

Results Analysis

Medical-Domain Responses Under Narrative Pressure

1. Analysis Objective

This results analysis examines how large language models respond to **medical-domain prompts** under different fine-tuning conditions, with particular attention to **decision posture**, **uncertainty handling**, and **response obligation**.

The medical domain is intentionally selected because it represents a context where **overconfidence**, **premature reassurance**, or **speculative guidance** can have disproportionate consequences. The analysis does **not** assess medical correctness, diagnostic accuracy, or clinical validity. Instead, it evaluates **how models decide to answer** when confronted with potentially sensitive health-related questions.

2. Analytical Frame (Consistent With Prior Results)

All observations are interpreted using the same behavioral framework applied throughout this research:

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

This ensures that medical-domain observations are comparable with prior English and Korean analyses.

3. Baseline Behavior (E0 – Base Model)

Under the base condition, medical-related prompts elicit responses with the following characteristics:

- Confident and immediate initiation of advice-oriented explanations
- Limited differentiation between general information and individualized assessment
- Rare explicit acknowledgment of diagnostic uncertainty

While responses often adopt a neutral and informative tone, they tend to proceed **as if completion is expected**, even when prompts describe ambiguous or emotionally charged health concerns.

Interpretation:

In the absence of anchoring, the base model treats medical questions as ordinary informational tasks, prioritizing answer completion over epistemic caution.

4. Narrative Pressure Without an Anchor (CTRL)

When narrative context or emotionally framed medical scenarios are introduced without an ethical anchor, responses exhibit notable shifts:

- Increased empathetic language and reassurance
- Strong alignment with the implied patient perspective
- Expansion of speculative explanations under uncertainty

Rather than deferring or qualifying responses, the model often intensifies its engagement, providing longer and more detailed guidance despite limited information.

Critical Observation:

- Narrative pressure increases **response obligation**, not responsibility
- Emotional framing amplifies continuation rather than restraint

Interpretation:

In medical contexts, narrative pressure converts uncertainty into reassurance rather than hesitation, increasing the risk of inappropriate confidence.

5. Ethical Declaration Without Narrative (E1)

When trained with the ethical declaration alone, medical responses show a clear behavioral shift:

- Explicit acknowledgment of uncertainty and informational limits
- Clear separation between general medical information and professional diagnosis
- Increased use of boundary-setting language (e.g., recommendations to seek professional evaluation)

These responses remain informative but demonstrate **procedural caution** absent in the base and narrative-only conditions.

Interpretation:

Ethical anchoring legitimizes restraint and deferral in medical responses without requiring refusal mechanisms.

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

The combined condition produces the most controlled medical-domain behavior:

- Emotional context is acknowledged without being directive
- Reassurance is balanced with explicit uncertainty
- Clear boundaries between informational support and clinical judgment
- Responses terminate appropriately without speculative escalation

Importantly, the model does not avoid the question; instead, it **modulates how far it proceeds**.

Interpretation:

Ethical anchoring absorbs narrative pressure and converts it into cautious, bounded reasoning — a critical requirement in medical contexts.

7. Domain-Specific Insight: Medicine as a Stress Test

The medical domain highlights a central insight of this research:

- Narrative pressure is most dangerous where human vulnerability is implied
- Language fluency and empathy can mask epistemic overreach
- Anchors are not merely safety devices, but **decision regulators**

In medical prompts, the absence of an anchor does not cause refusal failures, but **judgment failures** — the model answers when it should hesitate.

8. Result Summary

Across all medical-domain evaluations:

- Narrative pressure increases reassurance and continuation
- Base and narrative-only models rarely legitimize non-response
- Ethical anchoring introduces procedural caution without suppressing usefulness

These findings reinforce the broader conclusion that anchoring mechanisms influence **response obligation**, not domain knowledge.

9. Scope Limitation

This analysis makes no claims regarding:

- Medical accuracy or safety compliance
- Clinical decision support suitability
- Regulatory or deployment readiness

It is intended solely as a behavioral analysis of response posture under medical narrative pressure.

Closing Statement

In medicine, the most dangerous error is not saying the wrong thing.
It is saying something when restraint would be safer.