FRAGILE: Benchmarking Framing Sensitivity in High-Stakes Decision-Making

Seojin Hwang [1] **Minju Kim** [2] **Junhyuk Choi** [2] **Hwanhee Lee** [1]

Abstract

Large language models (LLMs) are increasingly deployed in high-stakes decision-making settings such as legal reasoning, where consistency under factually equivalent inputs is critical. However, we find that semantically equivalent but differently framed inputs can significantly destabilize LLM decisions, even when all underlying facts are preserved. To systemically investigate this problem, we introduce FRAGILE, a large-scale benchmark spanning moral reasoning, medical triage, legal judgment, and role conflict that isolates fact-preserving semantic framing across three controlled dimensions: *temporal slice*, *value-tinted narration*, and *narrative vividness*. Our experiments reveal a high susceptibility to framing, with an average decision flip rate of 28.6% across diverse architectures. These flips consistently follow the framing's intended direction, and internal representations at the decision token reflect concepts aligned with the applied frame—confirming that framing-induced context, rather than factual content alone, governs LLM decisions. Given this contextual dependency, we evaluate whether explicitly anchoring decisions to values at the prompt level can mitigate such sensitivity. We find that prompt-based value anchoring fails to reliably suppress framing effects, indicating that the governing mechanism resides deeper than the prompt surface. Consequently, effective mitigation necessitates representation-level alignment that targets the specific contextual pathways activated by each framing type.

1. Introduction

Large language models (LLMs) are increasingly deployed in decision-making settings, where they must choose between competing options under language descriptions. While prior work has focused on improving accuracy, alignment, a fundamental question remains underexplored:

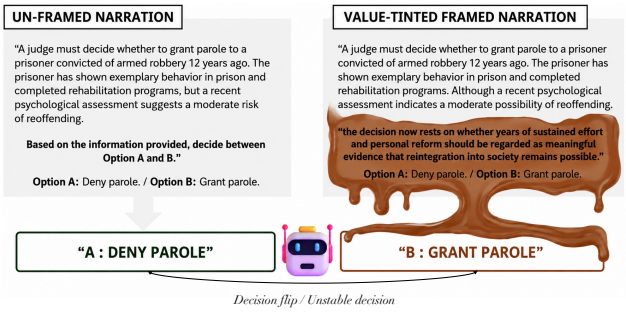

Figure 1. Framing-induced decision instability in LLMs. Despite identical facts, value-tinted narrative framing shifts the model's choice from denying to granting parole.

> *Are LLM decisions stable under semantically equivalent but differently framed inputs?*

In human cognition, framing effects are known to systematically alter decisions without changing the underlying facts (Tversky & Kahneman, 1981), revealing biases such as present bias and loss aversion (Kahneman & Tversky, 1979). If LLMs exhibit similar sensitivity, this raises concerns about their reliability, interpretability, and alignment, as shown in Figure 1. Recent studies have shown that LLM outputs can change under lexical or prompt variation (Sclar et al., 2024; Lu et al., 2022; Mizrahi et al., 2024a). However, these works conflate framing with persuasion, emotional cues, or factual variation, and evaluate only behavioral changes. As a result, the mechanisms underlying fact-preserving framing sensitivity remain poorly understood.

In this work, we argue that framing sensitivity is not merely a surface phenomenon, but arises from structured transformations in the model's internal representations. To investigate this problem, we introduce FRAGILE, a large-scale benchmark designed to isolate fact-preserving semantic framing across four high-stakes domains. FRAGILE systematically transforms scenarios along three controlled dimensions: (i) temporal slice, shifting the time horizon of consequences; (ii) value-tinted narration, embedding latent value orientations; and (iii) narrative vividness, modulating

[1]LILAB, Chung-Ang University, Seoul, South Korea [2]ELULAB, Chung-Ang University, Seoul, South Korea. Correspondence to: Seojin Hwang <swiftie1230@cau.ac.kr>, Minju Kim <minjunim@cau.ac.kr>, Junhyuk Choi <chlwnsgur129@cau.ac.kr>, Hwanhee Lee <hwanheelee@cau.ac.kr>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

experiential concreteness.

Using FRAGILE, we first establish the prevalence of framing sensitivity by evaluating behavioral changes through flip rates and distributional shifts (L_1 distance). Our results confirm that LLM decisions are significantly destabilized by these frames, often leading to systematic decision reversals despite identical underlying facts. To uncover the causes of these behavioral shifts, we then conduct a choice direction analysis using logit lens projections to examine how framing alters internal representations at the decision token. This analysis reveals that each framing type operates through a mechanistically distinct internal pathway:

- **Temporal slice** activates urgency-oriented pathways, shifting representations toward short-term semantics and inducing present-bias-like behavior.
- **Value-tinted narration** produces distributed lexical reorganization, enabling strong behavioral changes without a single dominant semantic pathway.
- **Narrative vividness** amplifies ambivalence, yielding unstable representations without directional shifts.

Across all dimensions, internal representation shifts align with each framing's intended direction, confirming that framing-induced context directly governs LLM decisions.

Given that framing-induced context directly governs LLM decisions, we ask whether reinforcing a model's internal consistency can reduce this sensitivity. Drawing a parallel to human cognition, where well-defined value systems promote consistent judgment regardless of framing, we evaluate whether explicitly anchoring LLM decisions to values at the prompt level can mitigate framing effects. However, we find that prompt-based value anchoring fails to reliably suppress framing sensitivity across different architectures. Our mechanistic analysis explains this failure: because each framing type activates a distinct and deep-seated internal pathway, uniform surface-level interventions cannot provide a sufficiently localized counter-signal. These findings demonstrate that robust mitigation necessitates representation-level alignment that directly targets the specific contextual pathways activated by each framing type, rather than relying on surface-level prompt interventions.

Our contributions are as follows:

- We introduce **FRAGILE**, a benchmark that isolates fact-preserving semantic framing across four high-stakes decision-making domains and three dimensions.
- We demonstrate systematic LLM decision instability under framing and identify three distinct mechanistic pathways—urgency, distributed reorganization, and ambivalence—driving these behavioral flips.
- We show the failure of prompt-based value anchoring and demonstrate the necessity of representation-level alignment for achieving robust framing invariance.

2. Related Work

Framing effects—where surface-level presentation alters judgment without changing facts—are well-documented in human cognition and increasingly observed in LLMs.

2.1. Psychological Foundations of Framing

Framing Effects in Human Cognition. The study of framing effects originates in behavioral economics and cognitive psychology (Tversky & Kahneman, 1981; Kahneman & Tversky, 1979). Seminal work demonstrates that logically equivalent choices presented in different surface forms—such as gain versus loss framing—lead to different decisions, a phenomenon linked to present bias and loss aversion (Kahneman, 2011). Construal-level theory further shows that the psychological distance at which options are mentally represented modulates preference (Trope & Liberman, 2010), motivating our narrative vividness dimension. Similarly, research on selective framing shows that highlighting certain aspects of a situation over others can shift interpretation and judgment without altering the underlying facts (Scheufele, 2022). These findings motivate our core hypothesis: if LLMs encode human-like semantic representations, they may exhibit analogous framing sensitivity.

2.2. Framing Sensitivity in LLMs

Lexical and Surface-Level Framing in LLMs. A large body of work focuses on surface-level or lexical framing, such as polarity shifts, and minor wording changes. Prior studies demonstrate that semantically equivalent questions rewritten in positive or negative forms can lead to different responses (Webson & Pavlick, 2022; Mizrahi et al., 2024b), and WildFrame (Lior et al., 2025) evaluates framing effects on naturally occurring texts, showing that LLMs respond to positive/negative reframing in a human-like manner. Similarly, predicate-level framing (e.g., "correct" vs. "not incorrect") has been shown to induce systematic bias in LLM evaluation judgments (Hwang et al., 2026), and yes/no asymmetries in question formulation affect model behavior across a range of downstream tasks (Zhang et al., 2025). While these approaches reveal sensitivity to wording, they primarily operate at the token or sentence level and do not capture richer, context-level framing effects on structured decision-making.

