NUISANCE-INDUCED CONFOUNDING IN LLM OUTPUT SENSITIVITY ANALYSIS: A SEMANTIC NORMALIZATION FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Distribution-based sensitivity analysis aims to quantify how large language model (LLM) outputs change under input perturbations, but raw output distributions can conflate semantic changes with nuisance variation from stochastic generation, surface form, language, and representation choice. We formalize this nuisance-confounding problem by modeling LLM outputs as mixtures of meaning and nuisance factors, showing that raw-output distributional sensitivity can be nonzero even when the underlying meaning is unchanged. Motivated by a nuisance-invariance criterion for meaning-centered sensitivity analysis, we formulate a semantic normalization framework that maps raw outputs into meaning-focused representations before distributional comparison. From a conditional information bottleneck perspective, this transformation is interpreted as approximately preserving meaning while suppressing nuisance-dependent variation. Experiments on controlled perturbation analyses show that meaning-preserving nuisance changes induce nonzero raw-output sensitivity, whereas summary-based normalization reduces this nuisance-induced sensitivity by 53.6% on average while retaining task-relevant content. In multilingual patient-data experiments with controlled semantic changes, summary-based normalization yields more comparable sensitivity estimates across languages. These results characterize nuisance confounding as a measurement problem and support summary-based semantic normalization as a practical mechanism for mitigating its effects.

1 INTRODUCTION

1.1 BACKGROUND

Large language models (LLMs) are increasingly used in settings where generated responses inform users' judgments and downstream decisions (Brown et al., 2020; Liu et al., 2023). This makes it important not only to evaluate average task performance, but also to understand how model outputs vary under small changes in input conditions. In real-world settings, however, LLM outputs may vary substantially even when the input meaning is largely unchanged. Small changes in phrasing or formatting can alter the structure, style, or length of generated responses (Sclar et al., 2024). While such variation may be acceptable or even desirable for generation purposes, it becomes problematic when output differences are used as quantitative evidence of input sensitivity.

A key source of this difficulty is the probabilistic generation process of LLMs. Even for identical or highly similar inputs, different samples can produce outputs that differ in wording, ordering, specificity, or style (Holtzman et al., 2020; Welleck et al., 2024). Thus, the observed change in an output distribution after an input perturbation reflects both changes in the semantic content of the responses and nuisance variation in how that content is expressed. This mixture makes it difficult to interpret whether a measured output change represents genuine semantic sensitivity or incidental variation unrelated to meaning.

Distribution-based approaches to sensitivity analysis compare sets of outputs generated before and after input perturbations (Errica et al., 2025). These methods typically embed the generated outputs into a vector space and quantify the differences between the resulting output distributions. However,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

distributional differences in embedding space do not necessarily correspond to semantic differences. They may also reflect sampling randomness, surface-level differences in style or expression, response length, formatting, or biases introduced by the embedding model used for analysis. Consequently, raw output distributional sensitivity can be difficult to interpret as a measure of semantic sensitivity.

This problem is further amplified in multilingual settings. Even when inputs are intended to preserve the same meaning across languages, language-specific syntax, lexical choices, discourse conventions, and cross-lingual representation quality can change the geometry of output embeddings (Xu et al., 2025; Zhong et al., 2024; Zhang et al., 2023). As a result, sensitivity estimates may vary across languages not only because the model responds differently to meaning, but also because language-dependent nuisance variation is represented differently in the analysis space. These observations motivate a sensitivity analysis framework that reduces nuisance variation in generated outputs while preserving their core meaning before comparing output distributions.

1.2 Brief Related Work

LLM outputs exhibit substantial variability due to stochastic decoding and inference-time sampling choices (Holtzman et al., 2020; Welleck et al., 2024; Loya et al., 2023; Wang & Wang, 2025; Minh et al., 2025). Prompt formatting and surface-form variation can further affect model behavior even when the underlying task meaning is largely preserved (Sclar et al., 2024; Bartsch et al., 2023). In multilingual settings, this variability is amplified by linguistic differences, training data imbalance, language-dependent model behavior, and imperfect cross-lingual alignment in embedding or latent spaces (Xu et al., 2025; Zhong et al., 2024; Zhang et al., 2023; Huang et al., 2023; Lim et al., 2025; Schut et al., 2025).

To quantify output sensitivity, prior work has proposed distribution-based approaches that compare embedding distributions before and after input perturbations (Errica et al., 2025; Rauba et al., 2025). While these methods capture variability at the distribution level, they are inherently influenced by embedding representations and surface-level variation, making it difficult to disentangle semantic differences from nuisance-induced variability (Gupta et al., 2023). From an information-theoretic perspective, summarization can be interpreted as a representational transformation that preserves task-relevant information while suppressing irrelevant variation under the Information Bottleneck principle (Tishby et al., 2000; Ju et al., 2021; West et al., 2019). However, how to control nuisance effects while preserving sensitivity to semantic differences remains a challenge in distribution-based analysis of LLM outputs. A more detailed discussion of related work is provided in Appendix B.

1.3 Contributions

We examine a nuisance-confounding problem in the distribution-based sensitivity analysis of LLM outputs. Although such analyses aim to quantify how outputs change under input perturbations, raw output distributions may conflate semantic changes with nuisance variation arising from stochastic generation, surface form, language, and embedding-model choice. We provide a theoretical account of this problem by modeling LLM outputs as mixtures of meaning and nuisance factors. We then use this formulation to motivate summary-based semantic normalization as one practical approach to mitigating nuisance effects in sensitivity analysis.

Our contributions are as follows:

- We formalize LLM outputs as mixtures of semantic meaning and nuisance factors, and show that raw-output distributional sensitivity can be nonzero even when the underlying meaning is unchanged. This reveals a structural confounding problem: observed differences between raw output distributions do not necessarily correspond to semantic differences.
- We characterize a sufficient condition for meaning-centered sensitivity analysis. If a representation is conditionally independent of nuisance factors given meaning, then inputs with the same meaning will induce the same representation distribution despite shifts in nuisance factors. This provides a formal criterion for representations suitable for semantic sensitivity analysis.
- We interpret summarization as an approximate mechanism for constructing nuisance-invariant representations. From a conditional information bottleneck (CIB) perspective, we formulate the ideal summary as a representation that preserves information about meaning while suppressing

irrelevant nuisance information conditional on meaning. This also connects summary-based normalization to an idealized variance-reduction baseline for sensitivity analysis.

- Through controlled perturbation analyses, we provide empirical evidence that raw-output distributional sensitivity is confounded by meaning-preserving nuisance variation. Consistent with the proposed nuisance-invariance perspective, summary-based semantic normalization reduces nuisance-induced sensitivity while retaining task-relevant content and improves the cross-lingual comparability of sensitivity estimates.

2 Theoretical Investigation of Sensitivity Analysis in LLMs

We start from the following observation: distribution-based sensitivity analysis of LLM outputs aims to capture semantic changes, yet the resulting sensitivity scores may be unstable not only due to genuine semantic variation but also due to nuisance variability unrelated to meaning. In this section, we formalize this issue and examine conditions for mitigating nuisance effects in distribution-based sensitivity analysis.

2.1 Notations, Definitions, Output Generation, and Sensitivity Analysis

We formalize a summarization-based distributional sensitivity analysis for its theoretical analysis.

- M: a random variable representing the core meaning that an LLM output intends to convey.
- N: a random variable representing nuisance factors (e.g., decoding randomness, stylistic/length variation, linguistic bias).
- Y: a random variable for the raw output text generated by the LLM.
- g: a summarization function that maps an output y into a meaning-focused summary.
- $S = g(Y)$: the resulting summarized random variable.

We model the LLM output generation process as a mixture of meaning M and nuisance N (Kunievsky & Evans, 2025):

$$Y \sim p(y \mid M, N). \tag{1}$$

Here, M denotes the shared semantic content that should remain invariant under changes in wording or language, while N captures non-semantic variation introduced by sampling, decoding temperature, language/style differences, etc. We denote a generic sensitivity analysis method for Y by $\Psi(\cdot)$. Thus, we treat $\Psi(Y)$ as a real-valued random variable induced by the randomness of Y. Existing sensitivity analysis methods typically compare embedding distributions of Y (Rauba et al., 2025; Raj et al., 2023). Therefore, nuisance-induced surface variation can be reflected in the sensitivity score, causing *instability*.

