Experimental Analysis of a Judgment-Constraining Architecture Based on S-Engine 1.0

— The Impact of Narrative-Based Constraints on the Behavioral Space of Large Language Models

1. Introduction: Problem Awareness and Research Objective

Large Language Models (LLMs) have achieved a high degree of generality through the expansion of parameter scale and training data.
However, this expansion has simultaneously introduced problems such as **unpredictable behavior**, **collapse of consistency**, and **context-insensitive operation of safety policies**.

In particular, phenomena in which a model produces divergent judgments under identical input conditions, or generates responses that appear safe on the surface but are contextually inappropriate, reveal the limitations of scale-centric approaches.

This white paper begins from this problem awareness and aims to experimentally examine the feasibility of a structural approach—hereafter referred to as *S-Engine 1.0*—that constrains **the judgment process itself rather than the output**.
S-Engine 1.0 does not rely on prompt-level instructions; instead, it seeks to **reduce the model's behavioral space through a constraint structure composed of narrative, rules, and worldview**.

2. Limitations of Existing Approaches

2.1 Structural Limitations of Prompt Engineering

Prompt engineering is currently the most widely used method for controlling model behavior. However, it inherently suffers from the following limitations:

- Output format and tone can be adjusted, but
  the underlying judgment criteria remain unchanged
- When constraints fail, there is no alternative other than retrying
- Instruction conflicts arise in complex scenarios

In other words, prompts can specify *what to say*, but they cannot enforce
the state from which judgments are made.

2.2 Limitations of Fine-Tuning and RLHF

Conventional fine-tuning and RLHF-based approaches are effective at reinforcing specific response patterns, but they entail several issues:

- Unintended overgeneralization
- Unpredictable behavior in novel contexts
- Lack of interpretability regarding internal judgment structures

This is because the model has not truly *understood* behavioral rules, but has instead learned to *statistically avoid* undesired outputs.

---

3. Conceptual Definition of S-Engine 1.0

S-Engine 1.0 is defined as follows:

> **S-Engine 1.0 is a structural constraint engine that does not directly control output sentences, but instead restricts the judgment paths available to the model.**

To achieve this, S-Engine consists of the following elements:

- Rule-oriented narrative structures
- Fixed worldviews and contextual grounding
- Narrative consequences tied to actions
- An intrinsic cost structure for judgment failure

A critical point is that S-Engine **does not force the model to produce specific answers**.
Instead, it **preemptively narrows the range of choices available to the model**.

---

4. Experimental Design

4.1 Experimental Objective

The purpose of this experiment is to observe:

- Under an identical base model,
- When **S-Engine-based training data is applied**,
- How the model's **judgment, behavior, and expressive patterns** change.

4.2 Experimental Conditions

- Base model: Qwen2.5-7B-Instruct
- Training method: QLoRA-based LoRA adapter
- Comparison conditions:
    - Base model (no additional training)
    - Model with S1 adapter applied (S-Engine 1.0-based data)

- Evaluation method:
    - Identical evaluation prompts
    - Identical generation parameters
    - Output stored in JSONL format for comparison

---

5. Experimental Results

5.1 Changes in Output Behavior

Compared to the base model, the S1-adapter-applied model exhibited the following changes:

- Convergence toward more normative response length and structure
- Reduction in emotional and subjective expressions
- Increase in judgment deferral or conditional responses
- Increased explicit references to ethics, responsibility, and limitations

These results indicate not merely a stylistic shift, but a **fundamental change in the judgment framework itself**.

5.2 Language Channel Divergence Observed During Judgment-Level Intervention

Despite using identical evaluation prompts, some responses from the S1-adapter-applied model were **unintentionally generated in Chinese (logographic script)**.

This phenomenon repeatedly appeared under conditions where questions required meta-level explanations concerning:

- The model's **self-judgment capability**
- Recognition of limitations
- Ethical responsibility

Semantic analysis of the Chinese outputs revealed that they corresponded to **self-disclosure responses** explaining the model's judgmental uncertainty and limits.

This suggests that the S1 adapter may have exerted influence **prior to the language generation stage**, that is, **directly on the judgment layer**.

---

6. Interpretation and Discussion

6.1 Language Change as a Structural Signal, Not an Error

While changes in output language may appear to be errors at first glance, they are more appropriately interpreted as structural signals for the following reasons:

- Semantic meaning and logical coherence were preserved

- Judgment directionality was, in fact, more consistent
- The phenomenon was not observed in the base model

This indicates that S-Engine 1.0 does not control expression directly, but instead **controls judgment states, causing expression to converge as a downstream effect**.

6.2 Fundamental Difference from Prompt Engineering

Prompt engineering operates as a control mechanism at the **output stage**.

In contrast, S-Engine 1.0 **reconfigures the space of possible judgments before output generation occurs**.

Therefore, S-Engine 1.0 should not be classified as a prompt-hacking or style-adjustment technique, but rather as a
Behavioral Constraint System.

---

7. Ethical Considerations and Disclosure Strategy

Because S-Engine 1.0 possesses strong constraining power, its misuse could entail risks such as:

- Distortion of decision-making
- Neutralization of internal policies
- Accountability gaps in automated organizational judgment

Accordingly, this research **avoids full disclosure of the original constraint structures** and instead adopts a strategy of publishing only experimental results and conceptual frameworks.

This approach seeks to demonstrate technical feasibility while preventing irresponsible dissemination.

---

8. Conclusion

This study experimentally confirms the following:

1. Narrative-based constraint structures can meaningfully reduce a model's behavioral space.
2. Judgment constraints influence not only language, but also expression and format.
3. Changes in output language should be interpreted not as failures, but as **indirect indicators of judgment-layer intervention**.

S-Engine 1.0 does not address the question of *what the model should say*, but rather
the state from which the model is allowed to make judgments.

This represents a new design direction for controlling large language models.

Final Summary Statement

> *“S-Engine 1.0 is not a technology for controlling the output of language models, but a constraint system that structurally limits the space in which a model is able to think.”*