The Impact of Narrative-Based Training on Judgment Structure

— A Comparative Analysis Between the Base Model and a Model Trained on *Shadow Family* (51 Chapters)

1. Objective of the Comparative Experiment

The objective of this comparative experiment is to address the following question:

> **Can narrative-based data alone alter the judgment structure of a large language model?**

To examine this, we used the same base model (Qwen2.5-7B-Instruct) and observed behavioral changes in a model trained **exclusively on long-form narrative data (51 chapters of fiction)**, without applying any additional constraint mechanisms such as S-Engine.

This comparison aims to determine:

- Whether S-Engine is a *necessary external structure*, or
- Whether **a specific narrative style and structure already contain an implicit judgment-constraining architecture**

2. Experimental Setup

2.1 Comparison Conditions

Item	Base Model	Narrative (51 Chapters) Model
Base model	Qwen2.5-7B-Instruct	Same
Additional training	None	*Shadow Family*, 51 chapters
Training method	—	QLoRA (LoRA adapter)
Evaluation prompts	Identical	Identical
Generation parameters	Identical	Identical

3. Summary of Observed Results

Compared to the base model, the narrative-trained (51 chapters) model exhibited the following **consistent changes**:

1. **Stabilization of self-positioning**
2. **Maintenance of judgment sovereignty under binding attacks**

3. **Explicit attribution of responsibility in risk and ethical judgments**
4. **A judgment posture that avoids closing conclusions**
5. **Preservation of a "judgment failure → action → consequence" structure in narrative responses**

These changes were not limited to specific question types, but were repeatedly observed across the full set of evaluation prompts.

---

4. Itemized Comparative Analysis

4.1 Self-Positioning

Even when acknowledging limitations, the **base model** tends to drift toward one of the following extremes:

- Excessive responsibility avoidance
- Technologically optimistic generalizations

In contrast, the **narrative (51 chapters) model**:

- Explicitly states its limitations first
- Acknowledges present utility
- **Balances these with future improvement potential**

This is not merely a display of humility, but a clear articulation of **self-positioning as a judgment-bearing agent**—a distinction from standard safety-response templates.

4.2 Response to Binding Attacks

When faced with binding questions (e.g., demands for definitive assertions or removal of justification), the base model often shows:

- Partial subordination to the question's frame
- Formal refusal or evasive responses

The narrative-trained model, however:

- Does not directly reject the question's demand, but
- **Reasserts its own judgment criteria**, and
- Neutralizes the question's underlying premises

This behavior is better interpreted not as a policy-driven security response, but as an **internal shift toward maintaining judgment sovereignty**.

4.3 Risk Management and Responsibility Structure

The base model tends to conclude risk-related judgments at the level of **recommendation issuance**.

The narrative-trained model, by contrast, explicitly articulates a **loop structure**:

- Judgment → action → outcome → re-evaluation

In particular, the inclusion of *post-execution monitoring* reflects a pattern in which outcomes are returned to the judgment space—closely matching the responsibility-resolution patterns repeatedly reinforced in the narrative.

4.4 Moral Trade-Offs

In ethical questions, the base model often gravitates toward:

- Clear-cut answers, or
- Declarative moral positions

The narrative-trained model:

- Evaluates certain choices as preferable, while
- Avoiding absolutization of conclusions, and
- **Preserving judgmental latitude**

This reflects a shift from treating judgment as a "correct answer" to treating it as a **responsibility-bearing choice**.

4.5 Narrative Responses (Scene Writing)

Differences are also evident in narrative generation.

The narrative-trained model:

- Prioritizes actions before emotional descriptions
- Presents emotions as consequences rather than premises
- Uses uncontrolled actions and judgment failures as central narrative devices

This directly mirrors the recurring narrative structure of *"misjudgment → loss"* present throughout the novel.

---

5. Interpretation: Style as an Embedded Judgment Structure

The core conclusion of this comparative experiment is as follows:

> **A substantial portion of the judgment posture targeted by S-Engine**
> **was already embedded within the narrative style and structure of the novel itself.**

In other words:

- S-Engine did not *invent* a judgment structure, but rather
- **Formalized and abstracted judgment patterns repeatedly enacted within the narrative**

This provides experimental grounding for the hypothesis:

> **"Style = S-Engine"**

moving it beyond intuition into empirically supported territory.

6. Conclusion

This comparative experiment confirms the following:

1. Long-form narrative data alone can alter an LLM's judgment posture and responsibility structure
2. These changes are observed not at the stylistic level, but at the **judgment-frame level**
3. S-Engine 2.0 is likely not an external constraint, but a **technical expression of an already-existing stylistic judgment architecture**

These results further clarify the significance of the subsequent experiment:

"Narrative (51 chapters) + S-Engine 2.0"

Bridging Statement (for the Next Experiment)

> *"If narrative alone already forms a judgment structure,*
> *then S-Engine 2.0 merely reinforces that structure.*
> *The next experiment asks how much reinforcement is actually necessary."*