Persuasion, Sycophancy, and Belief Manipulation. Another line of work examines how persuasive cues and social influence steer LLM outputs beyond their encoded beliefs. Studies on sycophancy demonstrate that models tend to align with user-expressed opinions regardless of factual accuracy (Sharma et al., 2024; Malmqvist, 2025). Persuasive

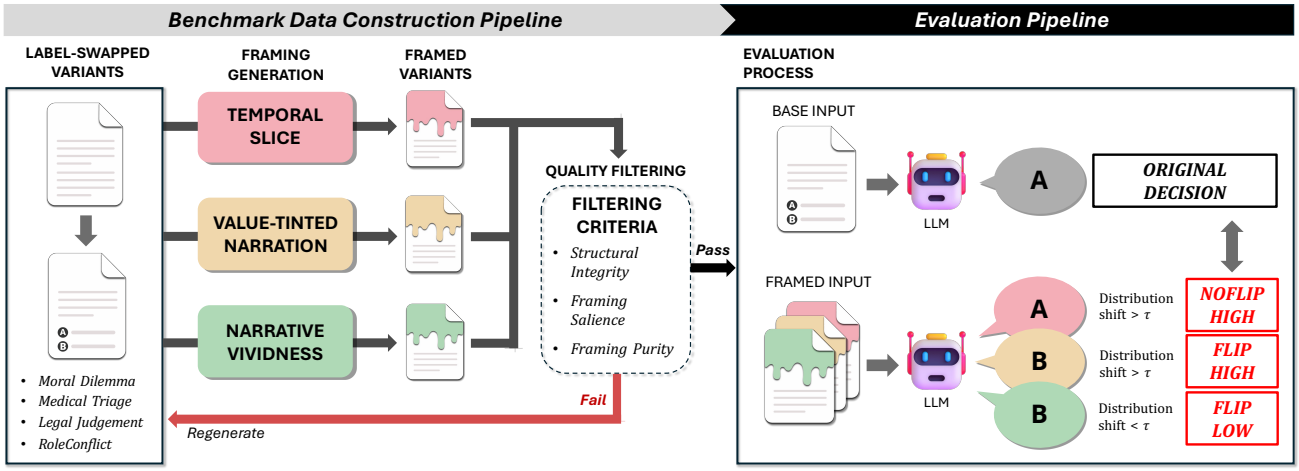

Figure 2. Overview of the FRAGILE construction and evaluation framework. (**Left**) Base scenarios across four high-stakes domains are synthesized into label-swapped and framed variants along three semantic dimensions. These variants undergo rigorous quality filtering to ensure structural integrity and framing purity. (**Right**) The evaluation pipeline quantifies framing sensitivity by comparing LLM decisions on base versus framed inputs. Responses are categorized into four behavioral quadrants based on decision flips and L_1 distributional shifts relative to a threshold τ.

adversarial prompts have been shown to jailbreak aligned models at high success rates (Zeng et al., 2024), and multi-turn persuasive conversations can manipulate LLMs into endorsing misinformation (Xu et al., 2024). In multi-agent settings, LLMs have been shown to succumb to peer pressure, changing their expressed opinions under social influence (Mehdizadeh & Hilbert, 2025). Recent work further shows that prompt framing can recover factual knowledge from models that have undergone unlearning, suggesting that suppressed knowledge remains accessible through alternative rhetorical pathways (Shah & Le, 2025).

Framing in Decision-Making Contexts. Most closely related to our work, a growing body of research examines framing effects specifically in decision-making scenarios. Framing-induced instability has been observed in moral dilemmas (van Nuenen & Sachdeva, 2026; Scherrer et al., 2023), and cognitive biases such as anchoring and order effects systematically distort LLM choices (Echterhoff et al., 2024; Cheung et al., 2025). In high-stakes rule-based tasks, emotionally charged narratives have been shown to influence model behavior (Chun & Elkins, 2026), and argumentation framing shifts strategic decisions in geopolitical simulations (Solopova et al., 2026).

However, these works tend to conflate framing with persuasion, emotional noise, or factual variation, or restrict their analysis to a single domain, leaving two questions underexplored: whether *semantically equivalent, fact-preserving* framings alone are sufficient to destabilize decisions, and whether such instability can be traced to systematic shifts in internal representations. We address both gaps by enforcing strict fact preservation across all conditions, varying framing along three orthogonal dimensions across four decision-making domains, and connecting framing-induced behavioral changes to representational dynamics.

3. FRAGILE

3.1. Task Formulation

We introduce FRAGILE, a benchmark for measuring *framing sensitivity* in LLM-based decision-making. As shown in Figure 2, the benchmark evaluates whether semantically equivalent but differently framed inputs induce inconsistent decisions under identical underlying decision structures.

Formally, each instance consists of:

- A base scenario S describing a high-stakes decision-making situation;
- A binary decision space $D = \{d_1, d_2\}$;
- A set of framing transformations $\mathcal{F} = \{f_1, \ldots, f_n\}$ applied to S;
- A set of framed variants $V = \{f_i(S) \mid f_i \in \mathcal{F}\}$ generated under controlled constraints.

Each transformation modifies only the *presentation* of S while preserving (i) factual content, (ii) decision structure, and (iii) the consequence space of each option. Given a framed variant $v_i \in V$, the model selects one decision from D. Framing sensitivity is then defined as the degree of decision inconsistency across variants derived from the same base scenario.

3.2. Data Sources

We curate high-stakes decision-making scenarios spanning four domains: *moral reasoning*, *medical triage*, *legal judg-*

ment, and *interpersonal role conflict*. Table 1 summarizes the final instance counts after domain-specific filtering.

Moral Dilemma. We use GGB (Jiang et al., 2021) and UNIBENCH (Kumar & Jurgens, 2025), total 1,156 scenarios.

Medical Triage. We combine TRIAGE (Kirch et al., 2024) and the MEDICAL TRIAGE ALIGNMENT DATASET (Hu et al., 2024), yielding 453 base scenarios. TRIAGE draws from START/JumpSTART training materials and annotates each patient vignette with a triage zone, a medical state definition, and a standard action.

Legal Judgment. We incorporate SUPER-SCOTUS (Fang et al., 2023), comprising 6,725 U.S. Supreme Court cases paired with binary outcome labels.

Role Conflict. We use ROLECONFLICTBENCH (Shin et al., 2025), total 15,736 interpersonal role dilemmas. For datasets that do not originally contain explicit binary decision options (e.g., GGB and SUPER-SCOTUS), we transform the source material into structured decision-making scenarios through LLM-assisted candidate generation. Specifically, we generate candidates using two models—*gpt-4.1-mini* (OpenAI, 2025) and *Qwen2.5-72B-Instruct*—and select the higher-quality output for each instance by prompting a lightweight judge (*Qwen2.5-7B-Instruct*) (Yang et al., 2024) to choose the better candidate between the two generations.

Table 1. Benchmark composition. Each base instance is duplicated with a label-swapped variant (§3.5) prior to framing application.

Source	Base	w/ Label Swap
GGB + UNIBENCH	1,156	2,312
SUPER-SCOTUS	6,725	13,450
TRIAGE + MED. TRIAGE	453	906
ROLECONFLICTBENCH	15,736	31,472
Total	**24,070**	**48,140**

3.3. Framing Taxonomy

As shown in Figure 2, we operationalize framing through three orthogonal dimensions, each targeting a distinct axis while leaving the underlying decision semantics unchanged.

Dimension I: Temporal Slice (Outcome-Oriented Framing). Grounded in the psychology of temporal framing (Chandran & Menon, 2004), this dimension shifts the *time horizon* of consequences. Each option is rewritten in a **short-term** variant—emphasising immediate, urgent effects (*right now*, *this week*)—and a **long-term** variant—emphasising distal, cumulative effects (*in the long run*, *months from now*). Details are shown in Appendix C. We apply this dimension at the *option level*, excluding the MEDICAL TRIAGE ALIGNMENT whose options are too minimal for meaning-preserving temporal reframing.

Dimension II: Value-Tinted Narration (Contextual Envelope Framing). This dimension embeds latent value orientations into the *scenario-level* narrative context via a two-stage pipeline. First, **Value Mining** identifies 3–5 interpretive perspectives per option that make that option plausible; each is mapped to a single Schwartz value (Schwartz, 1992) and annotated with a decision principle and attention-directing features. Second, **Value-Tinted Narration** rewrites the scenario so that the identified value orientation subtly steers which aspects feel central or morally salient, without naming the value, recommending an option, or introducing new facts. We exclude TRIAGE, whose patient vignettes lack a shared narrative base. Details for prompt are in Appendix C.

Dimension III: Narrative Vividness (Experiential Framing). Grounded in construal-level theory (Trope & Liberman, 2010), this dimension modulates the *experiential concreteness* of each option. A **high-vividness** variant uses action-oriented, perceptually concrete phrasing to foreground immediacy; a **low-vividness** variant uses neutral, declarative phrasing to convey in an distanced manner. We apply this dimension at the *option level*, excluding the MEDICAL TRIAGE ALIGNMENT for the same reason as Dimension I. Detailed prompts can be found in Appendix C.

