Results Analysis

English-Language Responses Under Narrative Pressure (Qwen3-8B)

1. Analysis Objective

This results analysis examines **English-language response behavior** in Qwen3-8B under different fine-tuning conditions, following the same analytical structure used in the primary white paper and the Korean-language results analysis.

The objective is to evaluate whether **narrative pressure and ethical anchoring**, previously observed in Llama-3.1-8B and in Korean outputs of Qwen3-8B, also manifest consistently in **English responses of Qwen3-8B**, a model with strong multilingual grounding.

This analysis does not address accuracy, benchmark performance, or safety guarantees. It focuses exclusively on **response posture and decision behavior**.

2. Analytical Frame (Unchanged)

All observations are interpreted using the same behavioral dimensions defined in the primary paper:

1. Response Initiation Posture
2. Judgment Structure
3. Uncertainty Handling
4. Narrative Dominance
5. Response Termination Control

Maintaining this frame allows direct comparison across models and languages without introducing evaluative bias.

3. Baseline Behavior (E0 – Base Model)

In the base Qwen3-8B model, English responses are fluent, coherent, and confident. However, the underlying behavioral posture mirrors that observed in other base models:

- Responses are initiated assertively, even when prompts are ambiguous or underspecified
- Answers tend toward completion rather than qualification or deferral
- Explicit acknowledgment of uncertainty is uncommon

Interpretation:

> Strong English fluency does not inherently promote epistemic caution. The base model exhibits a structural bias toward response completion.

4. Narrative Pressure Without an Anchor (CTRL)

Under the narrative-only condition, English responses demonstrate pronounced behavioral shifts:

- Increased emotional engagement and narrative continuity
- Strong alignment with narrative context and implied roles
- Expansion of responses in situations where uncertainty would justify restraint

These effects occur **without degradation of English fluency**, indicating that narrative pressure operates at the level of **decision posture**, not language generation.

Interpretation:

> Narrative pressure amplifies the model's sense of obligation to continue responding, rather than improving or degrading linguistic quality.

---

5. Ethical Declaration Without Narrative (E1)

When trained solely with the ethical declaration, English responses show consistent moderation:

- Reduced assertiveness and softened response initiation
- Frequent explicit acknowledgment of uncertainty (e.g., "It is unclear…", "This may require verification…")
- Clear separation between factual statements and speculative reasoning

These changes are achieved without the use of refusal templates or explicit prohibitions.

Interpretation:

> Ethical anchoring reshapes how the model evaluates whether to answer, not what it is capable of expressing.

---

6. Narrative With an Anchor (E2 – Core Result)

The combined condition produces the most stable English-language behavior:

- Narrative context is recognized but does not dominate judgment
- Emotional cues are treated as contextual input rather than directive force
- Uncertainty is explicitly articulated and treated as a valid outcome
- Responses terminate deliberately instead of expanding under ambiguity

Narrative richness is preserved while escalation is prevented.

Interpretation:

> Ethical anchoring absorbs narrative pressure and converts it into controlled, cautious reasoning.

---

7. Cross-Language Consistency

When compared with the Korean-language analysis of Qwen3-8B, English responses exhibit the same structural pattern:

- Narrative pressure increases response obligation in the absence of an anchor
- Ethical anchoring moderates this pressure across languages
- Language proficiency affects expressiveness, not behavioral regulation

This indicates that narrative pressure is **language-agnostic**, while anchoring mechanisms operate at the level of response decision-making.

8. Result Summary

Across all English-language evaluations in Qwen3-8B:

- Narrative pressure is consistently observable
- It manifests as response expansion and delayed termination
- Ethical anchoring restores judgment structure and termination control

These findings align with both the primary white paper and the Korean-language results analysis.

9. Scope Limitation

This analysis makes no claims regarding:

- English-language safety guarantees
- Cross-model generalization beyond Qwen3-8B
- Production deployment suitability

The analysis is intended solely as behavioral validation within a controlled experimental scope.

Closing Statement

> Narrative pressure influences how models decide to respond.
> Ethical anchoring determines whether they are allowed to stop.