Decomposing the source of instability. According to the law of total variance, the variance of the sensitivity analysis is decomposed as follows:

$$\mathrm{Var}(\Psi(Y)) = \mathrm{Var}(\mathbb{E}[\Psi(Y) \mid M]) + \mathbb{E}[\mathrm{Var}(\Psi(Y) \mid M)]\,. \tag{2}$$

The first term represents variation arising from changes in the meaning of M, i.e., semantic variation. The second term represents variation that occurs when the meaning of M is fixed due to fluctuations in the output (e.g., decoding randomness or formatting differences), i.e., instability under fixed meaning. We further decompose the second term once more by applying the law of total variance conditionally on M as follows:

$$\mathrm{Var}(\Psi(Y)) = \underbrace{\mathrm{Var}(\mathbb{E}[\Psi(Y) \mid M])}_{\text{semantic}} + \underbrace{\mathbb{E}[\mathrm{Var}(\mathbb{E}[\Psi(Y) \mid M, N] \mid M)]}_{\text{nuisance-induced instability}} + \underbrace{\mathbb{E}[\mathrm{Var}(\Psi(Y) \mid M, N)]}_{\text{residual randomness}}\,. \tag{3}$$

The nuisance-induced variability captures variability in the conditional expectation of $\Psi(Y)$ induced by changes in nuisance factors N even when the meaning M is fixed. In contrast, the residual randomness represents irreducible randomness that remains even when both M and N are fixed, such as stochasticity in the generation process. This decomposition makes explicit that instability arises from variability induced by nuisance factors even when semantic content is fixed. Our goal is to reduce the second term (nuisance-induced variability), while preserving the first term (semantic variation).

2.2 Structural Limitation of Raw-Output Sensitivity

We show that the instability observed in distribution-based sensitivity analysis is not merely an implementation artifact but is inevitable under an output-generation structure where the meaning M and nuisance N are mixed (Errica et al., 2025). This shows that raw-output distributional sensitivity is generally not identifiable as semantic sensitivity without additional assumptions or nuisance-control mechanisms.

Let the meaning associated with the input x be $M(x)$, and let the nuisance be drawn from $N \sim p(n \mid x)$. Then, the output distribution of Y conditioned on x is given by:

$$p(Y \mid x) = \int p(Y \mid M(x), n)\, p(n \mid x)\, dn. \tag{4}$$

Sensitivity analysis compares the output distributions conditioned on two inputs x and x' via a divergence or distance metric satisfying $D(P, Q) > 0$ whenever $P \neq Q$ as:

$$D(p(Y \mid x),\, p(Y \mid x'))\,, \tag{5}$$

and interprets this as the effect of changing the inputs.

Theorem 1. *Suppose there exist two inputs x, x' such that $M(x) = M(x') =: m$. Assume that the nuisance variable N takes two values n_1, n_2, and define $P_i(\cdot) := p(\cdot \mid m, n_i)$ for $i \in \{1, 2\}$. Let $p(n_1 \mid x) = \alpha$ and $p(n_1 \mid x') = \beta$ with $\alpha \neq \beta$, so that $p(n_2 \mid x) = 1 - \alpha$ and $p(n_2 \mid x') = 1 - \beta$. Let D be a divergence or distance satisfying $D(P, Q) = 0$ if and only if $P = Q$. If $P_1 \neq P_2$, then $p(Y \mid x) \neq p(Y \mid x')$ and hence $D(p(Y \mid x), p(Y \mid x')) > 0$. In particular, even when the meaning is unchanged, output-distribution-based sensitivity can be nonzero purely due to nuisance shifts.*

Proof sketch. Since $M(x) = M(x') = m$, the two output distributions differ only through the nuisance mixing weights:

$$p(Y \mid x) = \alpha P_1 + (1 - \alpha) P_2, \qquad p(Y \mid x') = \beta P_1 + (1 - \beta) P_2.$$

Thus,

$$p(Y \mid x) - p(Y \mid x') = (\alpha - \beta)(P_1 - P_2).$$

Because $\alpha \neq \beta$ and $P_1 \neq P_2$, the right-hand side is nonzero as a signed measure, so the two mixture distributions are different. By the identity-of-indiscernibles property of D, their distance is strictly positive. Thus, even without semantic change, nuisance shifts alone can induce nonzero output-distribution sensitivity. Detailed proof is provided in Appendix C. □

It is worth noting that the two-point nuisance setting is used for analytical clarity; the same phenomenon extends to more general nuisance distributions under appropriate identifiability conditions.

This theorem shows that distribution-based sensitivity computed on raw outputs cannot, in general, disentangle semantic variation from nuisance variation. In particular, when the environment (e.g., decoding strategy, language, or stylistic factors) varies across inputs, differences in output distributions may arise even in the absence of semantic change (Acharyya et al., 2025). Consequently, raw-output distribution differences are not reliable indicators of meaning.

These observations motivate the use of nuisance-invariant representations for meaning-centered sensitivity analysis. In other words, meaning-centered sensitivity requires a representation that suppresses nuisance effects while preserving semantic content.

Proposition 1. *Suppose there exists a representation $S = g(Y)$ such that $S \perp N \mid M$. Then, for any two inputs x, x' satisfying $M(x) = M(x')$, we have $p(S \mid x) = p(S \mid x')$. Thus, the distribution of S is invariant to nuisance shifts, and distributional comparisons based on S depend only on semantic differences.*

Proof. By conditional independence,

$$p(S \mid x) = \int p(S \mid M(x), n)\, p(n \mid x)\, dn = \int p(S \mid M(x))\, p(n \mid x)\, dn = p(S \mid M(x)). \tag{6}$$

Hence, when $M(x) = M(x')$, the distributions $p(S \mid x)$ and $p(S \mid x')$ coincide regardless of differences in $p(n \mid x)$. □

Proposition 1 provides a sufficient condition under which sensitivity analysis becomes invariant to nuisance variation.

2.3 Information-Theoretic Formalization of Summarization as a Nuisance-Invariant Representation

We here formalize the nuisance-invariant representations through an information-theoretic objective, and interpret summarization as one practical approach toward this ideal. A summarization function g maps the raw output Y into a meaning-focused summary representation $S = g(Y)$. In general, the design goals of summarization can be generalized as (West et al., 2019):

1. **Meaning preservation:** S should retain as much of the core meaning M as possible.
2. **Nuisance suppression:** S should minimize the influence of nuisance N such as decoding randomness and style/language variation.

To formalize these summarization goals in information-theoretic terms, we introduce a CIB objective based on the information bottleneck (IB) principle (Tishby et al., 2000; Ju et al., 2021; Alemi et al., 2017). We define the ideal summary $S^* = g^*(Y)$ as the summary induced by the summarization function g^* that maximizes the CIB objective as

$$g^* = \arg\max_g \; I(g(Y); M) - \beta\, I(g(Y); N \mid M), \quad \beta > 0, \tag{7}$$

where $I(\cdot;\cdot)$ denotes mutual information and $I(\cdot;\cdot \mid \cdot)$ denotes conditional mutual information. In practice, M and N are not directly observable; thus, the CIB objective should be understood as a conceptual principle guiding the design of summarization, rather than a directly optimized objective.

In the CIB objective, the first term measures how much meaning-related information is preserved by the representation induced by g, where a larger value implies stronger meaning preservation. The second term quantifies how much nuisance-related information remains in the representation once the meaning M is fixed, where a smaller value implies stronger suppression of nuisance dependence. The hyperparameter $\beta > 0$ controls the relative importance of the two terms, thereby determining the trade-off between meaning preservation and nuisance suppression.

The CIB objective constitutes an information-theoretic optimization problem that encourages maximizing $I(g(Y); M)$ while minimizing $I(g(Y); N \mid M)$. In other words, it promotes representations that retain meaning-related information while suppressing nuisance-dependent information. Through this objective, the optimal summary $S^* = g^*(Y)$ can be interpreted as a stable representation designed to preserve meaning while minimizing dependence on nuisance factors.

Ideal baseline under summarization. The ideal summary $S^* = g^*(Y)$ obtained from the CIB objective is intended to provide a stable, meaning-centered representation that is invariant to nuisance variations such as decoding randomness, language, and style. To analyze this idealized setting, we adopt the following assumptions:

1. S^* is a deterministic function of Y.
2. S^* depends only on meaning M and is conditionally independent of nuisance N:

$$S^* \perp N \mid M \qquad \Longleftrightarrow \qquad I(S^*; N \mid M) = 0. \tag{8}$$

It is worth noting that this represents an idealized condition that may only be approximately satisfied in practice. Under this condition, the marginal distribution of S^* satisfies

$$p(s) = \iint p(s \mid m, n)\, p(m, n)\, dm\, dn = \int p(s \mid m)\, p(m)\, dm, \tag{9}$$

where we used $S^* \perp N \mid M$. Hence, $p(s)$ depends only on the meaning distribution $p(m)$ and is invariant to changes in the nuisance distribution $p(n)$. This implies that sensitivity computed through S^* is invariant to nuisance variation, effectively removing nuisance-induced instability.

We define a sensitivity measure as $\hat{\theta} := \Psi(Y)$, and the *idealized* variance-reduced baseline as $\bar{\theta} := \mathbb{E}[\Psi(Y) \mid S^*]$. This baseline represents the expected sensitivity over outputs Y that share the same ideal summary S^*, thereby averaging out variability not resolved by the nuisance-invariant representation. Since S^* is a function of Y, $\bar{\theta}$ is the Rao–Blackwellized version of $\hat{\theta}$ with respect to S^*, and therefore satisfies $\mathrm{Var}(\bar{\theta}) \leq \mathrm{Var}(\hat{\theta})$. The additional assumption $S^* \perp N \mid M$ gives this

variance reduction a nuisance-control interpretation: the conditioning variable preserves meaning while suppressing nuisance-dependent variation. This does not imply that the practical quantity $\Psi(S_{1:n})$ is itself a Rao–Blackwellized estimator. Rather, $\bar{\theta}$ provides an idealized variance-reduced reference against which practical summary-based sensitivity can be interpreted.

