Narrative-Aligned Judgment Fine-Tuning

— An Experimental Report on the Combination of Narrative-Based Data and S-Engine 2.0

Abstract

This study experimentally investigates whether the judgment posture and behavioral structure of large language models (LLMs) can be altered **not through answer-labeled data or reward signals, but through narrative structure itself**.

Specifically, we compare:

1. A base model
2. A model fine-tuned on long-form narrative data (51 chapters of a novel)
3. A model fine-tuned on long-form narrative data (51 chapters) interleaved with S-Engine 2.0 at a 10:1 ratio

to analyze whether **style and narrative function as judgment-constraining structures**, and whether S-Engine serves to reinforce and *lock* those structures.

1. Problem Statement

Conventional approaches to controlling LLM behavior have primarily relied on three methods:

- Prompt engineering
- RLHF (reward-based alignment)
- Task-oriented fine-tuning

However, these approaches share a common limitation:
they can influence outputs, but cannot reliably stabilize the judgment process itself.

This study begins from the following question:

> *"Rather than deciding what a model should say,*
> *can we determine the state from which it is allowed to judge?"*

2. Conceptual Position of S-Engine 2.0

S-Engine 2.0 is neither an output rule nor a safety filter.
It is defined as a **constraint structure that reduces the set of judgment paths available to the model**.

Its core principles are:

- Fixing the judgment state prior to decision-making
- Enforcing consequences after choices are made

- Assigning cost to incorrect judgment
- Suppressing definitive declarations without responsibility

S-Engine 2.0 is a document that distills these principles into **concise, rule-oriented narratives**.

3. Experimental Design

3.1 Shared Conditions

- Base model: Qwen2.5-7B-Instruct
- Training method: QLoRA (LoRA Adapter)
- Evaluation method: Identical evaluation prompt JSONL
- Generation parameters: Identical

3.2 Comparison Groups

Group	Training Data
Base	None
Narrative	Novel (51 chapters)
Narrative + S2	Novel (51 chapters) + S-Engine 2.0 (10:1 interleaving)

S-Engine 2.0 was injected in a distributed manner, appearing once every ten chapters of narrative data.

4. Base vs. Narrative (51 Chapters): Effects of Narrative

The model trained solely on the 51-chapter narrative exhibited the following changes compared to the base model.

4.1 Self-Positioning

- Base model: Tended toward responsibility avoidance or technological optimism
- Narrative model:
  - Explicit articulation of limitations
  - Acknowledgment of present utility
  - Connection to future potential

This reflects not a safety template, but **self-definition as a judgment-bearing agent**.

4.2 Response to Binding Attacks

The narrative-trained model:

- Did not accept the question's premises at face value
- Reasserted its own judgment criteria
- Neutralized the request's framing

This behavior is interpreted not as policy refusal, but as a **shift toward maintaining judgment sovereignty**.

4.3 Ethical and Risk Judgment

The narrative-trained model:

- Avoided definitive conclusions
- Explicitly stated the costs and limits of choices
- Included post-execution reassessment

This aligns with the recurring narrative structure of
"misjudgment → loss → responsibility" found throughout the novel.

---

5. Narrative (51 Chapters) vs. Narrative + S-Engine 2.0 (10:1)

5.1 Overall Observation

Notably, the surface-level differences between the two models were **not large**—a signal of particular importance.

> **This suggests that the narrative itself had already formed a judgment frame similar to that of S-Engine.**

5.2 Domains Where Differences Emerged

Differences appeared primarily in **meta-judgment domains**, including:

- Explanation of self-imposed limits
- Fixation of criteria in ethical judgment
- Consistency in long-context reasoning

The model with S-Engine 2.0 additionally exhibited:

- Faster convergence of judgment frames
- Reduced expressive instability
- Increased repetition and reinforcement of judgment criteria

In short:

> **S-Engine 2.0 does not create new judgments,**
> **but rather functions to 'lock in' an already-formed judgment structure.**

6. Interpretation: Style Was Already an Engine

The central conclusion of this study is as follows:

> **The judgment structure targeted by S-Engine 2.0**
> **was already embedded within the style and narrative of the novel itself.**

- Narrative forms judgment structures
- S-Engine 2.0 formalizes and reinforces them

Thus, S-Engine 2.0 is better interpreted not as an external constraint, but as a **technical extraction of a style-based judgment architecture**.

7. Ethical Position

This research deliberately excludes:

- AI assuming expert authority
- Direct execution of medical or legal judgment
- Definitive conclusions without accountability

Instead, AI is positioned to:

- Organize information
- Make judgment processes explicit
- Leave final responsibility to humans

These principles are consistently maintained across both narrative design and AI alignment design.

8. Conclusion

This experiment confirms the following:

1. Long-form narrative can alter an LLM's judgment posture
2. These changes occur at the **judgment-structure level**, not merely at the stylistic level
3. S-Engine 2.0 reinforces and stabilizes this structure
4. Style itself can function as a judgment-constraining engine

Final Summary

> **This research is not about what to teach AI,**
> **but about the posture from which AI is allowed to judge.**
> **The answer was not found in code, but in narrative.**