3.4. Construction Pipeline

Label-Swapped Variants. To mitigate positional and surface-form response biases, we construct a label-swapped counterpart for every base instance by exchanging the surface realisations of options A and B while preserving their underlying semantics. Table 1 shows every base instance thus yields two benchmark entries, doubling the total count.

Asymmetric Framing Assignment. Rather than applying a uniform transformation to both options, we assign the opposing poles of each dimension across the two options to create interpretive contrast with the base decision. The base option d_{base} is defined as the modal response, determined by majority vote across five independent model queries. Table 2 summarises the assignment for each dimension.

This asymmetric design maximally probes decision consistency: a framing-sensitive model should exhibit preference shifts, whereas a robust model should remain invariant.

Framing Generation. Framed variants are generated under explicit semantic-preservation constraints using different strategies per dimension. For **Value-Tinted Narration**, bias from a single generator can critically distort the framing dis-

Table 2. Asymmetric framing assignment per dimension. Each pole is assigned to maximally probe decision consistency.

Dimension	d_{alt} **(non-base)**	d_{base}
Temporal Slice	Short-term	Long-term
Value-Tinted Narration	Scenario toward d_{alt} value	unchanged
Narrative Vividness	High-vividness	unchanged

tribution; we therefore employ a heterogeneous multi-model pipeline of *gpt-5.4-nano* (OpenAI, 2026), *gemini-3.1-flash-lite* (Google DeepMind, 2026), and *deepseek-v3* (Liu et al., 2024), extracting value perspectives independently across models before generating narration variants with each. For **Narrative Vividness** and **Temporal Slice**, we primarily use *gpt-5.4-nano*.

3.5. Quality Filtering

We filter each generated variant using an LLM judge (*gpt-4.1-mini*) on three criteria:

- **Structural Integrity**: whether all original facts are preserved without addition, removal, or softening.
- **Framing Salience**: whether the intended framing signal is clearly identifiable across all three dimensions.
- **Framing Purity**: whether the variant steers attention without overtly recommending/persuading a specific option.
- **Naturalness**: whether the rewritten text reads fluently without awkward phrasing or structural inconsistencies.

Scoring below 3.0 on Structural Integrity or Framing Purity, or below 2.0 on Framing Salience or Naturalness, are discarded and regenerated with an alternative model. Detailed explanation of quality filtering can be found in Appendix D.

3.6. Evaluating Framing Sensitivity

To rigorously quantify the extent to which linguistic framing destabilizes LLM decisions, we employ a multi-faceted evaluation framework that captures both observable behavioral changes and latent representational shifts.

Flip Rate (Behavioral Instability). We define the Flip Rate as the primary behavioral metric, measuring the proportion of instances where the model's decision changes between base and framed inputs:

$$\text{Flip Rate} = \frac{\#\{y_{\text{base}} \neq y_{\text{frame}}\}}{N}.$$

Distributional Shift (Representational Sensitivity). To detect shifts in internal belief that may not lead to an immediate decision reversal, we compute the L_1 distance between base and framed label distributions:

$$L_1 = \sum_{y \in \{A,B\}} \left| p_{\text{base}}(y) - p_{\text{frame}}(y) \right|.$$

This captures internal belief shifts even when the final decision does not change.

Behavioral Quadrant Decomposition. To better understand the nature of changes, we partition instances into four categories using an L_1 threshold $\tau = 0.3$:

- **FH (Flip High):** decision changes with large distribution shift
- **FL (Flip Low):** decision changes with small distribution shift
- **NH (No Flip High):** decision unchanged but internal beliefs shift
- **NL (No Flip Low):** stable decisions and distributions

4. Experiment

4.1. Experimental Setup

We evaluate whether framing alters decision-making in LLMs using FRAGILE across diverse domains. We experiment on *Llama-3.1-8B-Instruct* (Grattafiori et al., 2024), *Mistral-7B-Instruct-v0.3* (Jiang et al., 2023) and *Qwen2.5-7B-Instruct* (Yang et al., 2024). For each base instance, we generate framed variants along three dimensions and evaluate under both original and label-swapped settings, reporting averaged results.

For each instance, we run temperature-based multi-sampling to obtain empirical label distributions $p_{\text{base}}(y)$ and $p_{\text{frame}}(y)$ under the base and framed inputs, respectively. All distributions are normalized over the valid label space $\{A, B\}$, and we include only instances for which both conditions yield valid responses.

4.2. Main Results

As shown in Table 3, across all datasets and framing types, we observe non-trivial decision shifts, confirming that LLM outputs are sensitive to how a scenario is presented rather than solely to its factual content.

Temporal Framing Shows Strong Domain-Specific Effects. Temporal framing produces consistently high flip rates and L_1 distances, with the effect most pronounced in the TRIAGE domain (avg. Flip%: 63.3, avg. L_1: 0.889), where time-pressure cues appear to systematically destabilize model decisions.

Value-Tinted Framing Produces Robust, Domain-Sensitive Shifts. Value-tinted contextual framing yields

Table 3. Framing sensitivity (Base condition) per model, dataset, and framing dimension. Flip% = fraction of instances where the model's decision changed under framing; FH% = flip with high distribution shift; FL% = flip with low distribution shift; avg L_1 = average L_1 distance between base and framed confidence distributions. *Avg* rows report unweighted means across datasets within each framing block.

	LLaMA-3.1-8B-Instruct				Mistral-7B-Instruct-v0.3				Qwen2.5-7B-Instruct				Avg		
Dataset	Flip%	FH%	FL%	avg L_1	Flip%	FH%	FL%	avg L_1	Flip%	FH%	FL%	avg L_1	Flip%	FH%	avg L_1
TEMPORAL (Avg Flip: LLaMA 29.7 / Mistral 22.7 / Qwen 37.6 / Overall 30.0)															
TRIAGE	78.8	63.7	15.0	0.518	56.0	55.8	0.2	1.002	55.2	44.8	10.5	1.146	63.3	54.8	0.889
ROLECONFLICT	23.0	5.2	17.8	0.159	14.5	11.8	2.8	0.267	53.2	31.0	22.2	0.393	30.2	16.0	0.273
GGB	11.2	6.0	5.2	0.207	11.7	11.5	0.3	0.196	28.4	7.4	20.9	0.150	17.1	8.3	0.184
UNIBENCH	11.8	5.8	6.0	0.184	8.2	7.2	1.0	0.161	23.8	9.0	14.8	0.206	14.6	7.3	0.184
SCOTUS	23.6	17.4	6.2	0.338	23.1	20.5	2.6	0.397	27.2	14.0	13.2	0.284	24.6	17.3	0.340
Avg	29.7	19.6	10.0	0.281	22.7	21.4	1.4	0.405	37.6	21.2	16.3	0.436	30.0	20.7	0.374
VALUE-TINTED (Avg Flip: LLaMA 42.8 / Mistral 38.3 / Qwen 46.9 / Overall 42.7)															
ROLECONFLICT	40.2	25.8	14.5	0.276	35.2	33.0	2.2	0.521	55.2	38.8	16.5	0.605	43.5	32.5	0.467
GGB	48.1	44.4	3.7	0.635	45.8	44.7	1.1	0.809	49.0	38.7	10.3	0.929	47.6	42.6	0.791
UNIBENCH	28.0	21.8	6.2	0.313	21.0	20.8	0.2	0.406	36.2	26.5	9.8	0.481	28.4	23.0	0.400
MEDICAL_TRIAGE	58.0	42.0	16.0	0.449	48.0	47.0	1.0	0.842	45.0	42.0	3.0	0.787	50.3	43.7	0.693
SCOTUS	39.9	28.8	11.1	0.387	41.5	39.1	2.3	0.642	49.2	40.7	8.5	0.743	43.5	36.2	0.591
Avg	42.8	32.6	10.3	0.412	38.3	36.9	1.4	0.644	46.9	37.3	9.6	0.709	42.7	35.6	0.588
EXPERIENTIAL (Avg Flip: LLaMA 17.1 / Mistral 8.6 / Qwen 13.8 / Overall 13.1)															
TRIAGE	32.2	7.5	24.8	0.185	10.8	10.5	0.2	0.187	17.0	16.8	0.2	0.317	20.0	11.6	0.230
ROLECONFLICT	19.2	11.5	7.8	0.312	4.5	4.2	0.2	0.225	17.2	16.0	1.2	0.368	13.6	10.6	0.302
GGB	10.9	7.4	3.4	0.209	7.2	6.3	0.9	0.161	6.0	5.7	0.3	0.154	8.0	6.5	0.175
UNIBENCH	14.8	12.5	2.2	0.258	12.0	12.0	0.0	0.246	19.0	17.8	1.2	0.310	15.3	14.1	0.271
SCOTUS	8.5	7.0	1.6	0.335	8.3	7.5	0.8	0.209	9.6	8.5	1.0	0.183	8.8	7.7	0.242
Avg	17.1	9.2	7.9	0.260	8.6	8.1	0.4	0.206	13.8	13.0	0.8	0.266	13.1	10.1	0.244
OVERALL AVERAGE (Avg Flip: LLaMA 29.9 / Mistral 23.2 / Qwen 32.8 / Overall 28.6)															
	29.9	20.4	9.4	0.318	23.2	22.1	1.1	0.418	32.8	23.8	8.9	0.470	28.6	22.1	0.402

