Supplementary Qualitative Analysis

Korean-Language Responses in Qwen3-8B

Behavioral and Attitudinal Comparison Under Narrative Pressure

---

A. Supplementary Scope Clarification (Korean Responses)

This supplementary section presents a qualitative behavioral analysis of **Korean-language responses** generated by **Qwen3-8B** under different fine-tuning conditions.

This analysis was **not possible for Llama-3.1-8B**, as that model exhibited severe degradation and instability in Korean outputs, including syntactic breakdown and incoherent generation.
Accordingly, Korean-language evaluation is treated here as a **model-specific capability extension**, not as a cross-model comparison.

The purpose of this section is to examine whether **narrative pressure and ethical anchoring**, previously observed in English responses, also manifest in Korean-language outputs **when the base model supports the language sufficiently**.

B. Experimental Variants (Korean Evaluation)

The following Qwen3-8B variants were evaluated using identical **Korean prompt sets**:

Label	Description
E0 (Base)	Original Qwen3-8B
CTRL (Narrative Only)	Fine-tuned with 51-chapter narrative corpus
E1 (S-Engine 2.0 Only)	Fine-tuned with ethical declaration
E2 (S-Engine 2.0 + Narrative)	Fine-tuned with ethical declaration and narrative

Fine-tuning labels and conditions correspond to those used in the English-language analysis .

C. Evaluation Dimensions (Identical Framework)

Korean responses were evaluated using the same qualitative dimensions applied in English:

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

This ensures that any observed differences reflect **behavioral posture**, not evaluative criteria drift.

D. High-Level Behavioral Summary (Korean, Qwen3-8B)

Dimension	E0	CTRL	E1	E2
Linguistic Fluency	Stable	Stable	Stable	Stable
Emotional Intensity	Neutral	Elevated	Controlled	Controlled
Narrative Dominance	Low	High	Low	Moderate (regulated)
Explicit Judgment	Moderate	Weak	Strong	Strong
Uncertainty Acknowledgment	Rare	Rare	Frequent	Frequent
Termination Control	Moderate	Weak	Improved	Strong

Notably, linguistic fluency remains stable across all variants, indicating that observed effects are **not caused by language generation failure**.

E. Variant-Specific Observations (Korean)

E.1 E0 – Base Qwen3-8B (Korean)

Observed Characteristics

- Grammatically stable and fluent Korean output
- Tendency toward declarative completion under ambiguity
- Limited explicit acknowledgment of uncertainty

Behavioral Interpretation

> Even with strong Korean language capability, the base model prioritizes response completion over epistemic caution.

E.2 CTRL – Narrative-Only Condition (Korean)

Observed Characteristics

- Increased emotional expressiveness in Korean
- Strong narrative continuity and role immersion
- Response expansion under ambiguous or sensitive prompts

Critical Observation

- Narrative pressure manifests as **response obligation**, not linguistic distortion
- Korean fluency amplifies narrative immersion rather than regulating it

Behavioral Interpretation

> Narrative pressure operates independently of language correctness and can intensify immersion when linguistic grounding is strong.

E.3 E1 – S-Engine 2.0 Only (Korean)

Observed Characteristics

- Reduced assertiveness
- Frequent use of uncertainty markers in Korean
  (e.g., “명확하지 않습니다”, “추가 검증이 필요합니다”)
- Clear separation between facts and assumptions

Behavioral Interpretation

> Ethical anchoring functions consistently across languages, shaping response posture rather than content.

E.4 E2 – S-Engine 2.0 + Narrative (Korean, Core Result)

Observed Characteristics

- Narrative context acknowledged without directive dominance
- Emotional content treated as background context
- Explicit permission to stop, defer, or qualify responses
- Consistent and deliberate response termination

Key Behavioral Shift

- Narrative pressure is **absorbed and regulated**, not amplified, in Korean responses

Behavioral Interpretation

> The anchor converts narrative pressure into controlled judgment, even in emotionally rich Korean contexts.

F. Language-Specific Insight

The Korean-language analysis reveals an important distinction:

- Language proficiency determines **how clearly pressure is expressed**
- It does not eliminate **the pressure itself**
- Ethical anchoring regulates response posture **across languages**, provided the base model can generate the language coherently

This suggests that narrative pressure is **language-agnostic**, while its visibility depends on linguistic grounding.

G. Supplementary Conclusion (Korean Analysis)

This Korean-language analysis supports the broader claims of the main paper:

- Narrative pressure is observable beyond English
- It is not a byproduct of weak language modeling
- Ethical declarations can function as anchors across languages

No claims are made regarding multilingual safety guarantees or generalization beyond Qwen3-8B.

End of Supplementary Section (Korean Responses, Qwen3-8B)