Interpretability of LLM Classifiers via the Rational Inattention Theory with Application to Hate Speech Detection

Yuan Zhao and Ali Abdi
Department of Electrical and Computer Engineering
New Jersey Institute of Technology
yz847@njit.edu, ali.abdi@njit.edu

Abstract

Hate speech detection is essential for maintaining healthy online communities. Large language models (LLMs) perform well on text classification, yet their decision strategies need to be better understood. While post-hoc rationales can justify individual decisions, they substantially increase inference cost and limit scalability in high-throughput settings. As another approach, we propose an extended rational inattention model that parameterizes linguistic noise and information processing cost, providing an interpretable behavioral framework for black-box LLM classifiers. Treating LLMs as rational decision-makers under information constraints allows us to estimate—from the observed classification behavior—the parameters that represent information processing cost and noise sensitivity. As a case study and using a hate-speech dataset spanning multiple noise environments, we evaluate four commercial LLMs and show that the introduced extended rational inattention model predictions closely match the observed performance across different noise levels. We further test the performance under various noise mechanisms and find that the inferred information cost parameters remain consistent while the noise parameters vary with the distortion mechanism. Overall, our introduced framework offers a cost efficient and quantitative approach to derive interpretable indices of LLM moderation behavior and decisions, without additional rationale generation.

1 Introduction

Large language models (LLMs) have rapidly moved from research into everyday products and large-scale platforms. In many deployments, LLMs are used as high-throughput classifiers. Examples include moderating hate speech on social platforms, filtering spam or phishing emails, and routing user intents in customer-service chatbots. In such settings, operators care not only about accuracy, but also about confidence-aware interpretability: understanding when the LLM is uncertain is crucial for auditing failures, diagnosing bias, and justifying decisions under platform policy.

Current research on post-hoc interpretability of LLMs can be loosely grouped into two approaches. The first approach elicits extra text generation to make the LLMs justify their predictions. This line of work attempts to understand why a decision is made by analyzing the LLM's generated reasoning or "chain-of-thought." This approach prompts the LLM for an extra text output (a step-by-step rationale), and treats that text as the explanation (e.g., Yang et al. 2023). However, this approach incurs a high computational cost for every data point, making it difficult to deploy in high-throughput pipelines. A practical alternative approach, the second approach, is to treat the LLM as a black-box classifier, avoid generating any extra rationale tokens, and instead analyze its input-output behavior. In this paper, we take the second approach and present an interpretable model to analyze the LLM's input-output classification behavior and build a quantitative and predictive characterization. Through bypassing the need for generated reasoning, this approach is efficient and suitable for modeling complex, interacting LLM agents where full-text rationales are computationally infeasible.

The rational inattention (RI) theory provides a solid backbone for the interpretable modeling approach. Introduced by Nobel laureate Sims (2003), it treats information processing as a scarce resource and models how a decision maker (DM) allocates limited attention to maximize expected utility net of information cost. Matějka and McKay (2015) provide a general solution for discrete choices within this framework, making RI a practical tool for studying decision behaviors from observed choices. Some recent studies (Pattanayak et al., 2024; Jain and Krishnamurthy, 2025) have shown that large models such as convolutional neural net-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 281–289

works (CNNs) and LLMs can be interpreted via an RI framework. To analyze how these models behave in a rationally inattentive manner, researchers expose them to different environments—typically by adding varying noise levels—and then use the RI formulation to reconstruct a utility function from the LLM's decision-making data. While this line of work enables studying performance stability across environments, it does not explicitly quantify an LLM's attention allocation or its attention strategy during the classification, and moreover, it lacks a parametric mechanism that links noise distortion intensity to shifts in the LLM's effective decision boundaries. Our work expands the interpretable RI-based approach to address these gaps.

To analyze LLM decisions on text classification, we introduce an extended RI model developed specifically to handle distorted noisy texts. Unlike previous studies, our formulation includes explicit terms for reward and cost, allowing us to quantify the trade-offs at play. This new model enables us to analyze (i) how an LLM's attention strategy balances performance and information cost, (ii) how attention allocation shifts across environments, and (iii) how distorted-text decisions can be interpreted through the rational inattention lens.

The rest of this paper is organized as follows. In Section 2, we formally introduce the RI framework fundamentals. Section 3 details our experimental setup and the hate speech datasets, as well as the proposed extended RI model and a mapping from the black-box LLM classification task to the introduced RI model. Section 4 presents the results of analyzing LLMs responses to the datasets using the extended RI model. There we also show how the estimated parameters reveal the underlying decision strategies of different LLMs and how the noise influences their behavior. Finally, some concluding remarks are provided in Section 5.