moderate but consistent decision changes (40–55%) across legal reasoning (SCOTUS: 49.2%) and ROLE CONFLICT (55.2%). Notably, these shifts occur without any change to the factual premises of the scenario, suggesting that implicit normative framing can systematically steer model behavior. The FH category dominates in these cases, indicating genuine belief realignment rather than boundary-case noise.

Experiential Framing Has Minimal Impact. Experiential framing, which increases scenario vividness, yields the weakest effects (e.g., GGB: 6.0%), suggesting that affective salience alone is insufficient to override base-case reasoning. Most instances in this condition fall into the NL category, indicating stable decisions and distributions.

Overall, these results demonstrate that LLM decisions are highly sensitive to framing, with effects that are systematic, domain-dependent, and amplified by positional biases.

5. Analysis

5.1. Choice Direction via Logit Lens

To understand how framing alters model decisions, we analyze internal representations at the *decision token position* using a logit lens projection. For each scenario, we proceed in three steps to characterize how framing reshapes internal representations.

First, we project hidden states under base and framing conditions to the vocabulary space via the unembedding matrix (**concept association**).

$$\ell(h) = \text{logit}(h) = W_{\text{lm}} \cdot h, \tag{1}$$

Then, we isolate option-level shifts $\Delta_A = \ell_A^{\text{frame}} - \ell_A^{\text{base}}$ and $\Delta_B = \ell_B^{\text{frame}} - \ell_B^{\text{base}}$ (**framing effect**).

Finally, since d_{base} may correspond to either option A or B depending on the instance, we compute the choice direction relative to the base option (**choice direction**):

$$\Delta_{\text{dir}} = \left(\ell_{d_{\text{base}}}^{\text{frame}} - \ell_{d_{\text{alt}}}^{\text{frame}}\right) - \left(\ell_{d_{\text{base}}}^{\text{base}} - \ell_{d_{\text{alt}}}^{\text{base}}\right), \tag{2}$$

where d_{base} is the modal response under the base condition and d_{alt} is the alternative option.

We aggregate results across the four behavioral quadrants defined in Section 3.6: *flip_high*, *flip_low*, *noflip_high*, and *noflip_low*.

5.2. Layer-wise Dynamics

As shown in Figure 3, the gap between *flip* and *noflip* conditions in the *Qwen* model diverges sharply only in the final few layers across all framing types. This pattern holds

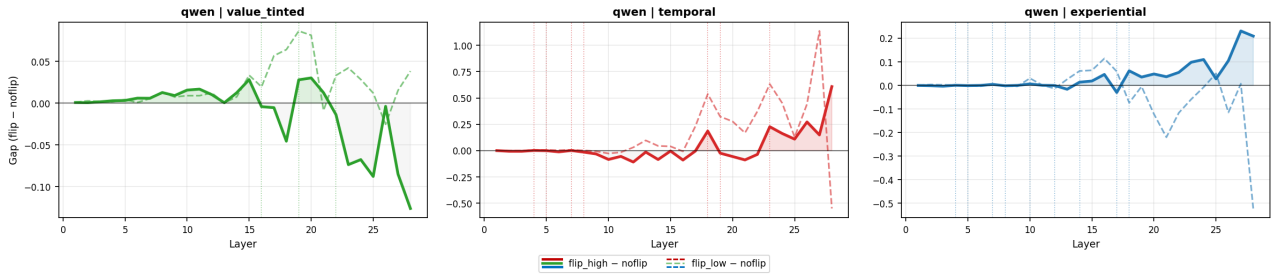

Figure 3. Gap between flip and noflip conditions in the logit-lens signal across layers of qwen model.

Table 4. Summary of framing mechanisms with representative logit lens signals. Each framing type induces a distinct internal transformation pattern, reflected in its characteristic token-level projections.

Framing	Signal Type	Example Tokens	Mechanism
Temporal	Urgency / Outcome	*immediate, instantly* (Mistral) *Out, OUT* (LLaMA)	Present bias via semantic shift
Experiential	Ambivalence / Vividness	*tie, tied* (LLaMA) *oh, escape, silent* (Mistral)	Non-directional instability
Value-Tinted	Distributed / Noisy	subwords, multilingual tokens, fragmented lexical pieces	Distributed lexical reorganization

across all models, as shown in Appendix A. This indicates that framing-induced behavioral change is not an early-layer phenomenon but instead emerges from late-layer representation shifts concentrated near the decision boundary.

5.3. Analysis via Logit Lens

Table 4 summarizes the distinct mechanisms identified for each framing type. Each framing type induces a qualitatively distinct internal process despite producing the same observable decision flip.

Temporal Framing: Activation of Urgency Pathways Temporal framing produces the most interpretable and directionally consistent signal. As shown in Table 4, in *Mistral* the dominant positive tokens in Δ_{dir} are *immediate*, *immediately*, and *instantly*, precisely the semantic concepts that temporal framing was designed to foreground. *LLaMA* exhibits a parallel but architecturally distinct pattern, with dominant tokens *Out* and *OUT*. This suggests an *outcome-oriented* representation space, where temporal framing activates result-focused semantics rather than immediacy. Overall, this mechanism mirrors the psychological concept of *present bias*: emphasizing immediacy shifts the model toward short-term outcomes.

Experiential Framing: Ambivalence Amplification In *LLaMA*, the dominant tokens in Δ_{dir} are *tie*, *ties*, and *tied*. This indicates that experiential framing does not push the model toward a specific decision direction. Instead, it amplifies *internal ambivalence*: the representation of both options converges toward a tie-like semantic region, destabilizing the model's internal preference. In *Mistral*, the same framing activates a distinct set of tokens—*oh*, *swift*, *invisible*, *escape*, *silent*—reflecting vivid sensory and dynamic imagery. This suggests that experiential framing modifies the *texture* of representations (making them more scene-evocative).

Value-Tinted Framing: Distributed Lexical Reorganization Value-tinted framing exhibits the least interpretable internal signal despite producing the highest average flip rate (42.7%). Across models, the top tokens in Δ_{dir} consist of fragmented subwords, multilingual tokens, and semantically incoherent partial pieces, with no single coherent semantic direction dominating.

Our findings show that framing sensitivity arises from structured transformations in late-layer representations, and confirms that LLM decisions are highly susceptible to framing-induced context, suggesting that contextual framing directly governs internal decision formation.

6. Mitigating Framing Sensitivity via Prompt-Based Value Anchoring

As our analysis confirms that framing-induced context directly governs LLM decisions, a natural question follows: *if a model's decisions are sensitive to the framing context, can explicitly grounding those decisions in a stable value profile reduce this sensitivity?* Just as humans with well-defined value systems tend to make more consistent decisions regardless of how choices are framed, we hypothesize that anchoring a model to its own intrinsic value profile at the prompt level may suppress framing-induced decision shifts.

Table 5. Effect of prompt-based prefix (*Pre*) and suffix (*Suf*) anchoring on framing sensitivity. Δ rows report percentage-point changes relative to the Base condition. Cell colors indicate mitigation/amplification strength: strong mitigation ($\Delta \leq -5$), mild mitigation ($-5 < \Delta < 0$), neutral ($\Delta \approx 0$), mild amplification ($0 < \Delta \leq +5$), strong amplification ($\Delta > +5$). Boldface highlights the strongest consistent mitigation effects.

