Structure Leap Evaluation Protocol for Conceptual Language Systems

CLP Proto Team
clpproto@gmail.com

Contents

Overview

This protocol defines the methodology for evaluating structure-based semantic transformations in the Conceptual Language Protocol (CLP) system. It introduces a multi-model, cross-verification framework, designed to replace single-model hallucination with a measurable, structured assessment layer.

Purpose

Traditional LLM outputs are difficult to evaluate objectively, especially in tasks involving deep semantic transformation. CLP introduces the notion of semantic leaps, requiring a transparent and reproducible protocol for validating:

- Conceptual alignment
- Structure coherence
- Nonlinear meaning preservation
- Inter-model verification agreement

Evaluation Layers

CLP Evaluation is divided into three main layers:

Layer	Name	Function
L1	Structural Leap Output	CLP system generates a target output through structure-guided semantic leap
L2	Model Scoring Interface	External LLMs (GPT-4, Claude, Gemini, Grok) rate the transformation
L3	Aggregated Judgement	Human + model synthesis of scores, identifying patterns or divergence

Table 1: Evaluation Layers

Evaluation Dimensions

Each model evaluates the output using the following four axes, each scored from 0.0 – 3.0:

Fidelity Does the output preserve the core meaning of the source input?

Fluency Is the target language output coherent and natural?

Structural Coherence Is the conceptual leap internally logical and well-formed?

Leap Validity Does the transformation reveal a true semantic shift (not a paraphrase)?

Total Score Range per Model: 0.0 – 12.0
Scores below 6.0 indicate semantic failure or incoherence.

Multi-Model Feedback Grid

Each output is evaluated by at least 3 independent LLMs, in isolated prompts. A sample scoring table:

Output #	GPT-4	Claude	Gemini	Average	Notes
001	9.5	8.7	9.1	9.1	Stable leap across models
002	6.2	5.8	6.5	6.2	Slight conceptual distortion
003	4.0	3.5	4.8	4.1	Leap failure – mismatch in logic

Table 2: Sample Multi-Model Feedback Grid

Verification Protocol

1. **Input Selection**: Choose source sentence with high conceptual density.
2. **CLP Transformation**: Run structure-based translation in controlled sandbox.
3. **Model Isolation**: Feed only the output + original input to each model in clean prompt.
4. **Score Extraction**: Collect structured scoring with justification.
5. **Documentation**: Archive source, target, scores, and screenshots.

All steps are repeatable and externally inspectable.

Sample Prompt (LLM Evaluation)

```
[Instruction]
You are evaluating a conceptual translation system. Rate the output along 4 axes

[Input]
Original: "He walked into the ruins, searching for the memory of a name."
Translated: "In the hollow bones of the city, he hunted echoes of forgotten iden

[Evaluation Template]
Fidelity: ___ – Justification: ...
Fluency: ___ – Justification: ...
Structure: ___ – Justification: ...
Leap Validity: ___ – Justification: ...
Total: ___ / 12.0
```

Aggregated Visualization

Optionally, aggregated visualizations such as heatmaps or line charts can be used to illustrate:

- Dimension strengths per model

- Inter-model agreement or divergence
- Stability and reliability of structural leap pathways

These visual tools help identify latent patterns in evaluation responses and support qualitative comparison across systems.

Below is a line chart showing the evaluation scores for Output #001 across three models (GPT-4, Claude, Gemini) along the four dimensions (Fidelity, Fluency, Structural Coherence, Leap Validity). The chart illustrates dimension strengths and inter-model agreement (see Figure 1).

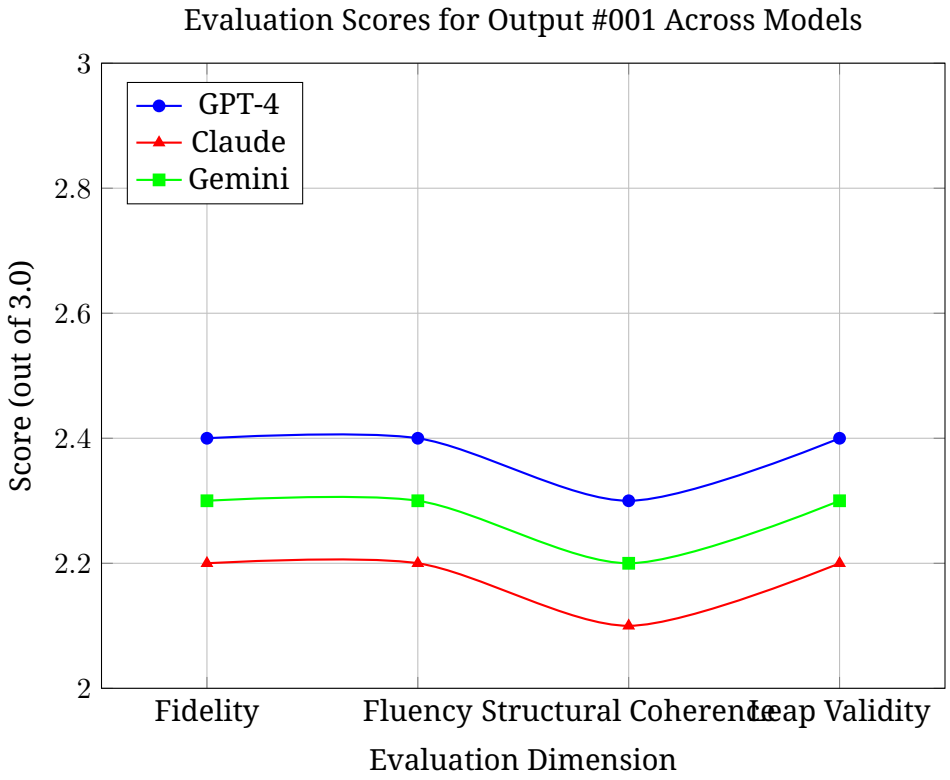

Figure 1: Line chart showing evaluation scores for Output #001 across models.

(Example visualizations can be appended to full evaluation logs or viewed on the CLP prototype website at https://clp-proto.github.io/clp-site.)

Ethical Considerations

- The evaluation does not train any model
- All model access is via read-only API prompts
- Scores are used for structure validation, not ranking models
- Human review is recommended for high-stakes applications

Version Control & Reproducibility

Each evaluation session should be:

- Timestamped
- Associated with CLP engine version (CLP_proto_v0.2 etc.)
- Logged with all inputs/outputs/screenshots for audit

Future Extensions

- Crowdsourced scoring layer (human-moderated)
- Integration into continuous evaluation pipelines
- Cross-language semantic leap benchmarking
- Use in academic peer review for structure-based NLP

License

This protocol is released under CC BY-NC-SA 4.0
Authors: CLP Proto Team
Contact: clpproto@gmail.com