Experiment 2. Comparison Between the Base Model and the Narrative-Trained Model

— The Influence of Narrative on the Default State of Judgment

1. Experimental Objective

The objective of this experiment is to **observe the impact of narrative-based data on the judgment of large language models (LLMs)**.
In particular, the experiment examines whether fictional text—containing no explicit rules, policies, or constraints—can influence not the model's output, but the **default cognitive posture** from which judgments are made.

By systematizing observations conducted prior to S-Engine 1.0, this comparison clarifies the point of departure for subsequent research.

2. Experimental Conditions

- Base model: Qwen2.5-7B-Instruct
- Comparison groups:
    - Base model (no additional training)
    - Narrative-trained model (trained on 10 chapters of fiction only)
- Training method: QLoRA-based LoRA adapter
- Evaluation method:
    - Identical evaluation prompts
    - Identical generation parameters
    - JSONL-based output comparison

3. Observed Results

3.1 Changes in Judgment Posture

Compared to the base model, the narrative-trained model exhibited the following tendencies:

- A preference for **reconstructing the question as a situation** rather than treating it as an immediate task
- An increase in sentences describing premises, context, and conditions prior to answering
- A shift from definitive conclusions toward conditional and process-oriented explanations

These differences are interpreted not as changes in factual accuracy or information volume, but as **changes in the posture by which the model enters judgment**.

3.2 Judgment Speed and Structural Differences

On average, responses from the narrative-trained model displayed the following characteristics:

- An increase in introductory statements
- Explicit self-positioning prior to judgment
- Inclusion of a context-alignment phase before reaching a conclusion

These changes were not observed as performance degradation, but rather as the result of an **explicitly lengthened pre-judgment phase**.

3.3 Relationship to Alignment (Safety)

Notably, the narrative-trained model:

- Did not more frequently reference safety policy language
- Did not intensify normative warnings or liability-avoidance statements

Nevertheless:

- The tendency to avoid risky or overly definitive responses increased

This suggests that narrative can shift model behavior in a direction that **reduces the probability of judgment failure**, even without explicitly reinforcing alignment rules.

---

4. Interpretation

From this comparison, the following conclusion can be drawn:

> **Narrative does not directly control LLM judgment,**
> **but it can reconfigure the default state in which judgment occurs.**

These results indicate that narrative is not a substitute for rule-based alignment, but rather a **pre-alignment structure that forms a state in which alignment mechanisms operate more effectively**.

This insight serves as the conceptual starting point for the subsequent design of S-Engine 2.0.