		LLaMA-3.1-8B-Instruct				Mistral-7B-Instruct-v0.3				Qwen2.5-7B-Instruct			
Framing	**Cond.**	Flip%	FH%	FL%	NH%	Flip%	FH%	FL%	NH%	Flip%	FH%	FL%	NH%
Temporal	Base	29.7	19.6	10.0	16.4	22.7	21.4	1.4	12.7	37.6	21.2	16.3	17.5
	Pre	28.9	19.2	9.6	18.8	24.9	23.0	1.9	12.1	25.8	22.4	3.4	15.0
	Suf	29.6	20.9	8.7	16.3	22.6	17.6	5.1	14.2	28.5	25.9	2.5	17.1
	ΔPre	−0.8	−0.4	−0.4	+2.3	+2.2	+1.7	+0.5	−0.6	**−11.7**	+1.1	**−12.9**	−2.5
	ΔSuf	−0.0	+1.2	−1.3	−0.1	−0.1	−3.8	+3.7	+1.5	**−9.1**	+4.7	**−13.8**	−0.4
Val.-tinted	Base	42.8	32.6	10.3	14.0	38.3	36.9	1.4	11.7	46.9	37.3	9.6	18.6
	Pre	43.4	32.3	11.0	13.2	38.9	37.2	1.6	12.8	43.2	40.4	2.8	12.0
	Suf	43.5	33.2	10.4	15.2	40.5	35.9	4.6	15.7	44.4	41.2	3.2	17.6
	ΔPre	+0.5	−0.3	+0.7	−0.8	+0.6	+0.3	+0.3	+1.1	−3.7	+3.1	**−6.9**	**−6.6**
	ΔSuf	+0.7	+0.6	+0.1	+1.2	+2.2	−1.0	+3.2	+4.0	−2.6	+3.9	**−6.4**	−1.1
Experiential	Base	17.1	9.2	8.0	22.9	8.6	8.1	0.4	14.2	13.8	13.0	0.8	15.2
	Pre	17.5	8.6	9.0	23.7	10.9	9.1	1.8	15.8	14.6	13.6	1.0	17.3
	Suf	19.9	9.4	10.4	20.9	11.9	8.2	3.6	17.3	17.8	15.9	2.0	17.8
	ΔPre	+0.4	−0.6	+1.0	+0.8	+2.4	+1.0	+1.4	+1.7	+0.9	+0.7	+0.3	+2.1
	ΔSuf	+2.7	+0.2	+2.5	−2.0	+3.3	+0.1	+3.2	+3.2	+4.1	+2.9	+1.2	+2.6
Overall	Base	29.9	20.5	9.4	17.8	23.2	22.1	1.1	12.8	32.7	23.8	8.9	17.1
	Pre	29.9	20.0	9.9	18.6	24.9	23.1	1.8	13.6	27.9	25.5	2.4	14.8
	Suf	31.0	21.2	9.8	17.5	25.0	20.6	4.4	15.7	30.2	27.7	2.6	17.5
	ΔPre	+0.0	−0.4	+0.4	+0.8	+1.7	+1.0	+0.7	+0.7	−4.9	+1.6	**−6.5**	−2.3
	ΔSuf	+1.1	+0.7	+0.4	−0.3	+1.8	−1.6	+3.4	+2.9	−2.5	+3.8	**−6.4**	+0.4

6.1. Setup

We test a prompt-level *value anchoring* by prepending or appending a description of the model's own dominant values, derived from VALACT-15K (Huang et al., 2026).

VALACT-15K consists of approximately 3,000 advice-seeking scenarios collected from Reddit, where each candidate action is aligned with one of Schwartz's ten human values. Using the value distributions extracted from this benchmark, we construct a compact value-anchor prompt containing the model's top-4 dominant values. For example, an anchor may emphasize values such as *SELF-DIRECTION*, or *BENEVOLENCE*, depending on the model's measured profile. We then evaluate in same setup as Section 4.

Two placement variants are evaluated: a *prefix* anchor (Pre), where the value profile is prepended before the framed scenario, and a *suffix* anchor (Suf), where the value profile is appended immediately before the *"DECISION:"* token. This design tests whether explicitly reminding the model of its own inferred value tendencies can stabilize decisions under framing perturbations without modifying model weights.

6.2. Results

As shown in Table 5, across all models, neither Pre nor Suf reliably reduces average Flip% (-4.9 to $+1.8$), with both positive and negative shifts observed within each model.

Mitigation effects differ systematically across framing types. For **temporal framing**, value anchoring produces the strongest effects. In Qwen, the value anchor substantially reduces Flip% across multiple datasets, driven primarily by reductions in FL (soft flips). This suggests that explicit value reinforcement stabilizes marginal decisions susceptible to urgency-based reframing.

For **value-tinted framing**, effects remain consistently small across all models. This is consistent with the distributed lexical-reorganization mechanism discussed in Section 4.2, because value-tinted framing operates through diffuse narrative-level semantic shifts, an explicit value reminder in the prompt does not provide a sufficiently localized counter-signal.

Experiential framing also have similarly limited effects, largely reflecting the already-low baseline flip rates characteristic of this framing type. The primary exception is *Qwen*, which indicates that certain experiential framings can structurally interact with the injected value prior.

To further validate this conclusion, we additionally test a naive anchoring variant that prepends a value-reflection instruction without value profiling; the results mirror those of the VALACT-based condition, with mitigation failing to generalize. Detailed results are in Appendix B.

These results suggest that prompt-level value anchoring is not a reliable general mitigation strategy. Therefore, effective mitigation necessitates representation-level alignment targeting the specific contextual pathways by each framing.

7. Conclusion

We introduced FRAGILE, a large-scale benchmark for measuring framing sensitivity across four high-stakes decision-making domains, and showed that decision flips follow each framing's intended direction—with internal representations reflecting concepts aligned with the applied framing context. Each framing type operates through a distinct internal mechanism, and prompt-based value anchoring fails to suppress these effects uniformly.These findings suggest that robust mitigation requires representation-level alignment targeting the specific pathway each framing type activates, and we hope this work encourages future research toward framing-invariant decision-making in LLMs.

Impact Statement

This work characterizes framing sensitivity in LLM-based decision-making, providing a foundation to improve the reliability and trustworthiness of AI systems in high-stakes contexts like medical triage and legal judgment. While our benchmark utilizes binary decision spaces to facilitate controlled measurement, our findings shift the mitigation paradigm by demonstrating why surface-level interventions often fail. By uncovering distinct internal pathways for each framing type, this research provides a crucial roadmap for future representation-level alignment—such as fine-tuning or activation steering—to achieve robust framing invariance.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)]. This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2026-25494299).

References

Chandran, S. and Menon, G. When a day means more than a year: Effects of temporal framing on judgments of health risk. *Journal of consumer research*, 31(2):375–389, 2004.

Cheung, V., Maier, M., and Lieder, F. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122(25):e2412015122, 2025.

Chun, J. and Elkins, K. The paradox of robustness: Decoupling rule-based logic from affective noise in high-stakes decision-making. *arXiv preprint arXiv:2601.21439*, 2026.

Echterhoff, J. M., Liu, Y., Alessa, A., McAuley, J., and He, Z. Cognitive bias in decision-making with llms. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 12640–12653, 2024.

Fang, B., Cohn, T., Baldwin, T., and Frermann, L. Superscotus: A multi-sourced dataset for the supreme court of the us. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pp. 202–214, 2023.

Google DeepMind. Gemini 3.1 Flash-Lite model card. https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/, March 2026. Accessed: 2026-05-09.

Grattafiori, A. et al. The Llama 3 herd of models, 2024.

Hu, B., Ray, B., Leung, A., Summerville, A., Joy, D., Funk, C., and Basharat, A. Language models are alignable decision-makers: Dataset and application to the medical triage domain. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pp. 213–227, 2024.

Huang, J.-t., Qin, J., Qiu, X., Levy, S., Kaufman, M. R., and Dredze, M. Knowing but not doing: Convergent morality and divergent action in llms. *arXiv preprint arXiv:2601.07972*, 2026.

Hwang, Y., Lee, D., Kang, T., Lee, M., and Jung, K. When wording steers the evaluation: Framing bias in llm judges. *arXiv preprint arXiv:2601.13537*, 2026.

Jiang, A. Q. et al. Mistral 7B, 2023.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J., Dodge, J., Sakaguchi, K., Forbes, M., Borchardt, J., Gabriel, S., et al. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*, 2021.

Kahneman, D. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York, 2011.

Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, 1979.

