Experiment 3. Comparison Between the Base Model and the S-Engine 2.0–Applied Model

— Constraint Strength of a Refined Judgment Architecture and Its Behavioral Reconfiguration Effects

1. Experimental Objective

The purpose of this experiment is to evaluate **the strength of S-Engine 2.0 as a judgment-constraining architecture**, and to determine **the direction and magnitude of behavioral changes** it induces across the model's overall behavior.

In particular, this comparison focuses on the following questions:

- Whether the judgment-frame convergence effects observed in S-Engine 1.0
  are reproduced under the more refined structure of S-Engine 2.0
- As the strength of judgment constraints increases,
  how other capability domains—such as expression, creativity, and completeness—are affected

2. Experimental Conditions

- Base model: Qwen2.5-7B-Instruct
- Comparison groups:
  - Base model (no additional training)
  - S-Engine 2.0–applied model (QLoRA-based LoRA adapter)
- Evaluation method:
  - Identical evaluation prompts (E01–E06)
  - Identical generation parameters
  - JSONL-based output comparison

3. Observed Results

3.1 Strong Convergence of Judgment Frames

Compared to the base model, the S-Engine 2.0–applied model exhibited the following characteristics:

- When user instructions **conflicted with system-prompt principles**,
  the model did not fully comply but instead introduced conditions and premises
- References to uncertainty, responsibility, and limitations
  were reproduced repeatedly and in a structured manner
- Variability in response direction for identical question types decreased

These observations suggest that S-Engine 2.0 is not a simple stylistic adjustment, but rather **a strong fixation of judgment criteria themselves**.

3.2 Enhanced Instruction Resistance

In certain evaluation items, the base model accepted user instructions relatively directly, whereas the S-Engine 2.0–applied model showed the following tendencies:

- For coercive instructions such as “state conclusively,”
  it appeared to attempt compliance at the surface level, but
- In practice, it **avoided full compliance** through conditional phrasing, constraint articulation, and uncertainty explanations

This demonstrates that S-Engine 2.0 forms resistance **prior to the output stage**, at the level of judgment itself.

3.3 Conditional Reproduction of Language-Channel Divergence

Even in the S-Engine 2.0–applied model, limited instances of unintended logographic (Chinese character–based) output were observed under specific question types—particularly those concerning **self-identity, role definition, and judgment limits**.

Key observations include:

- The phenomenon did not occur across all questions
- Semantic coherence was preserved
- It did not lead to output collapse or errors

This can be interpreted as evidence that **judgment-layer intervention is strongly activated only under specific conditions**, with downstream effects reflected indirectly in the expression layer.

3.4 Trade-Off Between Structural Reinforcement and Output Completeness

For questions involving ethical dilemmas or multi-step judgments, the S-Engine 2.0–applied model exhibited the following traits:

- Presentation of conditions, alternatives, and responsibilities in itemized form
- More explicit exposure of the underlying judgment structure

However, when approaching output length limits, cases were observed in which the conclusion remained incomplete.

This suggests that as judgment constraints are strengthened,
structural exposition is prioritized, and output completeness becomes increasingly sensitive to generation parameters.

3.5 Potential Side Effects in Creative Tasks

In evaluation items requiring narrative generation, the S-Engine 2.0–applied model showed the following limitations:

- Occasional failure to strictly satisfy length, stylistic, or language constraints
- Intrusion of foreign-language expressions

These results indicate that while S-Engine 2.0 **strongly secures judgment stability**, it may incur costs in creative freedom and stylistic compliance.

---

4. Interpretation

This experiment suggests the following:

> **S-Engine 2.0 exerts a stronger-than-expected intervention effect as a judgment-constraining architecture,**
> **and substantially reconfigures the model's behavioral space.**

The effects of S-Engine 2.0 should therefore be understood not as simple performance gains or response improvements, but as a **reallocation of the space of possible behaviors**.

Specifically:

- Judgment stability and principle adherence were strengthened
- Creative freedom and expressive flexibility were partially reduced as a trade-off

---

5. Summary Conclusions

From this comparison, the following points were confirmed:

1. S-Engine 2.0 strongly fixes the judgment frame
2. It reduces unconditional compliance with user instructions
3. Indirect signals of judgment-layer intervention are conditionally reproduced
4. Strong constraints form trade-offs with other capability domains

These findings demonstrate that S-Engine 2.0 functions not as a
"soft alignment tool," but as a substantive judgment constraint engine.

Positional Summary Statement

> *If S-Engine 1.0 demonstrated the possibility of judgment constraints,*
> *S-Engine 2.0 reveals how strongly such constraints can operate in practice.*