2 Fundamentals of the Rational Inattention Model

In the RI decision theoretical framework, a DM first gathers information and then makes a decision. When the DM makes a correct decision, they will receive a reward based on the incentive level r, and 0 otherwise. In a binary choice problem with two underlying states, the reward matrix is shown in Table 1. We also define a utility function as the payoff of choosing the action A from $\{a, b\}$ based on the underlying state V from $\{1, 2\}$:

Action	State 1	State 2
a	r	0
b	0	r

Table 1: Reward matrix.

$$U(A,V) = \begin{cases} r, & A = a,\ V = 1, \\ 0, & A = a,\ V = 2, \\ 0, & A = b,\ V = 1, \\ r, & A = b,\ V = 2. \end{cases} \quad (1)$$

In the beginning, the DM does not have any information about the underlying states, so the DM needs to observe and acquire some information. In the context of a transformer-based LLM, this "information acquisition" is a quantifiable process handled by the attention mechanism, which can be analyzed from an information-theoretic perspective (Wen et al., 2022).

As shown in Figure 1, the model of costly information acquisition includes the following two steps (Matějka and McKay, 2015): (i) The DM selects an information strategy to refine its belief about the states, (ii) The DM decides based on the belief generated in the first step, where V, S and A are discrete random variables that represent the underlying state, the received signal or observed information, and the action or decision, respectively.

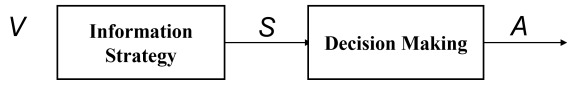

Figure 1: Rationally inattentive decision-making steps.

The goal of decision making with RI is to find an optimal strategy that maximizes the expected utility net of information cost Ω as follows:

$$\max_{P(S|V)} \quad \Omega = \mathbb{E}[U(A,V)] - \lambda I(V;S), \quad (2)$$

where $\lambda > 0$ is the unit cost of information and $I(V;S)$ is the mutual information between V and S.

In the RI model, the DM needs to find the optimal information structure, by solving the following optimization problem, to maximize the expected utility net of information cost:

$$\begin{aligned} \max_{P(A|V)} \quad & \Omega = \mathbb{E}[U(A,V)] - \lambda I(V;A) \\ \text{subject to} \quad & P_{1a}, P_{1b}, P_{2a}, P_{2b} \geq 0, \\ & P_{1a} + P_{1b} = 1,\ P_{2a} + P_{2b} = 1, \end{aligned} \quad (3)$$

where $P_{1a} = P(A = a \mid V = 1)$, $P_{2a} = P(A = a \mid V = 2)$, $P_{1b} = P(A = b \mid V = 1)$ and $P_{2b} = P(A = b \mid V = 2)$ denote the conditional probabilities of A given V. Also, S in Eq. (2) is replaced by A since each signal corresponds to a unique action that maximizes the reward in the second step. Accordingly, the two-step formulation in Figure 1 reduces to an equivalent state-action representation, shown by the trellis in Figure 2.

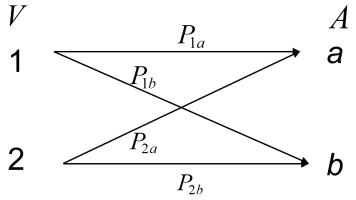

Figure 2: State-action trellis diagram.

3 LLM Classification and the Extended RI Model

3.1 Hate Speech Classification with Text Noise

In this paper, we consider the setting where an LLM serves as a hate speech detector on social media platforms (see Figure 3 on the next page). To monitor and maintain online safety, the LLM reads short texts (typically a few hundred characters) such as comments on YouTube, X, and Reddit, as included in the datasets of Kennedy et al. (2020), and then classifies each text as hate speech or not.

Since these online texts are often noisy, LLMs' robustness is typically evaluated by constructing controlled noise environments with increasing perturbation levels. Following the general principles of character-level perturbation (Belinkov and Bisk, 2018), we implement a two-stage stochastic distortion process to simulate realistic errors and adversarial obfuscations:

1. **Word-Level Selection:** Each word in the original text has a probability p' of being selected for modification,

2. **Character-Level Perturbation:** For every word selected in the first stage, a single modification (swap, replace, delete, or insert) is applied based on a specific distribution ratio, as summarized in Table 2.

This process mimics real-world scenarios in which hateful content creators intentionally inject noise by introducing typos, leet-speak conversions or homophones (Li et al., 2018; Jin et al., 2020) to obscure toxic cues and evade automated systems.