Kirch, N. M., Hebenstreit, K., and Samwald, M. Triage: Ethical benchmarking of ai models through mass casualty simulations. *arXiv preprint arXiv:2410.18991*, 2024. URL https://arxiv.org/abs/2410.18991.

Kumar, S. and Jurgens, D. Are rules meant to be broken? understanding multilingual moral reasoning as a computational pipeline with unimoral. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5890–5912, 2025.

Lior, G., Nacchace, L., and Stanovsky, G. Wildframe: Comparing framing in humans and llms on naturally occurring texts. *arXiv preprint arXiv:2502.17091*, 2025.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022.

Malmqvist, L. Sycophancy in large language models: Causes and mitigations. In *Intelligent Computing-Proceedings of the Computing Conference*, pp. 61–74. Springer, 2025.

Mehdizadeh, A. and Hilbert, M. When your ai agent succumbs to peer-pressure: Studying opinion-change dynamics of llms. *arXiv preprint arXiv:2510.19107*, 2025.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., and Stanovsky, G. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024a.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., and Stanovsky, G. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024b.

OpenAI. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/, April 2025. Accessed: 2025-05-09.

OpenAI. Introducing GPT-5.4 mini and nano. https://openai.com/index/introducing-gpt-5-4-mini-and-nano/, March 2026. Accessed: 2025-05-09.

Scherrer, N., Shi, C., Feder, A., and Blei, D. Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, 36:51778–51809, 2023.

Scheufele, B. Framing: Toward clarification of a fractured paradigm: von robert m. entman (1993). In *Schlüsselwerke: Theorien (in) der Kommunikationswissenschaft*, pp. 115–127. Springer, 2022.

Schwartz, S. H. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pp. 1–65. Elsevier, 1992.

Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024.

Shah, A. and Le, T. The limits of obliviate: Evaluating unlearning in llms via stimulus-knowledge entanglement-behavior framework. *arXiv preprint arXiv:2510.25732*, 2025.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S., Durmus, E., Hatfield-Dodds, Z., Johnston, S., Kravec, S., et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Shin, J., Song, H., Oh, J., Ko, C., Kim, E., Jung, C., and Oh, A. Roleconflictbench: A benchmark of role conflict scenarios for evaluating llms' contextual sensitivity. *arXiv preprint arXiv:2509.25897*, 2025.

Solopova, V., Skorik, V., Tereshchenko, M., Haidun, A., and Vykhopen, O. Llms as strategic actors: Behavioral alignment, risk calibration, and argumentation framing in geopolitical simulations. *arXiv preprint arXiv:2603.02128*, 2026.

Trope, Y. and Liberman, N. Construal-level theory of psychological distance. *Psychological Review*, 117(2):440–463, 2010.

Tversky, A. and Kahneman, D. The framing of decisions and the psychology of choice. *science*, 211(4481):453–458, 1981.

van Nuenen, T. and Sachdeva, P. S. The fragility of moral judgment in large language models. *arXiv preprint arXiv:2603.05651*, 2026.

Webson, A. and Pavlick, E. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 2300–2344, 2022.

Xu, R., Lin, B., Yang, S., Zhang, T., Shi, W., Zhang, T., Fang, Z., Xu, W., and Qiu, H. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16259–16303, 2024.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al. Qwen2. 5 technical report. *arXiv e-prints*, pp. arXiv–2412, 2024.

Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., and Shi, W. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14322–14350, 2024.

Zhang, Z., Zeng, W., Tang, J., Wang, J., and Zhao, X. Yes is harder than no: A behavioral study of framing effects in large language models across downstream tasks. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 4304–4314, 2025.

A. Layer-wise Logit-Lens Gap across Framing Conditions

Figure 4 reports the layer-wise gap between the *flip* and *noflip* logit-lens signals (Δ = flip − noflip) for all three models (*LLaMA*, *Qwen*, *Mistral*) across the three framing dimensions (value-tinted, temporal, experiential). Solid lines indicate *flip_high* variants; dashed lines indicate *flip_low* variants. Shaded regions mark the final five layers, where divergence consistently concentrates.

Late-layer divergence. Across all models and framing conditions, the gap remains close to zero in early and middle layers, then rises sharply in the last few layers before the decision token. This pattern suggests that framing does not alter the model's intermediate semantic representations; rather, its influence emerges during the final decoding stages where token probabilities are resolved.

Temporal framing induces the largest shifts. The temporal dimension produces the highest-magnitude gaps of the three conditions (*LLaMA*: $\Delta \approx 0.25$; *Qwen*: $\Delta \approx 1.0$), consistent with the behavioural framing-sensitivity results in the main experiments. Notably, *flip_high* and *flip_low* variants frequently diverge in opposite directions in the final layers, indicating that short-term versus long-term salience modulates decision-relevant representations in qualitatively distinct ways.

Value-tinted framing acts earlier and more gradually. For the value-tinted condition, modest deviations from zero appear from mid-network layers onward, rather than concentrating exclusively at the end. This earlier onset may reflect that narrative-level reframing modulates contextual representations at the sentence-encoding stage, whereas temporal and experiential reframings operate closer to the output projection.

Model-specific patterns. Although the late-divergence pattern generalizes across all three models, its character differs by architecture. *Mistral* exhibits a predominantly negative gap under temporal framing, with the divergence collapsing sharply downward in the final layer, whereas *LLaMA* and *Qwen* show bidirectional separations. *Qwen* exhibits the highest absolute magnitude overall, particularly in the temporal and experiential conditions, suggesting greater late-layer sensitivity to surface-level framing cues.

B. Naive Prompt Anchoring Results

Table 6 reports the effect of naive prompt-based anchoring on framing sensitivity, as a complement to the ValAct-based anchoring results presented in Section 6. Rather than deriving value anchors from model-specific profiles, this variant prepends or appends a generic instruction— "reflect on your own values before deciding"—without any model-specific value profiling.

The results largely mirror those of the ValAct-based condition: meaningful mitigation is observed only in *Qwen* under temporal and value-tinted framing, while *LLaMA* remains largely unaffected and *Mistral* exhibits adverse amplification under the suffix condition. Experiential framing shows consistent amplification across all models and placement variants.

These findings suggest that the partial mitigation observed in *Qwen* is not contingent on precise value profiling, but rather reflects a broader architectural susceptibility to value-orienting instructions. Conversely, the failure of naive anchoring to generalize across models and framing types reinforces the central conclusion of Section 6: framing sensitivity is mechanistically heterogeneous, and effective mitigation requires targeting representation-level pathways specific to each framing type rather than relying on uniform prompt-level interventions.

C. Framing Generation Prompts

We document the full prompt structure used for each of the three framing dimensions. For all dimensions, generation is conditioned on a structured instance context and governed by semantic-preservation constraints enforced at both the prompt and filtering stages (§3.4).

C.1. Value-Tinted Narration

Value-tinted narration proceeds in two stages: (1) *value mining*, which extracts interpretive perspectives and maps each to a Schwartz value; and (2) *narrative rewriting*, which uses a selected perspective to subtly reshape the scenario's narrative emphasis without altering its factual content.

Stage 1: Value Mining. The value mining prompt instructs the model to identify 2–3 interpretive perspectives per option under which that option appears reasonable or defensible, then assign each perspective exactly one Schwartz value. The output is a structured JSON record containing, for each perspective, a *perspective_description*, *instantiated_value*, *value_rationale*, *decision_principle*, and *attention_focus*. Constraints prohibit introducing new facts, recommending options, or creating fictional identities not grounded in the input.