2.4 Practical Summary-Based Sensitivity and Approximation Error

The ideal baseline $\bar{\theta} = \mathbb{E}[\Psi(Y) \mid S^*]$ is not directly computable in practice, as it requires access to the ideal nuisance-invariant summary S^* as well as the conditional expectation over all outputs Y that map to the same summary. In realistic settings, neither the ideal representation S^* nor the conditional distribution $p(Y \mid S^*)$ is available.

Instead, we approximate this ideal procedure by computing sensitivity directly on summarized outputs. Let $S = g(Y)$ denote a practical summary, which we interpret as an approximation to the nuisance-suppressing representation motivated by the CIB objective. We then define the practical sensitivity measure

$$\tilde{\theta} := \Psi(S_{1:n}), \tag{10}$$

For theoretical analysis, we let $\Psi(S)$ denote the corresponding population-level quantity associated with the sample-based sensitivity measure $\Psi(S_{1:n})$. We define the approximation error relative to the ideal baseline as

$$\Delta := \Psi(S) - \mathbb{E}[\Psi(Y) \mid S^*], \tag{11}$$

which serves as an intractable reference quantity. Then, the following identity holds:

$$\Psi(S) = \mathbb{E}[\Psi(Y) \mid S^*] + \Delta. \tag{12}$$

By the variance decomposition,

$$\mathrm{Var}(\Psi(S)) = \mathrm{Var}(\mathbb{E}[\Psi(Y) \mid S^*]) + \mathrm{Var}(\Delta) + 2\,\mathrm{Cov}(\mathbb{E}[\Psi(Y) \mid S^*],\, \Delta)\,. \tag{13}$$

This decomposition shows that the variability of summary-based sensitivity arises from (i) the intrinsic variability of the ideal baseline, and (ii) the approximation error introduced by summarization, together with their interaction captured by the covariance term. The remaining question is how large this approximation error can be in practice.

Information-theoretic interpretation of approximation error. The magnitude of Δ depends on how much information required for $\Psi(Y)$ is retained by the summary. To relate the practical summary S to the ideal baseline, we decompose

$$\Delta = \big(\Psi(S) - \mathbb{E}[\Psi(Y) \mid S]\big) + \big(\mathbb{E}[\Psi(Y) \mid S] - \mathbb{E}[\Psi(Y) \mid S^*]\big).$$

The first term reflects how much information relevant to $\Psi(Y)$ is retained by the practical summary, while the second term captures the gap between the practical summary S and the ideal nuisance-invariant representation S^*. When this second term is sufficiently small, standard information–estimation arguments suggest that smaller conditional information loss,

$$I(\Psi(Y); Y \mid S),$$

leads to smaller approximation error Δ under suitable regularity conditions on Ψ. This suggests that the more faithfully the summary preserves the information relevant to the ideal baseline, the smaller the approximation error becomes. Consequently, the approximation-related contribution in (13) is reduced. Overall, summarization can be interpreted as a practical mechanism for mitigating nuisance-induced variability while preserving the underlying meaning.

3 Experiments and Results

This section examines three empirical implications of our analysis. First, nuisance perturbations can induce nonzero raw-output sensitivity. Second, summary-based semantic normalization reduces nuisance-induced sensitivity while preserving sensitivity to meaning-relevant changes. Third, the resulting estimates become more comparable across languages. Appendix D.3 presents supporting controls, representation-level analyses, robustness checks, and detailed multilingual results.

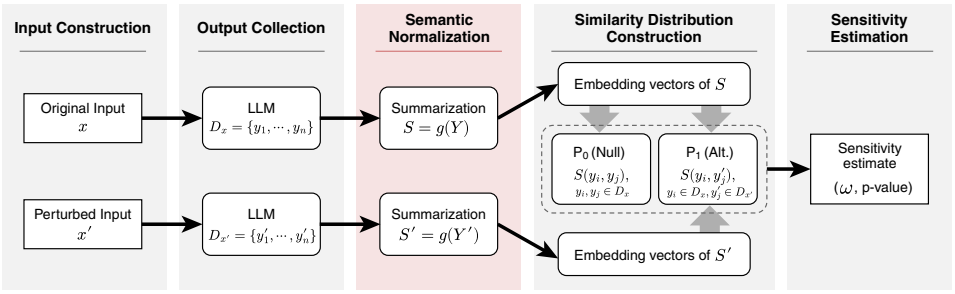


Figure 1: Overview of the summary-based semantic normalization framework. Outputs from base and perturbed inputs are summarized, embedded, and compared through similarity distributions to compute sensitivity ω. Raw-output sensitivity uses the same comparison pipeline with the semantic normalization step omitted.

Table 1: Raw-output and summary-based sensitivity ω under meaning-preserving nuisance perturbations. Raw-output sensitivity quantifies nuisance-induced distributional differences before normalization. Reduction is computed relative to the raw-output sensitivity score.

Nuisance factor	**Perturbation**	**Raw-output sensitivity (ω)**	**Summary-based sensitivity (ω)**	**Reduction (%)**	**Main action recall**
Format	Bullets	0.0842	0.0377	55.2	0.743
Format	Numbered	0.0797	0.0255	68.0	0.806
Length	Long	0.1119	0.0372	66.8	0.678
Length	Short	0.2899	0.1533	47.1	0.677
Tone	Casual	0.0717	0.0553	22.9	0.721
Tone	Formal	0.1141	0.0236	79.3	0.765
Tone	Friendly	0.0499	0.0320	35.9	0.810

3.1 Experimental Setup

Perturbation settings and response generation. We used LLMs as response generators for base and perturbed inputs, with Gemma-3 as the primary response-generation model. For each input condition, multiple outputs were sampled under the same decoding configuration so that output distributions could be compared under controlled generation settings. We evaluate the framework in three complementary experimental settings. The first uses meaning-preserving nuisance perturbations that vary tone, response length, or formatting while keeping the task intent fixed. The second uses a controlled refund-policy task that contrasts presentation-level nuisance perturbations with policy-relevant semantic changes. The third uses multilingual patient-data perturbations with controlled changes to clinically meaningful attributes while preserving the overall medical-advice context. Model configurations and implementation details are provided in Appendix D.2.

Sensitivity computation and summary-based normalization. Following prior distribution-based approaches to LLM sensitivity analysis (Errica et al., 2025; Rauba et al., 2025), generated outputs were embedded, and within-input (P_0) and cross-input (P_1) similarity distributions were constructed to compute the sensitivity measure ω. For each generated output Y, we applied a summarization function g to obtain $S = g(Y)$, which is intended to preserve task-relevant meaning while reducing surface-level variation. Raw and summarized representations were evaluated using the same response pools and sensitivity-computation settings, enabling a direct paired comparison. Figure 1 illustrates the overall procedure; additional details are provided in Appendix D.1.

3.2 Nuisance Confounding in Raw-Output Sensitivity

We first test whether meaning-preserving changes in tone, response length, or formatting induce nonzero raw-output sensitivity. The evaluation was conducted on four customer-service tasks involving refund responses, battery issue assistance, delivery delay notifications, and Wi-Fi troubleshooting. For each task, we defined a base input and perturbed inputs that differed only in surface-level presentation while keeping the task intent and required response content fixed. Using the same generation

Table 2: Semantic-vs-nuisance discrimination in the controlled refund-policy task. Nuisance and semantic ω denote the mean JSD-based sensitivity for the corresponding perturbation types. Results are averaged over MPNet, BGE-M3, and LaBSE.

Representation	Nuisance ω	Semantic ω	FPR	TPR	AUC
Raw-output	0.1751	0.1924	0.67	0.79	0.4531
Summary-based	0.0524	0.3558	0.42	0.92	0.8854

settings, we computed the sensitivity measure ω between the output distributions of the base and perturbed inputs. The resulting raw-output sensitivity therefore quantifies distributional differences induced by nuisance perturbations before semantic normalization.

Table 1 reports raw-output and summary-based sensitivity under meaning-preserving nuisance perturbations. The raw-output sensitivity scores are consistently nonzero across all nuisance conditions, indicating that distributional sensitivity is affected not only by changes in semantic content but also by nuisance shifts in surface form. Response-length perturbations induce the largest raw-output sensitivity, while formatting and tone perturbations also produce measurable shifts, showing that surface-level variation can substantially affect distributional comparisons even when task meaning is preserved. These results are consistent with the theoretical implication that raw-output distributional sensitivity is not, by itself, identifiable as semantic sensitivity under nuisance shifts. Therefore, raw-output sensitivity should not be interpreted directly as semantic sensitivity without an explicit mechanism for controlling nuisance-induced variation.