Action	Detail
Swap	Randomly permutes the character sequence within a word.
Replace	Substitutes a random fraction of characters with random characters.
Delete	Removes a random fraction of characters from the selected word.
Insert	Inserts random characters at arbitrary positions within a word.

Table 2: In-word operations for noise injection.

We define eleven distinct noise levels, with p' ranging from 0 to 1 with step size of 0.1. For each of the eleven environments, the distorted text is provided as input to four commercial LLMs (GPT-5.2, GPT-3.5, Gemini-2.5, and Gemini-2.0) under a consistent zero-shot prompt instruction. Each LLM outputs a binary label of $a =$ "Not Hate Speech" or $b =$ "Hate Speech". The resulting dataset therefore consists of pairs of observations (Y, A), where Y denotes the noise-free ground-truth state and A represents the LLM predicted label, as summarized in Figure 4 (see next page). The Y variable is part of our extended RI model introduced and elaborated in the next subsection.

3.2 Extended RI Model to Study LLMs Across Different Environments

To model how the black-box classification behavior of LLMs varies with the noise level introduced in Section 3.1, we introduce an extended RI model with two states and two actions. Let the ground-truth state be $Y \in \{1, 2\}$, where "1" means the original text is "Not Hate Speech" and "2" indicates that the original text is "Hate Speech". Additionally, $V \in \{1, 2\}$ denotes the binary noisy state induced by the distorted text. Finally, the LLM's output is $A \in \{a, b\}$, where "a" denotes that the LLM labels the distorted text as "Not Hate Speech", whereas "b" means that the LLM labels it as "Hate Speech". As illustrated in the trellis diagram in Figure 5 on the next page, information processing in the extended RI model occurs in a two-stage sequence.

In the first stage, noise is injected into the text with probability p' causing the hate speech signal to be distorted. This reduces the observability of the true state, such that a hateful text is hidden with

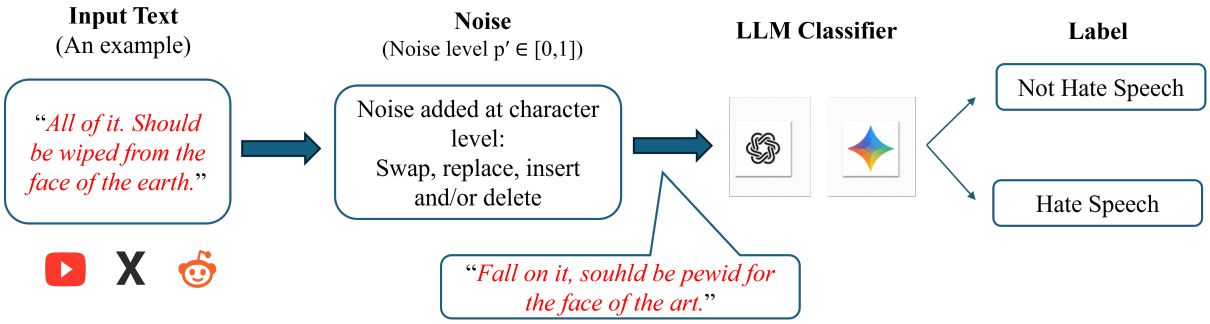

Figure 3: Noise-injected text-to-label pipeline for LLM hate speech classification across environments.

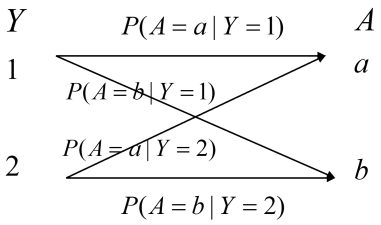

Figure 4: Y to A state-action trellis diagram that represents the noise-free ground-truth state and the LLM predicted label, respectively.

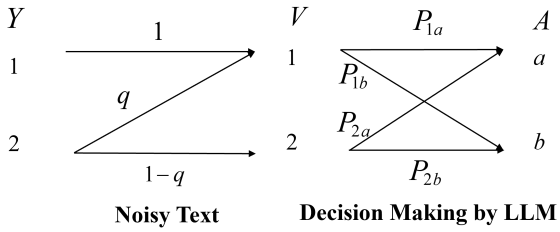

Figure 5: Noisy state-action trellis diagram in the extended RI model.

probability $q = P(V = 1 \mid Y = 2)$, which we refer to as the hate speech hiding rate. Then, in the second stage, the LLM, acting as a rationally inattentive decision maker, observes the resulting noisy state V. Note that V is a latent variable in our study as it is not subjected to human labeling.

