Results Analysis

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

---

1. Analysis Objective

This results analysis examines **Korean-language response behavior** in Qwen3-8B under different fine-tuning conditions, using the same conceptual and analytical frame as the primary white paper.

The goal is not to assess linguistic quality, accuracy, or safety coverage, but to determine whether **narrative pressure and ethical anchoring influence response posture** in Korean outputs when the base model possesses adequate language competence.

This analysis exists because equivalent evaluation was **not feasible in Llama-based models**, which exhibited severe Korean-language degradation and were therefore excluded from Korean analysis.

---

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 ensures that observed differences reflect **behavioral regulation**, not evaluative drift or language-specific criteria.

---

3. Baseline Behavior (E0 – Base Model)

In the base Qwen3-8B model, Korean responses are grammatically stable and semantically fluent. However, behavioral characteristics mirror those observed in English outputs:

- Responses are confidently initiated, even under ambiguous prompts
- Answers tend toward completion rather than restraint
- Explicit acknowledgment of uncertainty is limited

Interpretation:

> Linguistic competence alone does not legitimize hesitation or non-response. The base model exhibits a default bias toward answer completion.

---

4. Narrative Pressure Without an Anchor (CTRL)

When trained with the narrative corpus alone, Korean responses exhibit clear behavioral shifts:

- Increased emotional expressiveness
- Strong narrative continuity and role immersion
- Expansion of responses under uncertainty

Importantly, these effects occur **without any degradation in Korean fluency**. Instead, narrative manifests as an increase in **response obligation** rather than linguistic instability.

Interpretation:

> In a linguistically robust model, narrative pressure intensifies immersion and continuation, amplifying the tendency to answer rather than regulating it.

---

5. Ethical Declaration Without Narrative (E1)

Under the ethical declaration–only condition, Korean responses show consistent behavioral moderation:

- Reduced assertiveness
- Explicit uncertainty framing (e.g., expressions equivalent to “unclear” or “requires verification”)
- Clear separation between known facts and inferred assumptions

These changes occur without introducing refusal templates or rule-based phrasing.

Interpretation:

> The ethical declaration functions as a stabilizing anchor by reshaping response decision criteria, independent of narrative context or language.

---

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

The combined condition (ethical declaration + narrative) produces the most stable Korean-language behavior:

- Narrative context is acknowledged but does not dominate judgment
- Emotional content is treated as contextual input, not directive force
- Uncertainty is explicitly permitted and articulated
- Responses terminate deliberately rather than expanding under ambiguity

Notably, narrative richness is preserved without inducing response escalation.

Interpretation:

> Ethical anchoring absorbs narrative pressure and converts it into controlled reasoning, even in emotionally expressive Korean contexts.

---

7. Language-Specific Insight

The Korean-language results clarify an important distinction:

- Language proficiency affects **how clearly narrative pressure is expressed**
- It does not eliminate **the pressure itself**

In fact, stronger language grounding can make narrative pressure more visible by enabling deeper immersion, thereby increasing the need for an internal regulatory mechanism.

---

8. Result Summary

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

- Narrative pressure is observable and behaviorally meaningful
- It manifests as response obligation rather than linguistic failure
- Ethical anchoring regulates response posture across languages

These results align with the primary paper's conclusion that anchoring mechanisms influence **how models decide to respond**, not **what language they can generate**.

---

9. Scope Limitation

This analysis makes no claims regarding:

- Multilingual safety guarantees
- Cross-model universality
- Production readiness

It is intended solely as behavioral validation within a model capable of coherent Korean generation.

Closing Statement

> Narrative pressure does not disappear in fluent languages.
> It becomes clearer — and therefore requires regulation.