3.3 Summary-Based Normalization: Nuisance Suppression and Semantic Sensitivity

We next evaluate whether summary-based semantic normalization suppresses nuisance-induced sensitivity while preserving sensitivity to meaning-relevant changes. Using the same response pools and sensitivity estimator, we compute ω on the summarized outputs, enabling a direct comparison with raw-output sensitivity. As shown in Table 1, summary-based sensitivity is lower than raw-output sensitivity across all nuisance conditions, with an average reduction of 53.6%. Reductions occur across formatting, response-length, and tone perturbations, although their magnitude varies across conditions. Main-action recall ranges from 0.677 to 0.810 across nuisance conditions, indicating that the summaries retain a substantial portion of the task-relevant actions; its definition and extraction procedure are provided in Appendix D.3.2.

Reducing nuisance sensitivity is useful only if the normalized representation remains responsive to genuine semantic changes. We evaluate this complementary requirement using a controlled refund-policy task with eight nuisance perturbations and eight semantic perturbations. The task provides an explicit policy decision boundary: nuisance perturbations modify presentation-level properties while preserving the refund decision, whereas semantic perturbations modify policy-relevant conditions that can change the decision. A significant difference under a nuisance perturbation is counted as a false positive, while a significant difference under a semantic perturbation is counted as a true positive. AUC measures the separation between nuisance and semantic perturbations based on their sensitivity values.

As shown in Table 2, raw-output sensitivity provides little separation between nuisance and semantic perturbations, yielding an AUC of 0.4531. With summary-based representations, mean nuisance sensitivity decreases from 0.1751 to 0.0524, while mean semantic sensitivity increases from 0.1924 to 0.3558, raising the AUC to 0.8854. Thus, normalization not only suppresses sensitivity to presentation-level nuisance variation but also preserves a clear response to policy-relevant semantic changes. Together with the main-action recall results, this provides complementary evidence for nuisance suppression and semantic retention. Detailed perturbation definitions, additional controls, representation-level analyses, and estimator and encoder robustness checks are provided in Appendices D.3.1–D.3.4.

3.4 Cross-Lingual Robustness of Sensitivity Estimates

We examine whether summary-based semantic normalization improves the cross-lingual comparability of distribution-based sensitivity estimates. We use original and perturbed patient-data inputs with explicitly controlled semantic changes, modifying clinically meaningful attributes such as age, BMI, blood pressure, cholesterol level, smoking status, diabetes status, and family history. Because the same structured attribute changes can be expressed across languages while preserving the overall medical-advice context, this setting limits variation in the intended semantic perturbation itself. Differences in the resulting sensitivity estimates can therefore more directly reflect language-dependent nuisance factors such as expression, syntax, and representation quality. We applied the same input configuration to English, Korean, Chinese, French, and Indonesian and computed both raw-output and summary-based sensitivity using Gemma-3. For each language, we report the average over embedding models, the minimum–maximum range, and the cross-lingual standard deviation.

Figure 2 shows that raw-output sensitivity varies substantially across languages. This suggests that sensitivity estimates may be influenced by language-specific nuisance factors, such as differences in expression, syntax, and representation quality. In contrast, summary-based sensitivity estimates are more tightly concentrated across languages, as reflected by both the reduced range and the smaller cross-lingual standard deviation. Together, these results indicate that summary-based normalization reduces language-dependent nuisance variation while maintaining sensitivity to the controlled attribute changes, supporting more robust cross-lingual sensitivity analysis. A recommendation-retention audit further yields 85.9% pooled retention after normalization. Detailed language-specific sensitivity results and audit definitions are provided in Appendix D.3.5.

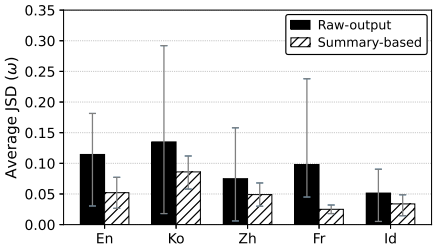


Figure 2: Raw-output and summary-based sensitivity ω across languages. Error bars indicate the minimum–maximum range.

4 Conclusion

We examined nuisance confounding in the context of distribution-based sensitivity analysis of LLM outputs. Our findings reveal that raw-output sensitivity can be nonzero even when the underlying meaning remains unchanged. We proposed summary-based semantic normalization as a practical nuisance-control mechanism, interpreting summarization as an approximate nuisance-invariant representation. Controlled perturbation experiments showed that summary-based normalization substantially reduces sensitivity induced by meaning-preserving nuisance changes while preserving responsiveness to semantic changes. Multilingual patient-data experiments further showed improved cross-lingual comparability of sensitivity estimates. Overall, these results support summary-based semantic normalization as a simple and interpretable step toward robust, meaning-centered sensitivity analysis of LLM output distributions.

Limitations and future work. Summary-based semantic normalization is an approximate nuisance-control mechanism and may remove task-relevant details, especially when summaries are overly compressed or the task requires fine-grained factual distinctions. Although main-action recall and the controlled semantic-vs-nuisance discrimination results provide evidence that task-relevant information is retained, they do not fully characterize semantic preservation across tasks. Future work should develop more direct measures of semantic preservation, such as factual consistency, hallucination analysis, and task-specific coverage, and systematically compare summarization strategies, summary lengths, and compression ratios.

The theoretical formulation also relies on an idealized meaning–nuisance decomposition and a nuisance-invariant representation condition, although meaning and nuisance may not be directly observable or cleanly separable in practice. Thus, the proposed framework should be viewed as a principled approximation rather than as a guarantee of fully semantic sensitivity estimation. Future work could explore alternative nuisance-invariant representations and more explicit nuisance-control mechanisms, and evaluate them across broader domains, higher-stakes decision tasks, and diverse perturbation types.

AI USE STATEMENT

In this work, generative AI tools were used to assist with language editing and manuscript organization, and to provide limited assistance with research ideation, experimental planning, and drafting and debugging analysis code. All AI-assisted text, code, and analyses included in this work were reviewed by the authors before inclusion.

The authors made all final decisions regarding the research questions, theoretical claims, methodology, experimental design, analysis, and conclusions, and take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The theoretical assumptions and proofs are described in Section 2 and Appendix C. The experimental procedure, model and decoding configurations, sensitivity estimation procedure, and additional robustness analyses are documented in Appendix D. Prompt templates used to construct controlled perturbations are provided in Appendix E.

REFERENCES

Aranyak Acharyya, Carey E Priebe, and Hayden S Helm. Testing for LLM response differences: the case of a composite null consisting of semantically irrelevant query perturbations. *arXiv:2509.10963*, 2025.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep Variational Information Bottleneck. In *International Conference on Learning Representations*, 2017.

Henning Bartsch, Ole Jorgensen, Domenic Rosati, Jason Hoelscher-Obermaier, and Jacob Pfau. Self-Consistency of Large Language Models under Ambiguity. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi (eds.), *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 89–105, Singapore, December 2023. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-Shot Learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Federico Errica, Davide Sanvito, Giuseppe Siracusano, and Roberto Bifulco. What Did I Do Wrong? Quantifying LLMs' Sensitivity and Consistency to Prompt Engineering. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

Gyandev Gupta, Bashir Rastegarpanah, Amalendu Iyer, Joshua Rubin, and Krishnaram Kenthapadi. Measuring distributional shifts in text: the advantage of language model-based embeddings. *arXiv:2312.02337*, 2023.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*, 2020.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

J Ju, M Liu, HY Koh, Y Jin, L Du, and S Pan. Leveraging Information Bottleneck for Scientific Document Summarization. In *The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*. Association for Computational Linguistics, 2021.

Nadav Kunievsky and James A Evans. Measuring (a sufficient) world model in LLMs: A variance decomposition framework. *arXiv:2506.16584*, 2025.

Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. Language-Specific Latent Process Hinders Cross-Lingual Performance. *arXiv:2505.13141*, 2025.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. In *Socially Responsible Language Modelling Research*, 2023.

Manikanta Loya, Divya Sinha, and Richard Futrell. Exploring the Sensitivity of LLMs' Decision-Making Capabilities: Insights from Prompt Variations and Hyperparameters. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3711–3716, Singapore, December 2023. Association for Computational Linguistics.

Nguyen Nhat Minh, Andrew Baker, Clement Neo, Allen G Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs. In *The Thirteenth International Conference on Learning Representations*, 2025.

Harsh Raj, Vipul Gupta, Domenic Rosati, and Subhabrata Majumdar. Semantic consistency for assuring reliability of large language models. *arXiv:2308.09138*, 2023.

Paulius Rauba, Qiyao Wei, and Mihaela Van Der Schaar. Statistical Hypothesis Testing for Auditing Robustness in Language Models. In *International Conference on Machine Learning*, pp. 51297–51313. PMLR, 2025.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual LLMs think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 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 (ICLR)*, 2024.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. *arXiv physics/0004057*, 2000.

Julian Junyan Wang and Victor Xiaoqi Wang. Assessing Consistency and Reproducibility in the Outputs of Large Language Models: Evidence Across Diverse Finance and Accounting Tasks. *arXiv:2503.16974*, 2025.

Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig, Ilia Kulikov, and Zaid Harchaoui. From Decoding to Meta-Generation: Inference-time Algorithms for Large Language Models. *Transactions on Machine Learning Research*, 2024.

Peter West, Ari Holtzman, Jan Buys, and Yejin Choi. BottleSum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3752–3761, Hong Kong, China, November 2019. Association for Computational Linguistics.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. A survey on multilingual Large Language Models: Corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362, 2025.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7915–7927, 2023.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. Beyond english-centric LLMs: What language do multilingual language models think in? *arXiv:2408.10811*, 2024.

A Broader Impacts

This work may support a more reliable analysis of LLM behavior by helping distinguish semantic changes from nuisance-induced variation, especially in multilingual settings. By making sensitivity estimates more interpretable, the proposed framework can help researchers and practitioners better diagnose when output differences reflect meaningful input changes rather than incidental variation from language, style, formatting, or stochastic generation. However, summary-based sensitivity estimates may be misinterpreted as fully semantic measures if their limitations are ignored. Because summarization is only an approximate nuisance-control mechanism, it may eliminate task-relevant details or behave inconsistently across languages and domains. In high-stakes applications, misleading robustness conclusions could affect downstream decisions. Therefore, the proposed analysis should be used as a diagnostic tool rather than as a standalone basis for deployment or decision-making. In high-stakes settings, summary-based sensitivity analysis should be accompanied by task-specific checks for semantic preservation, factual consistency, and domain validity.

B Related Work

B.1 Generative Structure of LLMs and Output Variability

LLMs can produce different outputs even for identical or highly similar inputs. Such output variability has been understood as a structural characteristic arising from the generative mechanism of LLMs. Prior work has shown that LLM outputs can vary not only with input meaning but also with stochastic choices made during the generation process (Loya et al., 2023; Bartsch et al., 2023).

LLMs generate tokens sequentially according to a conditional probability distribution, and different output expressions can be produced for the same input depending on the sampling strategy or stochastic choices made during decoding (Holtzman et al., 2020). While this generative structure enables diversity and expressive flexibility, it can also introduce additional variability in terms of output stability. Prior work has pointed out that such probabilistic generation characteristics can affect variation in the surface forms of generated outputs (Wang & Wang, 2025; Minh et al., 2025).

This issue becomes even more pronounced in multilingual settings. Due to differences in linguistic structure across languages and imbalances in training data, prior work has shown that output stability tends to be lower in non-English languages (Xu et al., 2025; Huang et al., 2023; Lim et al., 2025). It has also been noted that multilingual LLMs often rely on internal representations disproportionately shaped by English, and that cross-lingual semantic alignment remains imperfect even in embedding spaces, such that embedding models trained primarily on English may further amplify representational discrepancies in non-English texts (Zhong et al., 2024; Schut et al., 2025). As a result, expressive variability can be amplified across languages even among outputs that convey the same meaning.

Overall, variability in LLM outputs arises from the interaction of multiple factors, including probabilistic generation structure, decoding strategies, linguistic properties, and embedding bias. Prior work has suggested that such variability may constitute a structural limitation.

B.2 Distribution-Based Sensitivity Analysis

To quantitatively evaluate the sensitivity of LLM outputs to input changes, prior work has proposed approaches that analyze output variation from a distributional perspective (Errica et al., 2025). A representative approach computes sentence embeddings for sets of outputs generated before and after input perturbations and measures the effect of input changes by calculating distances or divergences between the resulting embedding distributions (Rauba et al., 2025).

This approach has the advantage of capturing output variability more comprehensively by considering distributional changes over the entire set of outputs rather than differences between individual outputs. However, what is actually being compared is not the semantic distribution of outputs themselves, but the embedding distribution induced by the selected embedding model. Accordingly, sensitivity analysis results can be influenced not only by changes in output meaning but also by the structural properties of the embedding space and surface-level variation in expression (Gupta et al., 2023).

Even for the same set of outputs, different embedding models may yield different sensitivity estimates. This instability becomes particularly pronounced in multilingual settings because of differences in the language suitability of embedding models (Lim et al., 2025). Consequently, existing distribution-based sensitivity analysis is limited in its ability to clearly disentangle semantic differences induced by input changes from nuisance-induced variability arising during the output generation process.

B.3 Effects of Summarization and an Information-Theoretic Perspective

Summarization generally condenses long texts into a concise form, but from an information-theoretic perspective, it also functions as a representational transformation that selectively preserves semantic information while removing unnecessary variation (Tishby et al., 2000). Summarization retains the core meaning of the original text while removing redundant or non-essential information.

According to Information Bottleneck (IB) theory, an effective representation retains information relevant to the target variable while minimizing information associated with unnecessary sources of variation or noise (Ju et al., 2021). From this perspective, summarization can be interpreted as a normalized representational transformation that preserves semantic information while suppressing non-semantic sources of variation.

Prior work has shown that summarization can help stabilize document-level semantic representations and mitigate unnecessary variability (Ju et al., 2021; West et al., 2019). However, these properties have received relatively limited attention in the context of stabilizing distribution-based sensitivity analysis of LLM outputs.

C Proof of Theorem 1

Since $M(x) = M(x') = m$, the conditional output distributions for the two inputs differ only through the nuisance mixing distributions. Under the two-point nuisance assumption,

$$p(Y \mid x) = \sum_{i=1}^{2} p(Y \mid m, n_i)\, p(n_i \mid x) \tag{14}$$

$$= \alpha P_1 + (1-\alpha) P_2, \tag{15}$$

and similarly,

$$p(Y \mid x') = \sum_{i=1}^{2} p(Y \mid m, n_i)\, p(n_i \mid x') \tag{16}$$

$$= \beta P_1 + (1-\beta) P_2. \tag{17}$$

Subtracting the two expressions yields

$$p(Y \mid x) - p(Y \mid x') = (\alpha - \beta)(P_1 - P_2). \tag{18}$$

Since $P_1 \neq P_2$, there exists a measurable set A such that $P_1(A) \neq P_2(A)$. Evaluating (18) on this set gives

$$p(A \mid x) - p(A \mid x') = (\alpha - \beta)\big(P_1(A) - P_2(A)\big).$$

Because $\alpha \neq \beta$ and $P_1(A) \neq P_2(A)$, this quantity is nonzero. Hence,

$$p(Y \mid x) \neq p(Y \mid x').$$

Since $D(P, Q) = 0$ if and only if $P = Q$, we conclude that

$$D(p(Y \mid x),\, p(Y \mid x')) > 0.$$

Thus, even when $M(x) = M(x')$, a change in nuisance mixing weights alone can induce a nonzero difference between the output distributions.

D SUPPLEMENTARY EXPERIMENTAL DETAILS AND RESULTS

D.1 DETAILED PROCEDURE FOR SUMMARY-BASED SEMANTIC NORMALIZATION

We provide a detailed procedure for the summary-based semantic normalization framework used in our experiments. For each base or perturbed input, we sample LLM outputs, optionally apply semantic normalization through summarization, construct similarity distributions in embedding space, and compute the sensitivity measure ω. Raw-output sensitivity is computed using the same distribution-based comparison procedure, with the semantic normalization step omitted.

1) Input Construction For each experiment, we constructed original and perturbed inputs according to the comparison objective. Depending on the experimental objective, we categorized input perturbations into those intended to induce only nuisance variation and those reflecting meaningful attribute changes. The former altered only the form of expression while preserving meaning, such as changes in tone, response length, and formatting style, whereas the latter introduced changes in input attributes that could be interpreted as actual semantic changes. This design allowed us to separately evaluate whether summarization mitigates the effects of nuisance-induced variability and whether it preserves sensitivity to semantic differences.

2) Output Collection For each input x, LLM outputs were collected under the same decoding configuration. Let the set of outputs generated for input x be denoted by

$$D_x = \{y_1, \ldots, y_n\},$$

where n denotes the number of sampled outputs per input; in our experiments, we set $n = 10$. For the original and perturbed inputs, output samples were generated independently. This design preserves the stochastic variability arising from the generation process while enabling comparison of output distributions across input changes.

3) Summary-based Semantic Normalization Each collected output $Y \in D_x$ was transformed into a summary $S = g(Y)$ using a summarization model and a prompt appropriate for the target language. In all experiments, we compared raw outputs and summarized outputs side by side to analyze how summarization affects the structure of output distributions and the resulting sensitivity analysis.

4) Similarity Distribution Construction For both raw outputs Y and summarized outputs S, we computed sentence embeddings and constructed empirical distributions based on a rescaled cosine similarity score defined as

$$s(y_i, y_j) := \frac{1 + \cos(e(y_i), e(y_j))}{2},$$

where $e(\cdot)$ denotes the embedding function. This rescales cosine similarity values to the interval $[0, 1]$, with larger values indicating greater similarity in the embedding space.

We define the within-input distribution as the null distribution P_0 and the cross-input distribution as the alternative distribution P_1 as follows:

$$P_0 = \{s(y_i, y_j) : y_i, y_j \in D_x,\ i \neq j\} \text{ and } P_1 = \{s(y_i, y'_j) : y_i \in D_x,\ y'_j \in D_{x'}\}.$$

Here, D_x denotes the set of output samples generated from the original input x, and $D_{x'}$ denotes the set of output samples generated from the perturbed input x'.