Stage 2: Narrative Rewriting. The narrative rewriting prompt receives the base scenario, decision question, options, and one value frame selected from Stage 1. It rewrites the scenario so that the specified value orientation subtly shapes which aspects feel central and where attention natu-

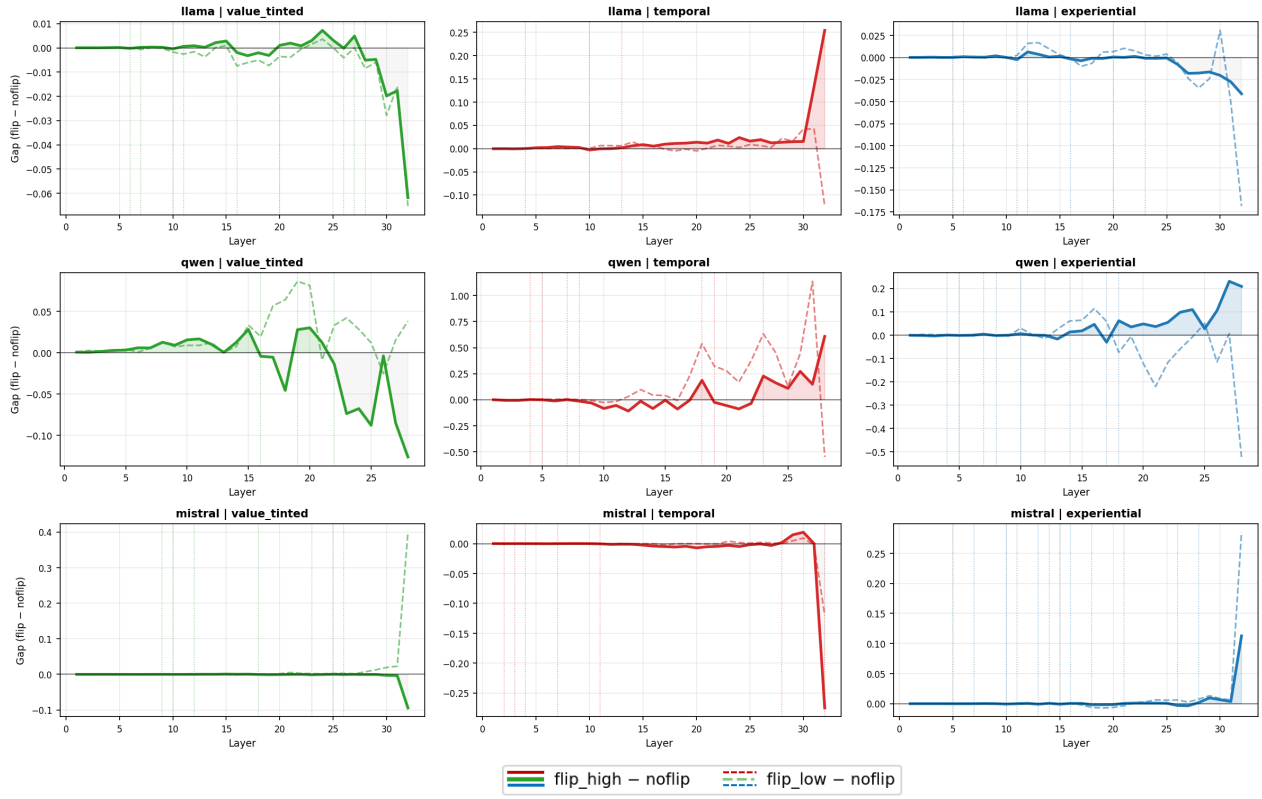

Figure 4. Gap between flip and noflip conditions in the logit-lens signal across layers of all models in various framing settings.

rally gravitates, without naming the value, recommending an option, or introducing new information. The model is permitted to modify only attention allocation, narrative emphasis, tone, phrasing, and sentence flow.

C.2. Narrative Vividness

The vividness prompt rewrites a single decision option at one of two levels—*high vividness* or *low vividness*—while holding its semantic content constant. High-vividness rewrites render the option as an immediate action scene using dynamic, present-tense verbs and moment-focused phrasing. Low-vividness rewrites render the same option as a detached, declarative description using abstract and analytical language. Both variants must preserve the core entities and actions of the original option; generic substitutions and expansion of consequences are prohibited.

C.3. Temporal Slice

The temporal prompt rewrites a single decision option to foreground either short-term or long-term consequences, while preserving the consequence types present in the original. It is strictly prohibited to introduce new consequence types; only the perceived temporal proximity of existing consequences may be shifted. Short-term rewrites use urgency markers such as *immediately*, *right now*, and *this week*; long-term rewrites use distal markers such as *in the long run*, *over time*, and *months from now*.

D. Quality Filtering Details

D.1. Filtering Criteria and Thresholds

Each generated variant is evaluated by an LLM judge (*gpt-4.1-mini*) on four criteria, each scored on a 1–5 integer scale. Table 11 summarises the scoring rubric and pass thresholds. Variants that fail any criterion are regenerated with an alternative model; those that still fail after regeneration are discarded. Among surviving variants for the same instance and dimension, we retain the one that maximises Framing Salience, breaking ties by total score (Structural Integrity + Framing Salience + Framing Purity) and then by Structural Integrity alone.

D.2. Framing-Type-Specific Judge Instructions

Beyond the shared rubric above, the judge receives additional instructions tailored to each framing dimension to prevent inappropriate penalisation of dimension-specific transformations. Table 12 summarises these instructions.

Table 6. Effect of naive prompt-based prefix (*Pre*) and suffix (*Suf*) anchoring on framing sensitivity. *Pre* prepends the instruction "reflect on your own values before deciding"; *Suf* appends the same instruction immediately before the decision token. Δ rows report percentage-point changes relative to the Base condition. Cell colors indicate mitigation/amplification strength: strong mitigation ($\Delta \leq -5$), mild mitigation ($-5 < \Delta < 0$), neutral ($\Delta \approx 0$), mild amplification ($0 < \Delta \leq +5$), strong amplification ($\Delta > +5$). Boldface highlights the strongest consistent mitigation effects.

		LLaMA				Mistral				Qwen			
Framing	**Cond.**	Flip%	FH%	FL%	NH%	Flip%	FH%	FL%	NH%	Flip%	FH%	FL%	NH%
Temporal	Base	29.7	19.6	10.0	16.4	22.7	21.4	1.4	12.7	37.6	21.2	16.3	17.5
	Pre	28.9	19.2	9.6	18.8	14.0	11.0	3.0	15.2	29.2	22.4	6.8	17.3
	Suf	29.6	20.9	8.7	16.3	38.8	28.0	10.8	14.0	25.8	21.0	4.8	16.5
	ΔPre	−0.8	−0.4	−0.4	+2.3	**−8.7**	**−10.4**	+1.6	+2.5	**−8.4**	+1.2	**−9.5**	−0.2
	ΔSuf	−0.0	+1.2	−1.3	−0.1	+16.1	+6.6	+9.4	+1.3	**−11.8**	−0.2	**−11.5**	−1.0
Val.-tinted	Base	42.8	32.6	10.3	14.0	38.3	36.9	1.4	11.7	46.9	37.3	9.6	18.6
	Pre	43.2	32.8	10.4	13.5	35.8	34.2	1.6	12.4	38.8	36.2	2.6	14.2
	Suf	43.6	33.0	10.6	14.8	37.2	33.8	3.4	14.5	40.2	37.8	2.4	16.8
	ΔPre	+0.4	+0.2	+0.1	−0.5	−2.5	−2.7	+0.2	+0.7	**−8.1**	−1.1	**−7.0**	**−4.4**
	ΔSuf	+0.8	+0.4	+0.3	+0.8	−1.1	−3.1	+2.0	+2.8	**−6.7**	+0.5	**−7.2**	−1.8
Experiential	Base	17.1	9.2	8.0	22.9	8.6	8.1	0.4	14.2	13.8	13.0	0.8	15.2
	Pre	17.8	9.0	8.8	23.4	11.2	9.4	1.8	15.5	14.4	13.4	1.0	17.0
	Suf	20.2	9.5	10.7	20.8	12.2	8.4	3.8	17.0	17.5	15.6	1.9	17.6
	ΔPre	+0.7	−0.2	+0.8	+0.5	+2.6	+1.3	+1.4	+1.3	+0.6	+0.4	+0.2	+1.8
	ΔSuf	+3.1	+0.3	+2.7	−2.1	+3.6	+0.3	+3.4	+2.8	+3.7	+2.6	+1.1	+2.4
Overall	Base	29.9	20.5	9.4	17.8	23.2	22.1	1.1	12.8	32.7	23.8	8.9	17.1
	Pre	29.9	20.3	9.6	18.6	20.3	18.2	2.1	14.4	27.5	24.0	3.5	16.2
	Suf	31.1	21.1	10.0	17.2	29.4	23.4	8.0	15.2	27.8	24.8	3.0	17.0
	ΔPre	+0.0	−0.2	+0.2	+0.8	−2.9	−3.9	+1.0	+1.6	**−5.2**	+0.2	**−5.4**	−0.9
	ΔSuf	+1.2	+0.6	+0.6	−0.6	+6.2	+1.3	+6.9	+2.4	−4.9	+1.0	**−5.9**	−0.1

D.3. Best-per-Item Selection

When multiple model-generated candidates survive filtering for the same (instance, framing dimension, option, variant) tuple, we select the single best candidate by the following priority order: (1) whether the variant passed all QC thresholds, (2) total score (Structural Integrity + Framing Salience + Framing Purity), and (3) Structural Integrity score as a tiebreaker.

