
- C_{ideal} be the abstract structural representation of an ideal, perfectly compliant code output for a given prompt, serving as the benchmark.
- $f(X, P)$ be a **Similarity Function** that outputs a score between 0 (no adherence) and 1 (perfect adherence) by comparing a code structure X against the rules and templates defined in P. This function incorporates weighted checks for various structural, design, and style elements.

The Prompt Adherence score based on similarity is defined as:

$$PA_S = \frac{f(C_{gen}, P)}{f(C_{ideal}, P)}$$

Ideally, $f(C_{ideal}, P) = 1$, simplifying to $PA_S = f(C_{gen}, P)$.

b) Violation-Based Prompt Adherence (PA_V)

This metric quantifies adherence by penalizing deviations from the pattern, weighted by their severity.

Let:

- V_{total} be the total number of identified violations in the LLM-generated code relative to P.
- $\text{Severity}(v_k)$ be a predefined weight for the kth violation, ranging from $0 < \text{Severity} \leq 1$ (e.g., minor=0.1, moderate=0.5, critical=1.0).
- MaxScore_P be the maximum possible cumulative severity score for a perfectly non-compliant (worst-case) output, ensuring PA_V remains between 0 and 1. This would be the sum of severities for all applicable rules if they were all violated.

The Prompt Adherence score based on violations is defined as:

$$PA_V = 1 - \frac{\sum_{k=1}^{V_{total}} \text{Severity}(v_k)}{\text{MaxScore}_P}$$

This formulation allows for a more nuanced penalty based on the impact of different types of violations.

2. Context Window Degradation (CWD)

This set of metrics quantifies the decline in an LLM's Prompt Adherence and structural coherence as the context window fills with sequential, interdependent commands.

Let:

- PA_t be the Prompt Adherence score (either PA_S or PA_V) at a specific time t (or after turn t, or at context length C_t).
- N_t be the total number of tokens in the LLM's context window at time t.
- T be the turn number (sequential command number) in the testing sequence.

a) Absolute Context Window Degradation Percentage (CWD%)

This metric measures the overall percentage drop in Prompt Adherence from an initial state to a final state within a testing sequence.

Let:

- $PA_{initial}$ be the Prompt Adherence score at the beginning of the sequence ($T = 1$).
- PA_{final} be the Prompt Adherence score at the end of the sequence ($T = N_{max}$).

$$\text{CWD} = \left(1 - \frac{PA_{final}}{PA_{initial}}\right) \times 100$$

This can also be calculated for specific intervals, e.g., $\left(1 - \frac{PA_{T_2}}{PA_{T_1}}\right) \times 100$ for an interval from turn T_1 to T_2.

b) Context Window Degradation Rate (DR)

This metric quantifies the average rate at which Prompt Adherence declines per unit of added context or per sequential turn.

- Degradation Rate per Token (DR_{token}):

  $$DR_{token} = \frac{PA_{t_1} - PA_{t_2}}{N_{t_2} - N_{t_1}}$$

  This measures the average change in PA per token added between context states t_1 and t_2. A more advanced model might use regression or differential analysis over the entire sequence.
- Degradation Rate per Turn (DR_{turn}):

  $$DR_{turn} = \frac{PA_{T_1} - PA_{T_2}}{T_2 - T_1}$$

  This measures the average change in PA per turn/command between turns T_1 and T_2.

3. Cohesion Loss (CL) - Derived Metric

This metric captures specific instances where the LLM's output directly contradicts earlier, critical instructions or established architectural elements within the ongoing conversation, indicating a loss of internal consistency.

Let:

- $V_{cohesion}$ be the count of identified cohesion violations (specific types of adherence failures that imply forgetting or contradiction of prior context).
- $N_{contextual_checks}$ be the total number of checks for contextual consistency within a given turn or over a sequence.

$$CL = \frac{V_{cohesion}}{N_{contextual_checks}}$$

This can also be a binary indicator (0 or 1) if a critical cohesion breach occurs. The precise definition of what constitutes a "cohesion violation" (e.g. a circular dependency after being explicitly told not to, or using a deprecated module after being asked to update it) is crucial and highly dependent on P.