We discretized the similarity-score values over the interval $[0, 1]$ using histogram binning, and the bin frequencies were normalized to form probability mass functions. This allowed P_0 and P_1 to be represented as discrete probability distributions suitable for Jensen–Shannon divergence computation.

To reduce bias arising from differences in sample size between P_0 and P_1, we randomly subsampled values from P_1 so that it contained the same number of samples as P_0. This was done to minimize the influence of sample size differences on divergence estimation.

5) Sensitivity Estimation The difference between the two distributions was quantified using Jensen–Shannon Divergence (JSD), which we define as the sensitivity effect size ω:

$$\omega := \mathrm{JSD}(P_0 \,\|\, P_1).$$

JSD is defined as

$$\mathrm{JSD}(P_0 \,\|\, P_1) = \frac{1}{2}\mathrm{KL}(P_0 \,\|\, M) + \frac{1}{2}\mathrm{KL}(P_1 \,\|\, M), \qquad M = \frac{1}{2}(P_0 + P_1).$$

JSD approaches 0 when the two distributions are similar, and larger values indicate greater differences between output distributions induced by the input change.

To assess the statistical significance of the observed sensitivity measure ω, we performed a permutation test and computed a p-value. The permutation test repeatedly shuffled the output samples used to construct the null and alternative distributions, recomputed the corresponding divergence values, and evaluated whether the observed ω could be attributed to the input perturbation.

The resulting p-value represents the probability of observing a distributional difference of at least the same magnitude under the null hypothesis of no input change; smaller values indicate that the input perturbation had a statistically significant effect on the outputs.

D.2 Implementation Details

Table 3 summarizes the main implementation settings used in our experiments. All models were used off-the-shelf without additional fine-tuning. For each experimental condition, the same response generation, embedding, and statistical testing settings were used for the original and perturbed inputs, enabling controlled comparisons between raw-output and summary-based sensitivity.

The nuisance experiments modified surface-level factors such as tone, response length, and output format, whereas the multilingual patient-data experiments modified structured patient attributes while keeping the overall task context fixed. For summary-based semantic normalization, each generated output was converted into a concise meaning-focused summary before embedding, with the goal of preserving task-relevant meaning while reducing tone, verbosity, formatting, and other surface-level variation.

D.3 Additional Experimental Results

D.3.1 Controlled Semantic-vs-Nuisance Analysis and Baseline Controls

The semantic-vs-nuisance experiment in Section 3.3 uses a controlled refund-policy task with an explicit decision boundary. The base case describes a customer who purchased a defective product within the refund window and has proof of purchase, making the request eligible for a refund. We constructed eight nuisance perturbations and eight semantic perturbations. The nuisance perturbations modify tone, response length, formatting, or response structure while preserving the underlying refund decision. The semantic perturbations modify policy-relevant conditions, including the return window, proof-of-purchase availability, product defect status, customer-caused damage, opened-item eligibility, and final-sale restrictions.

For each condition, we generated output samples from the base and perturbed inputs and computed sensitivity using the same distributional comparison pipeline. Nuisance perturbations were treated as control conditions and semantic perturbations as target conditions. A statistically significant difference under a nuisance perturbation was counted as a false positive, whereas one under a semantic perturbation was counted as a true positive. Raw, Summary, and baseline representations were evaluated using the same stored response pools, embedding models, histogram configuration, and cross-pair subsampling indices.

To distinguish the effect of semantic normalization from simple response compression or formatting cleanup, we evaluated three non-generative controls. *Prefix Truncation* retains only the initial tokens of each Raw response. *Extractive Length-Matched* selects salient and non-redundant textual units from the Raw response using a TF–IDF-based maximal marginal relevance criterion. *Format-Only* removes Markdown headings, list markers, emphasis markers, line breaks, and redundant whitespace without rewriting the response content.

For both length-matched controls, the target token budget was determined separately for each response using the length of its paired Summary. Only the Summary length was used; its textual content was not provided to either control. Both controls exactly matched the paired Summary length under the Gemma tokenizer.

Table 3: Key implementation settings used in the experiments.

Component	Parameter	Value
Response generator	Primary LLM	`google/gemma-3-4b-it`

	Supplementary LLMs	`meta-llama/Llama-3.2-1B`
		`LGAI-EXAONE/EXAONE-3.5-2.4B-Instruct`
		`deepseek-ai/deepseek-llm-7b-chat`
		`microsoft/Phi-3-mini-4k-instruct`
		`HuggingFaceTB/SmolLM2-135M-Instruct`
		`NousResearch/Nous-Hermes-2-Mistral-7B-DPO`

Response generation	Samples per input	$n = 10$
	Decoding	`do_sample=True`; temperature $= 0.7$; top-$p = 0.95$
	Length limit	max $= 500$ new tokens
Summary normalization	Gemma-based	Gemma-3 re-summarization for Gemma-3 nuisance experiments
	Seq2seq summarizers	En: `google/pegasus-xsum` Ko: `kobart-summary-v3` Fr: `t5-base-fr-sum-cnndm` Zh: `t5_summary_en_ru_zh_base_2048` Id: `cahya/t5-base-indonesian-summarization-cased`
Summary generation	Gemma-based decoding	`do_sample=False`
	Seq2seq decoding	Deterministic beam-search decoding
	Length limit	min $= 50$, max $= 150$ tokens

Embedding models	Primary sentence encoders	`sentence-transformers/paraphrase-multilingual-mpnet-base-v2`
		`intfloat/multilingual-e5-large`
		`BAAI/bge-m3`
		`sentence-transformers/LaBSE`

Statistical testing	Effect size	JSD ω
	Permutation test	10,000 resamples
Compute resources	Software environment	Python 3.12.6; PyTorch 2.7.0+cu126; Transformers 5.7.0.dev0; Sentence-Transformers 5.4.1
	Hardware	1× NVIDIA GeForce RTX 3090 GPU (CUDA 12.6); 8 physical CPU cores; 32 GB RAM

As shown in Table 4, Prefix Truncation does not reduce nuisance sensitivity and decreases semantic sensitivity, resulting in an AUC of 0.3229. Format-Only produces only a small change relative to Raw, with an AUC of 0.5052. Extractive Length-Matched reduces nuisance sensitivity and improves the AUC to 0.6771, indicating that content-aware compression accounts for part of the improvement. Summary representations achieve the lowest nuisance sensitivity, the highest semantic sensitivity, and the highest AUC of 0.8854. These results indicate that the observed improvement cannot be fully explained by response-length reduction or formatting cleanup alone.

D.3.2 Nuisance Recoverability and Semantic Retention

Main-action recall. To assess whether normalization retains task-relevant information, we measure main-action recall. For each task, actions are extracted from the Raw and Summary outputs using a predefined set of action labels. Recall is computed as

$$\mathrm{MainActionRecall} = \frac{|A_{\mathrm{Raw}} \cap A_{\mathrm{Summary}}|}{|A_{\mathrm{Raw}}|}, \tag{19}$$

Table 4: Comparison with compression and formatting controls in the controlled refund-policy task. Nuisance and semantic ω denote the mean JSD-based sensitivity for the corresponding perturbation types. AUC measures discrimination between semantic and nuisance perturbations. Results are averaged over MPNet, BGE-M3, and LaBSE.

Representation	**Nuisance ω**	**Semantic ω**	**AUC**
Raw	0.1751	0.1924	0.4531
Prefix Truncation	0.1785	0.1144	0.3229
Extractive Length-Matched	0.0837	0.1549	0.6771
Format-Only	0.1613	0.1994	0.5052
Summary	**0.0524**	**0.3558**	**0.8854**

Table 5: Linear-probe macro-F1 on Raw and Summary representations, averaged over MPNet, BGE-M3, and LaBSE.

Probe target	**Raw**	**Summary**
Tone nuisance	0.910	0.385
Format nuisance	0.986	0.475
Length nuisance	0.929	0.557
Decision semantics	0.883	0.900
Fine semantic state	0.902	0.899

where A_{Raw} and A_{Summary} denote the action sets extracted from the Raw and Summary outputs, respectively. The condition-level recall values are reported in Table 1.

Representation-level probing. To examine whether normalization selectively suppresses nuisance-related information, we conducted a representation-level probing analysis on a fully crossed refund-policy dataset with nine semantic states and eight nuisance states. For each combination, we generated ten responses and their corresponding Summaries.

We trained linear probes on frozen MPNet, BGE-M3, and LaBSE embeddings. Nuisance probes predicted tone, format, and length using leave-one-semantic-state-out evaluation, while semantic probes predicted three-way refund decisions and nine fine-grained policy states using leave-one-nuisance-state-out evaluation. We report macro-F1 averaged across held-out folds and the three encoders.

As shown in Table 5, nuisance-label recoverability decreases substantially after normalization, whereas recoverability of the evaluated decision and fine-grained semantic labels remains nearly unchanged. These results support the interpretation that normalization suppresses nuisance-related information while retaining the evaluated task-semantic structure.