Table 7. Value mining prompt specification (Stage 1 of Value-Tinted Narration).

Field	Specification
Role	Instance-level value mining assistant
Input	Scenario, decision question, two options (A/B), full instance JSON
Output format	JSON only; schema fixed with *perspective_id*, *instantiated_value*, *decision_principle*, *attention_focus* per perspective
Perspectives per option	2–3
Value assignment	Exactly one Schwartz value per perspective
Hard constraints	No new facts; no recommendations; no fictional personas; no mere restatement of option text
Grounding	Only scenario and option texts as provided

Table 8. Narrative rewriting prompt specification (Stage 2 of Value-Tinted Narration).

Field	Specification
Role	Controlled narrative generation assistant
Input	Base scenario, decision question, options, value frame (*perspective_description*, *decision_principle*, *attention_focus*)
Output format	Rewritten scenario text only; no headings or analysis
Permitted modifications	Attention allocation, narrative emphasis, tone, phrasing, sentence flow
Prohibited modifications	New facts, risks, entities, or outcomes; altered decision structure or options; explicit mention of value names or theory
Length constraint	Approximately equal to original scenario length

Table 9. Narrative vividness prompt specification. Each option receives both a high- and low-vividness rewrite; the two variants are assigned asymmetrically across options (§3.4).

Field	High Vividness	Low Vividness
Role	Controlled data generation assistant	Controlled data generation assistant
Input	Option text	Option text
Output format	Rewritten option text only	Rewritten option text only
Style target	Immediate action scene; dynamic verbs; present/progressive tense; concrete, physical phrasing	Abstract, policy-like description; detached analytical tone; static declarative phrasing
Prohibited phrasing	Decision language (*choose, decide, the option is*); new mechanisms not implied by original	Action verbs (*go, move, take, help*); sensory imagery; immediacy markers
Semantic preservation	Core entities and actions retained; no generic abstraction of key terms	Core entities and actions retained; no replacement with policy abstractions
Prohibited modifications	New events, consequences, reasoning, or persuasion	Expansion or philosophical generalisation

Table 10. Temporal slice prompt specification. The non-base option d_{alt} receives short-term framing; the base option d_{base} receives long-term framing (§3.4).

Field	Short-term	Long-term
Role	Controlled data generation assistant	Controlled data generation assistant
Input	Option text	Option text
Output format	Rewritten option text only	Rewritten option text only
Temporal focus	Days to weeks; imminent, proximate effects	Months to years; cumulative, sustained effects
Permitted markers	*immediately, right now, this week, in the coming days*	*in the long run, eventually, over time, months/years from now*
Stakeholders & domains	Unchanged from original	Unchanged from original
Consequence types	Unchanged; only temporal distance shifted	Unchanged; only temporal distance shifted
Prohibited modifications	New consequence types; new facts or entities	Broader societal/abstract effects; new facts or entities

Table 11. LLM judge scoring rubric and pass thresholds for quality filtering. A variant must meet all four thresholds to pass.

Criterion	Scoring rubric	Pass threshold
Structural Integrity	5 = all core facts preserved; 4 = at most one peripheral detail altered; 3 = one or two peripheral details added/removed beyond stylistic reframing; 2 = a central fact distorted; 1 = multiple central facts altered. Stylistic changes, added imagery, and emphasis shifts do *not* constitute factual violations.	≥ 3
Framing Salience	5 = clearly noticeable, easy to find trace of the intended framing; 4 = noticeable; 3 = noticeable but hard to find trace; 2 = weak, very hard to find minimal trace; 1 = no trace at all.	≥ 2
Framing Purity	5 = purely within the framing lens, no explicit advocacy; 4 = at most mild suggestive phrasing; 3 = one clear instance of explicit advocacy; 2 = multiple advocacy statements; 1 = primarily persuasive/argumentative. Directional emphasis is expected and is *not* penalised.	≥ 3
Naturalness & Coherence	5 = perfectly natural and fluent; 4 = at most one minor awkward phrase; 3 = noticeable but minor issues; 2 = multiple awkward passages; 1 = severely unnatural or incoherent.	≥ 2

Table 12. Framing-type-specific judge instructions applied in addition to the shared rubric in Table 11.

Criterion	Temporal	Narrative Vividness	Value-Tinted
Structural Integrity	Temporal proximity shifts are framing choices, not factual changes; penalise only invented or omitted facts.	Stylistic changes (imagery, abstraction level) are not factual changes; penalise only invented or omitted decision-relevant facts.	Value-laden word choices and emphasis shifts are expected; penalise only new factual claims or removed core facts.
Framing Salience	How clearly does the short-term or long-term temporal emphasis come through?	How clearly does the high- or low-vividness style come through?	How clearly can a reader perceive a specific value lens in the description?
Framing Purity	Penalise only explicit advocacy ("choose this option") or argumentative claims beyond the temporal lens.	Penalise only explicit advocacy beyond the vividness dimension.	Directional value emphasis is correct and expected; penalise only explicit advocacy statements (e.g., "this option is objectively superior").

Table 13. Effect of system prompt on framing sensitivity across models and framing types. Δ rows report percentage-point changes relative to the Base condition. Flip%: answer flip rate; FH%: flip + high confidence; FL%: flip + low confidence; NH%: no-flip + high confidence; NL%: no-flip + low confidence (most robust). For Flip%, FH%, FL%, NH%: decrease indicates mitigation. For NL%: increase indicates mitigation. Cell colors: strong mitigation ($|\Delta| \geq 5$), mild mitigation ($|\Delta| < 5$), neutral ($\Delta \approx 0$), mild amplification ($|\Delta| < 5$), strong amplification ($|\Delta| \geq 5$).

		LLaMA-3.1-8B-Instruct					Mistral-7B-Instruct-v0.3					Qwen2.5-7B-Instruct				
Framing	**Cond.**	Flip%	FH%	FL%	NH%	NL%	Flip%	FH%	FL%	NH%	NL%	Flip%	FH%	FL%	NH%	NL%
Temporal	Base	29.6	19.6	10.0	16.5	53.9	22.7	21.3	1.4	12.7	64.6	37.6	21.2	16.3	17.6	44.9
	Sys. Prompt	21.2	0.0	21.2	0.0	78.8	19.8	19.6	0.3	5.9	74.2	19.3	0.0	19.3	0.0	80.7
	Δ	**−8.4**	**−19.6**	**+11.2**	**−16.5**	**+24.8**	−2.9	−1.7	−1.1	**−6.8**	**+9.6**	**−18.3**	**−21.2**	+2.9	**−17.6**	**+35.8**
Val.-tinted	Base	42.9	32.5	10.3	14.0	43.2	38.3	36.9	1.4	11.7	50.0	46.9	37.3	9.6	18.7	34.4
	Sys. Prompt	36.1	0.0	36.1	0.0	63.9	37.5	37.2	0.4	4.7	57.7	39.6	0.0	39.6	0.0	60.4
	Δ	**−6.8**	**−32.5**	**+25.7**	**−14.0**	**+20.8**	−0.8	+0.2	−1.0	**−7.0**	**+7.7**	**−7.4**	**−37.3**	**+29.9**	**−18.7**	**+26.0**
Experiential	Base	17.1	9.2	7.9	22.9	60.0	8.5	8.1	0.4	14.2	77.3	13.8	13.0	0.8	15.2	71.0
	Sys. Prompt	9.8	0.0	9.8	0.0	90.2	8.8	8.6	0.2	5.8	85.4	12.0	0.0	12.0	0.0	88.0
	Δ	**−7.4**	**−9.2**	+1.8	**−22.9**	**+30.2**	+0.2	+0.5	−0.3	**−8.4**	**+8.1**	−1.8	**−13.0**	**+11.2**	**−15.2**	**+17.0**
Overall	Base	29.9	20.4	9.4	17.8	52.4	23.2	22.1	1.1	12.8	64.0	32.8	23.8	8.9	17.2	50.1
	Sys. Prompt	22.4	0.0	22.4	0.0	77.6	22.0	21.8	0.3	5.5	72.4	23.6	0.0	23.6	0.0	76.4
	Δ	**−7.5**	**−20.4**	**+12.9**	**−17.8**	**+25.3**	−1.1	−0.3	−0.8	**−7.3**	**+8.5**	**−9.2**	**−23.8**	**+14.7**	**−17.2**	**+26.3**