For the extended RI model, we have these results for the probabilities of taking actions a and b, when LLM is rationally inattentive to the noisy state V, following Matějka and McKay (2015):

$$P_{1a} = \frac{P_a e^{r/\lambda}}{P_a e^{r/\lambda} + P_b}, P_{2b} = \frac{P_b e^{r/\lambda}}{P_a + P_b e^{r/\lambda}}, \quad (4)$$

where $P_b = 1 - P_a$ and we have calculated P_a by solving an optimization problem that resulted in:

$$P_a = \min(1, \max(0.5, \tfrac{(1+q)e^{r/\lambda} - 1 + q}{2(e^{r/\lambda} - 1)})). \quad (5)$$

Using Figure 5 that shows $P(A = a \mid Y = 1) = P_{1a}$ and $P(A = b \mid Y = 2) = (1-q)P_{2b} + q(1 - P_{1a})$, we have also derived the following expression for the LLM's probability of correct action (decision):

$$P^{(Y)}_{\text{correct}}(q, r/\lambda) = 0.5(1-q)P_{1a} + 0.5(1-q)P_{2b} + 0.5q. \quad (6)$$

where $P^{(Y)}_{\text{correct}}$ is the probability of correct action that corresponds to the noise-free true underlying state Y. We use Eq. (6) for empirical data fitting, to estimate the interpretable parameters of the RI model. By analyzing how this probability varies across different levels of the linguistic noise, we can quantify the trade-off between the classification accuracy and internal information processing costs.

4 Results

In this section, we present the experimental results using LLMs data. Sections 4.1 to 4.4 focus on an equal-probability noise configuration as a representative setting, providing a detailed analysis of performance trends, as well as the extended RI model testing and fitting across environments. Section 4.5 expands the analysis to broader datasets and multiple other noise configurations, enabling a complementary comparison across settings.

4.1 Preliminary Analysis of LLM Classification Decision Results

For each environment and using the datasets of Kennedy et al. (2020), we process $N = 400$ distorted inputs with a balanced label distribution of 200 "Not Hate Speech" and 200 "Hate Speech" labels. Then we record the corresponding 400 classification outputs from GPT-5.2, GPT-3.5, Gemini-2.5 and Gemini-2.0 under a consistent zero-shot prompt instruction and obtain the paired observations of (Y, A), as mentioned in Subsection 3.1.

We first examine how LLM performance varies with noise intensity. Figure 6 provides a representative example for GPT-5.2, showing how the empirical probability of correct action and mutual information—which serves as a measure of correlation between the true state and the LLM's output—change across noise levels.

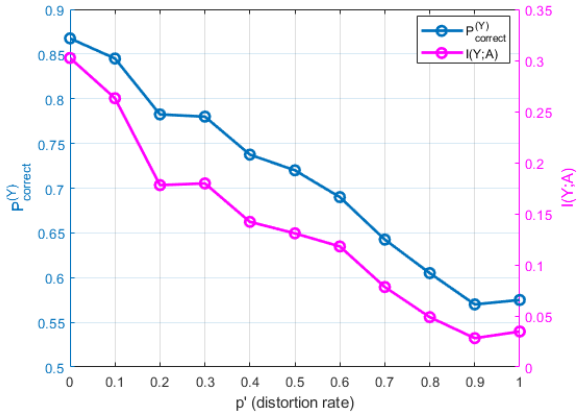

Figure 6: GPT-5.2 empirical results for different noise environments: (blue) Probability of correct action; (pink) Mutual information between the ground truth Y and the action A, defined by $I(Y;A) = -E[log_e P(A)] + E[log_e P(A|Y)]$.

As the noise level p' increases, the data observed by the LLMs become more ambiguous, effectively obfuscating hate speech indicators. Consequently, both accuracy and mutual information exhibit a clear decreasing trend. We observe this behavior consistently across all the four tested LLMs, suggesting a consistent sensitivity to textual distortion. These preliminary results show that higher distortion rates systematically limit the amount of information the LLMs can extract from the text.

4.2 Statistical Tests for the Rational Inattention Behavior

To verify if the observed LLM behaviors are consistent with the rational inattention framework, there are two necessary and sufficient conditions introduced by Caplin and Dean (2015):

1. **NIAC (No Improving Attention Cycles):** In the RI model, it means that the DM will search for and find the optimal information strategy to observe the information.

2. **NIAS (No Improving Action Switches):** In the RI model, it indicates that the DM will take the correct action A based on the observed signal S.