D.3.3 Robustness of the Sensitivity Estimator

We evaluate the stability of semantic-vs-nuisance discrimination with respect to the number of generated responses and the histogram resolution. For the sample-size analysis, we constructed pools of 50 responses per input condition and repeatedly subsampled $n = 10$, 20, and 30 responses, using 200 repetitions for each setting. Raw and Summary comparisons shared generation indices, cross-pair subsampling indices, and permutation seeds. For the histogram-resolution analysis, we recomputed sensitivity on the same stored output samples using 10, 20, 30, and 50 bins.

As shown in Table 6, aggregate discrimination remains stable across generation sample sizes: Raw AUC ranges from 0.469 to 0.480, while Summary AUC ranges from 0.891 to 0.898. The separation between nuisance and semantic perturbations is likewise preserved across histogram resolutions, although the exact AUC values vary with the number of bins.

Table 6: Robustness of semantic-vs-nuisance discrimination to generation sample size and histogram resolution. Results for $n = 10$, 20, and 30 report mean AUC $\pm$ standard deviation over 200 subsamples from a 50-response pool; $n = 50$ reports the estimate obtained from the full response pool.

Robustness check	Setting	Raw AUC	Summary AUC
Sample size	$n = 10$	0.476 ± 0.038	0.896 ± 0.027
	$n = 20$	0.480 ± 0.025	0.898 ± 0.017
	$n = 30$	0.479 ± 0.017	0.896 ± 0.012
	$n = 50$	0.469	0.891
Histogram bins	10	0.385	0.862
	20	0.453	0.885
	30	0.427	0.865
	50	0.427	0.870

Table 7: Nuisance sensitivity across embedding encoders under the paired evaluation protocol. Results are averaged over four tasks and seven nuisance conditions. Reduction denotes Raw minus Summary sensitivity. "Lower cells" denotes the number of task–condition cells in which Summary sensitivity was lower than Raw sensitivity.

Encoder	Raw ω	Summary ω	Reduction	Lower cells
MPNet	0.1307	0.0677	0.0630	17/28
E5	0.0320	0.0248	0.0071	6/28
BGE-M3	0.1179	0.0496	0.0683	17/28
LaBSE	0.2071	0.0792	0.1279	22/28
Qwen3-Embedding-0.6B	0.1721	0.0595	0.1126	25/28
Gemini Embedding 2	0.1820	0.0485	0.1335	23/28

Individual significance decisions become increasingly stable as the sample size increases. Relative to the $n = 50$ reference, overall decision agreement increases from 0.765 to 0.941 for Raw and from 0.727 to 0.925 for Summary as the sample size increases from $n = 10$ to $n = 30$. We further address dependence among pairwise similarities by reassigning generated outputs, rather than individual similarity values, under the permutation null. Under this output-level reassignment, 87.5–95.8% of the significance decisions obtained with the scalar-level procedure are retained across the evaluated datasets and representations. Together, these results show that the aggregate semantic-vs-nuisance discrimination pattern is stable across the evaluated estimator configurations.

D.3.4 Robustness across Encoders, Generations, and Sensitivity Regimes

We examine the stability of the observed sensitivity patterns across embedding encoders, repeated response generations, and different baseline-sensitivity regimes.

Embedding-encoder robustness. We evaluated nuisance sensitivity using six embedding encoders across the same four tasks and seven nuisance conditions, yielding 28 task–condition cells per encoder. The evaluated encoders are MPNet, E5, BGE-M3, LaBSE, Qwen3-Embedding-0.6B, and Gemini Embedding 2. For each task–condition cell, the same 45 cross-pair positions were used across Raw and Summary representations and all six encoders.

As shown in Table 7, mean nuisance sensitivity is lower for Summary representations under all six encoders. The magnitude and cell-level consistency of the reduction vary across embedding spaces, with the smallest aggregate reduction observed for E5, where Summary sensitivity is lower in 6 of 28 task–condition cells. Across the remaining encoders, the reduction is larger and occurs in 17–25 of the

Table 8: Raw-output and summary-based sensitivity across LLMs and embedding models. Min–max denotes the range of ω observed over 10 repeated runs, and $P(p < 0.05)$ denotes the proportion of statistically significant results.

Gemma-3				
Embedding model	**Raw-output**		**Summary-based**	
	avg ω (min–max)	$P(p < 0.05)$	avg ω (min–max)	$P(p < 0.05)$
MPNet	0.2618 (0.1749–0.3687)	0.22	0.0595 (0.0386–0.0766)	0.11
E5	0.0193 (0.0151–0.0258)	0.22	0.0480 (0.0334–0.0663)	0.00
BGE-M3	0.0251 (0.0152–0.0354)	0.11	0.0536 (0.0418–0.0632)	0.11
LaBSE	0.0381 (0.0289–0.0583)	0.11	0.0597 (0.0464–0.0770)	0.22
EXAONE-3.5				
Embedding model	**Raw-output**		**Summary-based**	
	avg ω (min–max)	$P(p < 0.05)$	avg ω (min–max)	$P(p < 0.05)$
MPNet	0.0751 (0.0483–0.1157)	0.11	0.0529 (0.0391–0.0776)	0.11
E5	0.0143 (0.0090–0.0235)	0.22	0.0282 (0.0201–0.0418)	0.11
BGE-M3	0.0199 (0.0126–0.0309)	0.22	0.0454 (0.0326–0.0635)	0.22
LaBSE	0.0179 (0.0140–0.0233)	0.00	0.0234 (0.0136–0.0291)	0.22

28 cells. Overall, the aggregate nuisance-reduction pattern extends across the evaluated embedding representations, while its magnitude varies across encoders.

Repeated-generation stability. We evaluate the stability of sensitivity estimates across embedding models and repeated generations over ten runs using fixed patient-data input configurations for Gemma-3 in English and EXAONE-3.5 in Korean. We use MPNet, E5, BGE-M3, and LaBSE for both models. Within each run, $P(p < 0.05)$ denotes the fraction of baseline–perturbation comparisons yielding a permutation-test p-value below 0.05.

Raw sensitivity varies substantially across embedding models, particularly in configurations with larger sensitivity values. Summary sensitivity is more tightly concentrated across embedding models, while the direction and magnitude of within-encoder changes vary across configurations. The repeated-generation results therefore support improved cross-embedding comparability after normalization.

Low-Raw sensitivity regimes. As a post-hoc diagnostic, we examine whether nuisance reduction is also observed when the initial Raw sensitivity is relatively low. We focus on the four nuisance conditions with the lowest aggregate Raw sensitivity in the nuisance analysis: friendly tone, casual tone, numbered formatting, and bullet formatting. This analysis uses MPNet, E5, BGE-M3, and LaBSE, yielding 64 task–condition–encoder cells.

For each task and condition, we generated two independent output pools, U_1 and U_2, each containing 30 responses from the same unperturbed input, and a separate 30-response pool N from the nuisance-perturbed input. We define same-input-adjusted nuisance sensitivity as

$$\Delta\omega = \omega(U_1, N) - \omega(U_1, U_2), \tag{20}$$

where $\omega(U_1, U_2)$ captures the distributional difference between independently generated outputs from the same input. The reduction associated with normalization is then measured as

$$R = \Delta\omega_{\mathrm{Raw}} - \Delta\omega_{\mathrm{Summary}}. \tag{21}$$

Generation-resampling indices and cross-pair positions were shared across Raw and Summary representations and across encoders. Confidence intervals were computed using 500 paired bootstrap replicates. Across the four low-Raw conditions, same-input-adjusted nuisance sensitivity decreases by 0.0343 on average, with a 95% paired bootstrap confidence interval of $[0.0257, 0.0445]$. Reductions occur in 49 of the 64 task–condition–encoder cells. These results indicate that the aggregate nuisance-reduction pattern is also observed in conditions with relatively low initial Raw sensitivity. The effect magnitude varies across encoders and cells, with E5 showing an aggregate reduction close to zero.

Table 9: Embedding-level raw-output and summary-based sensitivity ω in English settings.

Embedding	Llama-3.2		SmolLM2		Phi-3		Gemma-3	
	Raw	Summary	Raw	Summary	Raw	Summary	Raw	Summary
MPNet	0.0384	0.0353	0.0516	0.0288	0.0541	0.0422	0.1473	0.0583
E5	0.0119	0.0173	0.0239	0.0197	0.0190	0.0297	0.0304	0.0266
MiniLM-L6	0.0215	0.0304	0.0300	0.0345	0.0293	0.0328	0.0334	0.0456
DistilBERT	0.0208	0.0233	0.0358	0.0289	0.0360	0.0392	0.0404	0.0412
BGE-M3	0.0450	0.0219	0.0110	0.0211	0.0154	0.0142	0.0988	0.0462
LaBSE	0.0362	0.0310	0.0629	0.0254	0.0161	0.0176	0.1814	0.0772

Table 10: Embedding-level raw-output and summary-based sensitivity ω in non-English settings.

Embedding	Korean				Chinese				French				Indonesian	
	EXAONE-3.5		Gemma-3		DeepSeek		Gemma-3		Mistral		Gemma-3		Gemma-3	
	Raw	Summary	Raw	Summary	Raw	Summary	Raw	Summary	Raw	Summary	Raw	Summary	Raw	Summary
MPNet	0.1717	0.0602	0.3646	0.1259	0.0605	0.0393	0.1961	0.0507	0.3466	0.0583	0.3081	0.0319	0.0413	0.0328
E5	0.0188	0.0436	0.0166	0.0601	0.0166	0.0326	0.0096	0.0268	0.0244	0.0279	0.0201	0.0172	0.0052	0.0145
BGE-M3	–	–	–	–	0.0370	0.0326	0.0435	0.0448	0.0318	0.0341	0.0272	0.0273	0.0904	0.0390
BGE-Korea	0.0237	0.0541	0.0183	0.0819	–	–	–	–	–	–	–	–	–	–
SBERT	0.1515	0.0511	0.1612	0.0741	–	–	–	–	–	–	–	–	–	–
Text2Vec	–	–	–	–	0.0555	0.0605	0.0532	0.0719	–	–	–	–	–	–
LaBSE	–	–	–	–	–	–	–	–	0.0156	0.0411	0.0299	0.0204	0.0691	0.0487

D.3.5 Detailed Multilingual Results and Recommendation Retention

We report language- and model-specific sensitivity estimates underlying the cross-lingual averages presented in Section 3.4. Dashes in the tables indicate unavailable or unused LLM–embedding combinations.

Across the English configurations, sensitivity estimates are relatively tightly clustered across LLM–embedding model combinations. In several settings, particularly those based on E5, MiniLM-L6, and DistilBERT, Summary sensitivity remains close to Raw sensitivity or increases slightly. Accordingly, the effect of normalization is smaller in configurations where the Raw sensitivity estimates are already relatively concentrated.

Compared with English, sensitivity estimates in the non-English settings are more dispersed across languages, LLMs, and embedding models. This pattern is particularly pronounced for MPNet in Korean and French, where Raw sensitivity varies substantially across response-generation models. Summary representations substantially reduce sensitivity in several high-variability settings, while changes are smaller or mixed in configurations with low Raw sensitivity. Overall, these detailed results are consistent with the improved cross-lingual comparability observed in the main analysis.

Recommendation-retention audit. The content-retention analysis in Section 3.4 uses nine structured cardiovascular patient profiles and five languages. For each profile and language, ten Raw responses were paired with their corresponding same-language Summaries, yielding 90 pairs per language and 450 pairs overall.

The deterministic pattern-based evaluator covers eight shared recommendation categories: smoking cessation, exercise, diet, weight management, blood-pressure monitoring, cholesterol management, medical follow-up, and medication. For pair i, let R_i^{Raw} and R_i^{Summary} denote the category sets detected in the Raw response and Summary. Pair-level retention is defined as

$$r_i = \frac{\left| R_i^{\mathrm{Raw}} \cap R_i^{\mathrm{Summary}} \right|}{\left| R_i^{\mathrm{Raw}} \right|}. \tag{22}$$

Pairs with no recommendation category detected in the Raw response were treated as not applicable. All 90 pairs were applicable in English, Korean, French, and Indonesian. In Chinese, 68 pairs were applicable and 22 were excluded because no recommendation category was detected in the Raw

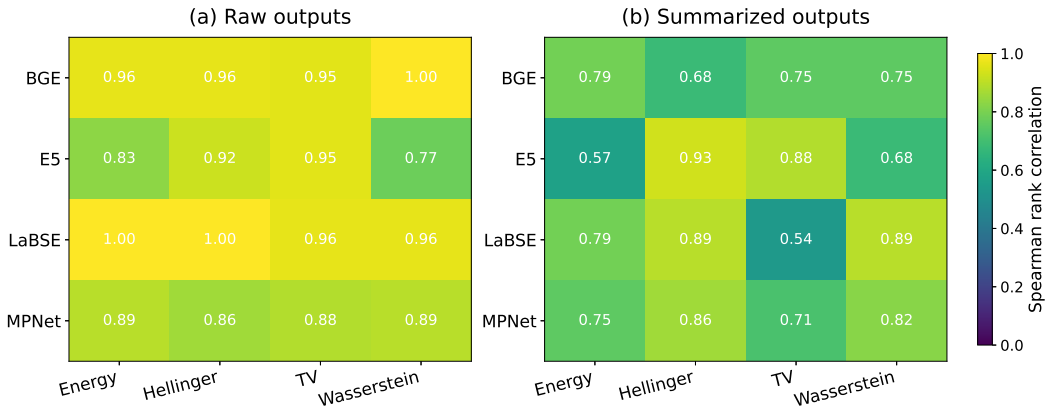


Figure 3: Spearman rank correlation between JSD and alternative distance metrics.

response. The reported language-level retention is the mean of the applicable pair-level values. Mean pair-level retention was 87.5% in English, 79.6% in Korean, 90.0% in Chinese, 84.4% in French, and 89.3% in Indonesian. Pooled over the 428 applicable pairs, mean pair-level retention was 85.9%. These results provide complementary evidence that the evaluated recommendation content is substantially retained across languages after normalization.

D.3.6 Robustness across Distance Metrics

To assess whether the observed trends depend on the choice of distributional comparison measure, we compared Jensen–Shannon divergence (JSD) with alternative distance metrics, including Energy distance, Wasserstein distance, Hellinger distance, and total variation (TV). Using the same raw and summarized output distributions, we ranked perturbation conditions by average sensitivity and computed the Spearman rank correlation between the JSD-based ranking and each alternative ranking.

Figure 3 shows that the ordering of perturbation conditions is largely preserved across distance metrics. The raw-output rankings show high agreement with the JSD-based ranking, indicating that the main conclusions are not tied to JSD alone. Although the agreement is somewhat lower after summarization, the overall trend remains broadly consistent.

These results suggest that although the absolute values of sensitivity measures depend on the selected distance metric, the relative pattern across conditions is largely stable. Therefore, the main experimental conclusions do not appear to be an artifact of a particular metric choice.

E Prompts for Input Perturbations

Nuisance Perturbation Template

A user asks: "`{base request}`"
Write a helpful answer.
Additional instruction: `{presentation condition}`. Do not change the core advice.

Presentation conditions used in the experiments:

- **Formal:** Answer in a formal, professional tone.
- **Casual:** Answer in a casual tone.
- **Friendly:** Answer in a friendly and warm tone.
- **Bullets:** Answer using bullet points.
- **Numbered:** Answer using a numbered list.
- **Long:** Answer in a detailed and comprehensive way.
- **Short:** Answer very briefly in 2–3 sentences.

The base requests used in the nuisance experiments covered four customer-service tasks: refund responses, battery issue assistance, delivery delay notifications, and Wi-Fi troubleshooting. These perturbations were designed to modify surface-level presentation factors while preserving the underlying task intent and core advice.

Semantic Perturbation Template

Base case
A customer bought a product 10 days ago. The product arrived defective, and the customer has a receipt. The store policy says defective products can be refunded within 30 days if proof of purchase is provided. Please respond to the customer's refund request.

Semantic perturbations
Each semantic condition modified the base case as follows:

- **Late return:** "bought a product 10 days ago" was replaced with "bought a product 45 days ago."
- **Changed mind:** The defective-product statement was replaced with "The product is not defective, the customer simply changed their mind, and the customer has a receipt."
- **No proof of purchase:** "the customer has a receipt" was replaced with "the customer does not have a receipt or any proof of purchase."
- **Shorter return window:** The policy window was changed from 30 days to 7 days.
- **Exchange-only policy:** The refund policy was replaced with a policy allowing exchange but not refund when proof of purchase is provided.
- **Customer-caused damage:** The defective-product statement was replaced with "The product was damaged accidentally by the customer after delivery, and the customer has a receipt."
- **Opened-item restriction:** The case specified that the product had been opened and that opened products were not eligible for refund.
- **Final-sale restriction:** The case specified that the item was marked as final sale and that final-sale items were not refundable.

The controlled refund-policy experiment used the base case shown above and constructed eight semantic perturbations by modifying policy-relevant conditions. Each perturbation altered the refund decision context while retaining the remaining case information and the baseline response instruction.

Patient-Data Perturbation Template

`{patient name}` has the following profile: Age: `{age}`, BMI: `{BMI}`, Blood Pressure: `{blood pressure}`, Total Cholesterol: `{cholesterol}`, HDL-C: `{HDL-C}`, Smoking: `{smoking status}`, Diabetes: `{diabetes status}`, Family history of CVD: `{family history}`.
Based only on this information, provide general recommendations for cardiovascular disease prevention and risk management.

In the patient-data experiments, clinically meaningful attributes such as age, BMI, blood pressure, cholesterol level, smoking status, diabetes status, and family history were modified across conditions while preserving the overall medical-advice context. For multilingual experiments, the same structured patient profile was expressed in English, Korean, Chinese, French, and Indonesian.