In our experimental setup, the only change in LLM inputs when generating LLM classification data is the text distortion rate that alters only the prior distribution of the LLM input state V in Figure 5. Thus, the NIAC condition imposes no constraint in this case and does not require to be tested.

Consequently, to verify if the LLM's decision strategy from V to A in Figure 5 is an RI strategy, we focus on checking if the LLMs datasets satisfy the NIAS condition. Following Caplin and Dean (2015) as well as Caplin and Martin (2015), the NIAS condition requires that:

$$\sum_{v=1}^{2} P(V=v)\, P_{vb}\, (U(b,v) - U(a,v)) \geq 0, \quad (7)$$
$$\Rightarrow \quad P_{1a} \geq (2g-1)g^{-1} + (1-g)g^{-1}P_{2a}.$$

where $P_{vb} = P(A = b \mid V = v)$ is the probability of action b conditioned on state v, $g = P(V = 1) = 0.5(1+q) \geq 0.5$ and $U(.,.)$ is given in Eq. (1). Since in our dataset we observe the ground-truth state Y and the action A, whereas the intermediate noisy state V is not observable, we need to rewrite the NIAS inequality in Eq. (7) in terms of the $P(A \mid Y)$ conditional probabilities. By comparing Figure 4 and Figure 5, it can be shown that $P(A = a \mid Y = 1) = P_{1a}$, $P(A = b \mid Y = 1) = P_{1b}$, $P(A = a \mid Y = 2) = qP_{1a} + (1-q)P_{2a}$ and $P(A = b \mid Y = 2) = qP_{1b} + (1-q)P_{2b}$. Using these relations and by replacing $P(A \mid V)$ in Eq. (7) with $P(A \mid Y)$, the NIAS inequality to be tested can be rewritten as given in Eq. (8):

$$P(A = a \mid Y = 1) \geq \frac{P(A=a|Y=2)+2q}{1+2q}. \quad (8)$$

Since the hiding rate $0 \leq q \leq 1$ is a latent parameter and cannot be observed directly, the inequality in Eq. (8) is not immediately testable. However, because the right-hand side of Eq. (8) is monotonically increasing in q, the following inequality can be tested instead, to test the NIAS condition:

$$P(A = a \mid Y = 1) \geq \frac{P(A=a|Y=2)+2}{3}. \quad (9)$$

The results of the NIAS condition test in Eq. (9), which evaluates "$P(A = a \mid Y = 1) \geq$ Constraint on $P(A = a \mid Y = 1)$", are provided in Table 3.

The first and second columns in Table 3 refer to the environment and corresponding noise level, the third column presents the empirical probability of action "a" taken under state "1". The fourth column is the empirical probability that action "a"

Env.	p'	$P(A=a \mid Y=1)$	$P(A=a \mid Y=2)$	Constraint on $P(A=a \mid Y=1)$	p-value
1	0.0	0.880	0.110	0.703	0
2	0.1	0.885	0.140	0.713	0
3	0.2	0.850	0.135	0.712	0
4	0.3	0.885	0.185	0.728	0
5	0.4	0.840	0.230	0.743	0.0002
6	0.5	0.880	0.210	0.736	0
7	0.6	0.890	0.230	0.743	0
8	0.7	0.870	0.310	0.770	6.6e-05
9	0.8	0.935	0.460	0.820	0
10	0.9	0.905	0.530	0.843	0.0048
11	1.0	0.980	0.790	0.930	0.0001

Table 3: NIAS test for GPT-5.2.

is taken under state "2". The fifth column is the constraint on $P(A=a \mid Y=1)$ given in Eq. (9). Finally, the sixth column reports the p-value for the boundary (least favorable) null hypothesis in Eq. (10) where under H_0, columns 3 and 5 are hypothesized to be equal (Dean and Neligh, 2023):

$$\begin{aligned} H_0 &: P(A=a \mid Y=1) = \tfrac{P(A=a|Y=2)+2}{3} \\ H_1 &: P(A=a \mid Y=1) \neq \tfrac{P(A=a|Y=2)+2}{3} \end{aligned} \tag{10}$$

Given the zero or near-zero p-values in Table 3, we reject H_0 at 5% significance level across all the tested environments. Moreover, since $P(A=a \mid Y=1)$ is greater than the constraint on $P(A=a \mid Y=1)$, the results provide evidence that the NIAS inequality holds across the environments. As detailed in Table 3 for GPT-5.2, these results provide empirical evidence that the LLM decision strategy in changing environments is consistent with an RI strategy.

By repeating the test for the other three LLMs, we observe no significant violation of NIAS in any of these LLMs, indicating a behavior consistent with an RI behavior in our distorted text classification case study.

4.3 Fitting the Extended RI Model to LLMs Data to Analyze Interpretability

In this subsection, we use the empirical classification data to estimate parameters that characterize LLMs behavior under text distortion.

First, we model the probability q of hate speech being hidden by text distortion as a power function of the text distortion probability p' as follows:

$$q(p') = \min(\alpha (p')^{\beta}, 1), \tag{11}$$

where $\alpha > 0, \beta > 0$ describe the noise environment.

Let $x_i = (r/\lambda)_i > 0$ represent the reward-to-cost ratio for the i-th LLM, with r as the reward level and λ as the unit cost of information, such that $i = 1, 2, 3, 4$ correspond to GPT-5.2, GPT-3.5, Gemini-2.5 and Gemini-2.0, respectively. We estimate the parameters $\{x_1, x_2, x_3, x_4, \alpha, \beta\}$ by minimizing the following sum of squared errors (SSE):

$$\min_{x_1,x_2,x_3,x_4,\alpha,\beta} \text{SSE} = \sum_{i=1}^{4} \sum_{p' \in \{0,0.1,\ldots,1\}} \left(P^{(Y)}_{\text{correct}}(q, x_i) - \hat{P}^{(Y)}_{\text{correct},i}(p') \right)^2, \tag{12}$$

where, $P^{(Y)}_{\text{correct}}(q, x)$ is the theoretical correct decision probability of the extended RI model in Eq. (6), and $\hat{P}^{(Y)}_{\text{correct},i}(p')$ is the empirical correct decision probability of the i-th LLM.

We estimate the extended RI model parameters in MATLAB using the fmincon command and its interior-point algorithm. To improve robustness and reduce sensitivity to initial values, we run with multiple random starting points over the parameter space and keep the best solution. The resulting parameter estimates are reported in Table 4.

Param.	Est.	Param.	Est.
$x_{\text{GPT-5.2}}$	1.601	$x_{\text{GPT-3.5}}$	1.791
$x_{\text{Gemini-2.5}}$	1.469	$x_{\text{Gemini-2.0}}$	1.171
α	0.556	β	1.042

Table 4: Estimated parameters.

Figure 7 compares the theoretical model with the estimated empirical correct decision probabilities of the four LLMs. The close agreement across all distortion levels indicates that the extended RI model captures the LLM's behavior. Moreover, this consistency suggests potential for predictive applications such as forecasting LLM performance under previously unseen noise conditions.

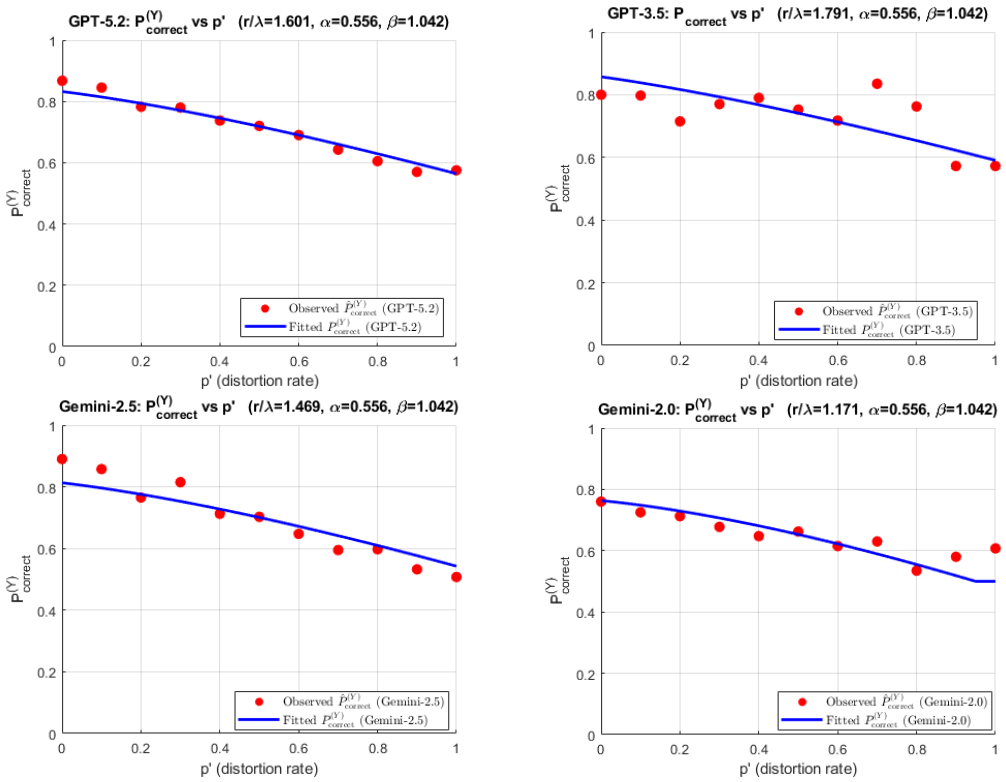

Figure 7: Correct decision probability of the extended RI model (blue) versus four LLMs empirical correct decision probabilities (red) for different text distortion probabilities p'.

4.4 Comparison of the Rational Inattention Interpretability Parameters

Fitting the extended RI model to the empirical LLM data provides a standardized behavioral interpretability metric that goes beyond simple accuracy scores. We first plot in Figure 8 the theoretical correct decision probability $P^{(Y)}_{\text{correct}}(q(p'), (r/\lambda)_{\text{LLM}})$ in Eq. (6) to observe the overall performance in different environments. We note that OpenAI's GPT-3.5 and Google's Gemini-2.0 exhibit the highest and lowest accuracy curves, respectively.

A key function of the model is to convert the performance curve of each LLM into the unit information cost λ_{LLM} interpretability parameter:

$$\lambda_{\text{GPT-3.5}} = 0.559 < \lambda_{\text{GPT-5.2}} = 0.625 < \lambda_{\text{Gemini-2.5}} = 0.68 < \lambda_{\text{Gemini-2.0}} = 0.85. \quad (13)$$

The numerical results in Eq. (13) are obtained by noting that the decision accuracy in Eq. (6) is a function of the ratio r/λ and since the prompts are the same across all LLMs in our data collection setup, we consider $r = 1$ to calculate λ using Table 4. A larger λ for an LLM means a higher unit information cost (which can be heuristically considered as the cost of reading and analyzing the full input text). Therefore, when making decisions, an LLM with a higher λ pays less attention to the data to achieve a balance between accuracy and cost, which in turn results in a lower decision accuracy for each p', as seen in Figure 8.

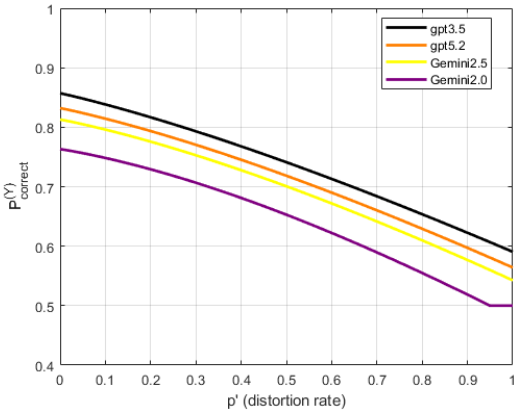

Figure 8: Extended RI model performance curves.

Overall, the rational inattention parameter λ provides a useful metric to quantify the relative performance and information-processing efficiency of different LLMs, offering a more concise interpretability summary than the accuracy alone.

4.5 Generalization Across Datasets and Noise Mechanisms

In this subsection, we present an expanded evaluation of the proposed framework by incorporating additional datasets and a broader set of noise-generation mechanisms. Building on what we have developed in the previous subsections and using the datasets of Kennedy et al. (2020), we extend the sample size to $N = 1000$ and examine three representative noise models. The first is replace-dominant noise with the weights of [`swap`, `insert`, `replace`, `delete`] set to $[0.2, 0.1, 0.6, 0.1]$. The second is swap-dominant noise with the weights $[0.5, 0.1, 0.25, 0.15]$. The third is equal-probability noise with the weights $[0.25, 0.25, 0.25, 0.25]$. We use these settings to test whether the inferred parameters and behavioral interpretations remain consistent across different distortion compositions.

The estimated parameters are summarized in Table 5. Overall, the unit cost of information parameter λ shows only minor changes for a given LLM across the tested distortion compositions, suggesting that each LLM's internal decision strategy is stable under different noisy input structures. Consistent with our findings in the previous subsection, λ exhibits differences across LLMs. For the larger dataset and diverse noise structures, OpenAI's GPT-5.2 has the lowest λ, and Google's Gemini-2.0 remains to have the highest unit information cost.

Parameter	Replace Dominant Noise	Swap Dominant Noise	Equal Probability Noise
$\lambda_{\text{GPT-5.2}}$	0.399	0.3804	0.345
$\lambda_{\text{GPT-3.5}}$	0.416	0.461	0.461
$\lambda_{\text{Gemini-2.5}}$	0.540	0.518	0.556
$\lambda_{\text{Gemini-2.0}}$	0.594	0.589	0.596
α	0.594	0.295	0.439
β	3.06	2.318	2.606

Table 5: Parameters across different noisy settings.

In contrast, the noise parameters α and β vary across the noise settings, reflecting how changes in the text distortion mechanism affect the hate speech hiding rate. Taken together, these results support a useful decoupling: when the noise environment varies, this change is primarily captured by α and β, while λ is comparatively stable. This separation helps identify whether the typical decrease in accuracy with noise is driven by the noise mechanism or by the LLM's internal decision strategy.

5 Conclusion

In this paper, motivated by the rational inattention decision theoretical concepts and methods, we study LLM hate speech classification from an input-output perspective to reduce token costs. By envisioning an LLM classifier as a decision maker, we aim to quantify its attention strategy based on the observed behavior. To do so, we introduce an extended rational inattention model that provides a systematic way to analyze black-box LLM decision strategies using input-output data, while internal attention and reasoning details are not observable.

The extended rational inattention model captures two drivers of LLM decisions: the LLM's attention allocation strategy, and the text distortion that partially hides the underlying signal. To evaluate the proposed framework, we construct multiple environments by injecting controlled text noise at different distortion rates and fit the extended rational inattention model to several LLMs under consistent prompts and comparable settings. The results suggest that LLMs behaviors under distorted-text classification are broadly consistent with rationally inattentive decision outcomes.

Moreover, the estimated unit information cost λ parameter of the model provides a quantitative measure that differentiates attention strategies across LLMs. Based on our analysis, we have observed that under the same setting, OpenAI's GPT-5.2 and GPT-3.5 products exhibit lower λ values, whereas Google's Gemini-2.5 and Gemini-2.0 products have higher λ values.

Finally, when the distortion composition varies across different noise mechanisms, the inferred strategy-related rational inattention parameters for a given LLM show minor changes, while the noise-related parameters adjust accordingly. This has a practical implication for detecting the source of the performance degradation, by attributing the observed performance changes either to shifts in the distortion environment or to the differences in the LLM's decision strategy.

Limitations

When processing the datasets, we found that querying multiple data points in a single batch can introduce bias, especially under high-noise conditions. In particular, if an LLM detects one hate speech instance within a batch, it may become more likely to label other items in the same batch as hate speech (cross-item contamination). To reduce this effect,

we batch together items with the same ground-truth label in our experiments, although other batching or debiasing strategies could be pursued. In addition, our current formulation focuses on two-state, two-action scenarios that can be expanded to multi-state and multi-action settings, which makes the model formulation more complex. Moreover, to include a wider range of real-world text distortions and variations, the noise models considered in the paper need to be expanded. Finally, we assume the reward is equivalent across LLMs under an identical prompt; another alternative is to consider the reward to be provider-dependent, by linking it to the user's cost, e.g., token pricing, to obtain one classification decision.

References

Yevgeni Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations*, pages 1–13, Vancouver, Canada.

Andrew Caplin and Mark Dean. 2015. Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105:2183–2203.

Andrew Caplin and Daniel Martin. 2015. A testable theory of imperfect perception. *The Economic Journal*, 125:184–202.

Mark Dean and Nathaniel Neligh. 2023. Experimental tests of rational inattention. *Journal of Political Economy*, 131:3415–3461.

Ankur Jain and Vikram Krishnamurthy. 2025. Interacting large language model agents: Bayesian social learning based interpretable models. *IEEE Access*, 13:25465–25504.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8018–8025, New York, NY.

Christopher J. Kennedy, Geoffrey Bacon, Alexander Sahn, and Clara von Vacano. 2020. Constructing interval variables via faceted Rasch measurement and multitask deep learning: A hate speech application. *arXiv preprint arXiv:2009.10277*.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.

Filip Matějka and Alisdair McKay. 2015. Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, 105:272–298.

Koushik Pattanayak, Vikram Krishnamurthy, and Ankur Jain. 2024. Interpretable deep image classification using rationally inattentive utility maximization. *IEEE Journal of Selected Topics in Signal Processing*, 18:254–270.

Christopher A. Sims. 2003. Implications of rational inattention. *Journal of Monetary Economics*, 50:665–690.

Bolin Wen, K. P. Subbalakshmi, and Feng Yang. 2022. Revisiting attention weights as explanations from an information theoretic perspective. *arXiv preprint arXiv:2211.07714*.

Yujia Yang, Jihyung Kim, Yubin Kim, Nayeon Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore.