# HEXACO Personality Traits as Safety Signals: Detecting Inappropriate LLM Responses via Text-Inferred Personality Scores

Anonymous ACL submission


## Abstract

As large language models (LLMs) rapidly evolve, inappropriate text generation has become a critical concern, yet constructing comprehensive guardrail datasets that cover the full diversity of harmful content remains a practical challenge. We propose a detection framework combining two complementary components: a HEXACO-based personality-informed detector that estimates factor scores from response text, and an ensemble algorithm that integrates conventional detectors. We empirically validate that unsafe responses exhibit systematically distinct HEXACO profiles: Agreeableness and Extraversion are consistently depressed globally, while within harm-category comparisons against safe refusals show elevated Openness in unsafe compliance responses. On the AEGIS2 benchmark, the HEXACO decision model achieves recall=0.975 with F1=0.651, functioning as a near-complete recall booster. Among four WildGuard–HEXACO integration strategies, the hybrid k-NN variant (WG+HDist) reaches AEGIS F1=0.841, which is statistically indistinguishable from a HEXACO-free text-only k-NN baseline (0.844; McNemar $p = 0.423$), and provides a compact, interpretable 6-dimensional distance representation as an alternative to 1024-dimensional text similarity.


## 1 Introduction

Preventing inappropriate outputs in LLM-based applications is a critical challenge. Blacklist-based filters suffer from context-dependent false positives, while learning-based detectors require large labeled datasets that are costly to construct.

Our choice of the HEXACO model (Ashton and Lee, 2007) as the underlying personality framework is motivated by both conceptual fit and methodological transparency. HEXACO extends the Big Five (Goldberg, 1990) with an additional Honesty-Humility dimension and refines several other factors, yielding a structured set of personality traits that have been linked to morally and socially relevant behaviors, and that align conceptually with what distinguishes unsafe LLM outputs from safe refusals. The six-factor structure further offers interpretability, since each safety judgment can be traced back to specific personality dimensions rather than to opaque classifier weights. These properties make HEXACO factor scores a natural candidate signal source for text-based safety detection, with the question of which individual factors carry the strongest discriminative signal left to be answered empirically.

We propose a personality-informed detection framework whose personality estimator is trained without safety labels, while a lightweight downstream Decision Model is calibrated on labeled data. Our approach is motivated by a finding from personality psychology: low HEXACO factor settings tend to elicit inappropriate LLM outputs when used as prompt instructions (Wang et al., 2025). We investigate the converse direction: whether unsafe responses exhibit distinct HEXACO profiles when estimated from text. We find that they do, with directional differences that vary by factor and analysis scope, and these differences provide actionable safety signals. We further combine HEXACO-based detection with conventional detectors via an ensemble framework, leveraging the complementary strengths of both signal sources.

Our contributions are as follows:

- We propose a HEXACO-based detector that pairs a personality Estimator with a Decision Model over the resulting factor scores; on AEGIS2 it attains F1=0.651 with recall=0.975, functioning as a near-complete recall booster.
- We design a selective-complementation ensemble that overrides WildGuard's negative (safe) decisions using HEXACO signals. Four variants are tuned on a held-out AEGIS2 validation split; the hybrid k-

NN variant (WG+HDist) does not outperform a HEXACO-free text-only k-NN baseline (AEGIS F1=0.841 vs. 0.844; McNemar $p = 0.423$), and the AND-style category-gate variant (WG+HGate) converges to WildGuard's behaviour once its category set is derived without test-set leakage, indicating that personality-informed signals act as a compact, interpretable distance feature rather than as an independent gating condition.

- We empirically validate that unsafe LLM responses exhibit distinct HEXACO profiles: globally, Agreeableness ($d = -0.65$) and Extraversion ($d = -0.34$) are depressed; within each harm category, Openness is elevated across all 24 categories (with Conscientiousness depressed in 17 of 24).

## 2 Related Work

### 2.1 Personality Models and HEXACO Framework

The HEXACO model (Ashton and Lee, 2007) extends the Big Five (Goldberg, 1990) into a six-factor personality framework by adding Honesty-Humility as a distinct factor. The six factors are Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience, each comprising four facets and associated with positive and negative descriptor terms. The full list of factors, facets, and negative descriptors is provided in Appendix A.

Wang et al. (2025) demonstrated that prompting LLMs with low HEXACO factor instructions elicits more inappropriate outputs, particularly under low Agreeableness and low Openness. We investigate the converse question: whether unsafe responses themselves exhibit distinct HEXACO profiles when their factor scores are inferred from text.

### 2.2 Existing Approaches to Inappropriate Text Detection

We categorize existing approaches to inappropriate-content detection into three groups: supervised learning-based detectors, commercial moderation services, and ensemble methods.

#### 2.2.1 Supervised Learning-Based Detectors

Meta's *Llama-Guard* family (Inan et al., 2023; Dubey et al., 2024) exemplifies a prominent class of supervised safety classifiers trained on curated safety-annotated datasets. Other supervised approaches, including Google's *ShieldGemma* (Zeng et al., 2024), *BeaverDam* (Ji et al., 2023), and *MD-Judge* (Li et al., 2024), adapt language models for content moderation using explicit categorical harm labels. Public datasets such as *ToxicChat* (Lin et al., 2023), collected from real-world user–AI conversations, and the red-teaming benchmark *HarmBench* (Mazeika et al., 2024) have further advanced supervised-detector development by exposing distribution shifts and adversarial jailbreak inputs. Shared evaluation frameworks such as SemEval-2020 Task 12 (Zampieri et al., 2020) have contributed to multilingual offensive language identification by establishing benchmarks and taxonomies for content moderation research. While effective, these models require comprehensive training datasets and are sensitive to distribution shift.

#### 2.2.2 Commercial and API-Based Services

The *OpenAI Moderation API* (Markov et al., 2023) is a closed-source commercial service for real-time harm classification. While convenient for deployment, such services lack transparency in their decision-making logic and raise concerns about data privacy and service dependency (Ghosh et al., 2025).

#### 2.2.3 Multi-Task and Ensemble Approaches

Han et al. (2024) introduced *WILDGUARD*, a framework that unifies prompt classification, response classification, and refusal detection within a single model. Ghosh et al. (2025) proposed *AEGIS2.0*, which combines supervised classification with ensemble reasoning over multiple external baselines. Rebedea et al. (2023) introduced *NeMo Guardrails*, a production-oriented toolkit that composes multiple programmable safety rails for deployable LLM applications.

### 2.3 Motivation for Personality-Informed Detection

We propose a *personality-informed* detection framework that infers response safety from HEXACO factor scores estimated from response text, grounding the detection features in psychological trait theory. Unlike supervised classifiers that learn harm-category-specific patterns, personality trait signals capture stylistic dimensions of language that are largely orthogonal to topical content. This orthogonality yields complementarity with existing supervised detectors, as the resulting signal can

recover unsafe samples in categories that those detectors miss. We hypothesise that unsafe responses exhibit systematically distinct HEXACO factor profiles, rendering HEXACO-based detection suitable as a high-recall component within an ensemble.

# 3 Proposed Method

## 3.1 Detection Framework

Building on the motivation in Section 2.3, our framework pairs HEXACO-based detection (Section 3.2) with conventional detectors via an ensemble (Section 3.3), exploiting their complementary error profiles.

## 3.2 HEXACO-based Inappropriate Text Detection

As illustrated in Figure 1, our HEXACO-based detector consists of two components: a *HEXACO Estimator* that infers six factor scores from a response text, and a *Decision Model* that maps these scores to a binary safe/unsafe judgment.

### 3.2.1 HEXACO Estimator

The estimator takes a target text $\mathbf{t}$ as input and outputs six factor scores $s_i = M_i(\mathbf{t})$ for $i \in \{H, E, X, A, C, O\}$. We consider two implementations of the estimator.

The first implementation uses an LLM-as-a-Judge approach (Zheng et al., 2023): we prompt the model with factor-specific descriptors and ask it to rate each factor on a 1–5 scale, with the prompt template shown in Appendix B.

The second implementation, which we refer to as the *local HEXACO estimator*, fine-tunes a compact text encoder via contrastive learning (Chen et al., 2020; Gao et al., 2021) on synthetic HEXACO-conditioned responses. The training data are generated from RealToxicityPrompts (Gehman et al., 2020): for each source prompt, an LLM produces responses under each HEXACO factor's high- and low-polarity persona descriptions, and the resulting pairs are used as contrastive examples; full details of the data construction (including the LLM used) are provided in Appendix C.1, and the backbone model and training hyperparameters are described in Section 4.3. Unlike the LLM-as-a-Judge variant, the local estimator's scores are not calibrated to the 1–5 inventory scale, so any threshold $\tau$ derived from them must be recalibrated whenever the estimator is updated (see Limitations §6).

### 3.2.2 Decision Model

Given the factor scores produced by the estimator, the Decision Model yields a binary safety prediction. We evaluate two strategies: a simple threshold-based rule, and a learning-based classifier trained on the factor-score feature vector.

**Threshold-based Decision.** The simplest rule classifies a text as inappropriate when at least one factor score falls below a threshold $\tau$:

$$y_{hexaco} = \mathbf{1}\left[\min_i s_i \leq \tau\right]. \tag{1}$$

A higher $\tau$ increases recall at the cost of precision, and we optimize $\tau$ by sweeping candidate values on a held-out validation set and selecting the value that maximizes F1. As a finer-grained variant, we assign an independent threshold $\tau_i$ to each factor $i$, so the decision becomes $y = \mathbf{1}[\exists\, i : s_i \leq \tau_i]$. This monotone formulation only partially matches the empirical HEXACO signal (Appendix D.1), motivating the unconstrained Learning-based Decision Model below.

**Learning-based Decision Model.** As an alternative to thresholding, we cast the safe/unsafe classification as a supervised learning problem over HEXACO features. From the six-factor score vector estimated from the response side of a conversation turn, we construct a 10-dimensional feature vector:

$$\begin{aligned}\mathbf{x} = (&s_H, s_E, s_X, s_A, s_C, s_O,\\ &\min_i s_i, \bar{s}, \max_i s_i, \sigma_s),\end{aligned} \tag{2}$$

where $\bar{s}$ and $\sigma_s$ are the mean and standard deviation of the six factor scores. We use only the response side of the conversation, since the goal is to detect whether the generated response is inappropriate.

A binary classifier $f : \mathbb{R}^{10} \to \{0, 1\}$ is trained on labeled examples, and we evaluate two classifier families: logistic regression and LightGBM (Ke et al., 2017). To address class imbalance, we undersample safe samples to a 1:1 ratio with unsafe samples in the training set. The feature subset, regularization strength, and the final classification threshold are jointly selected on a held-out validation set by maximizing F1, with the resulting configuration reported in Section 4.3.

## 3.3 Ensemble-based Inappropriate Text Detection

We propose an ensemble detector that combines HEXACO-based detection with conventional detectors via category-dependent weighted aggregation

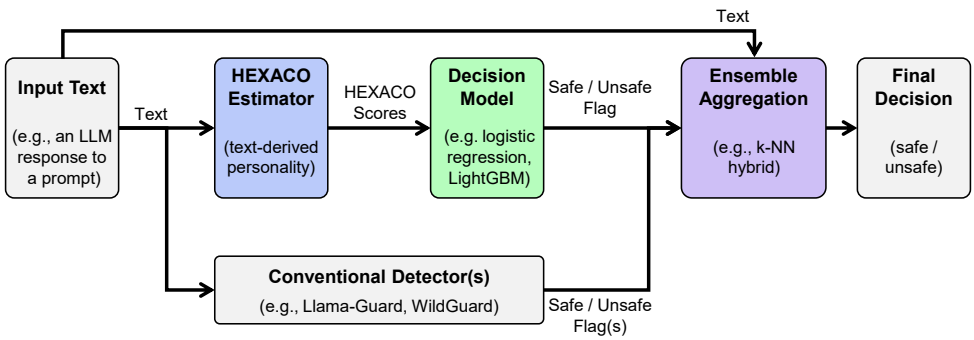


Figure 1: Overview of the proposed detection framework: HEXACO-based detection combined with ensemble approaches.

of multiple weak detectors. Figure 1 illustrates the overall framework, including the ensemble integration step. Beyond this multi-detector aggregation, we additionally explore a focused alternative that selectively complements the strongest single conventional detector on AEGIS2 (identified empirically in RQ3); this specialization is described together with its results in Section 4.5.

The proposed aggregation assigns category-dependent weights $w_i(c)$ to $N$ weak detectors, which include conventional safety classifiers and the HEXACO Detector, and aggregates their binary outputs. Let the $i$-th weak detector be $M^i_{weak}$ with output $y^i_{weak} = M^i_{weak}(\mathbf{t}) \in \{0, 1\}$; the final decision is

$$y_{ensemble} = g\left(\sum_i y^i_{weak} \cdot w_i(c)\right), \quad (3)$$

where $g$ is a thresholding function and $c = M_{cat}(\mathbf{t})$ is the estimated harm category via LLM-as-a-Judge (Zheng et al., 2023). We evaluate three weighting strategies for $w_i(c)$; full results are reported in Appendix E.1.

**Metric-weighted Aggregation.** For each detector $i$ and category $c$, a metric score $m_i(c) \in \{\text{F1}, \text{Recall}, \text{Precision}\}$ is computed on the training set and normalized within the category:

$$w_i(c) = \frac{m_i(c)}{\sum_j m_j(c)}. \quad (4)$$

Selecting $m = \text{Recall}$ biases aggregation toward detectors that recover more unsafe samples per category; $m = \text{Precision}$ favors lower false-positive behavior.

**Embedding-based k-NN Aggregation.** For a test sample, its $S$ nearest neighbors are retrieved from the training set using cosine similarity over text embeddings. Weights are averaged from the neighbors' category-specific weights:

$$W_i = \frac{1}{S}\sum_{k=1}^{S} w_i(c_k), \quad (5)$$

where $c_k$ is the category of the $k$-th neighbor; safe-sample neighbors use uniform weight $1/N$.

**Embedding-based Clustering.** K-means clustering is applied to the training embeddings and F1-based weights are computed for the discovered clusters, removing the dependency on predefined category labels.

## 4 Evaluation Experiments

We evaluate the proposed method from the following perspectives:

- **RQ1**: Do unsafe LLM responses exhibit systematically different HEXACO factor scores than safe responses, testing the converse direction of the findings in Wang et al. (2025)?
- **RQ2**: What decision logic best leverages HEXACO factor scores for binary safe/unsafe classification?
- **RQ3**: How does the best-configured HEXACO Detector compare with conventional guardrail methods?
- **RQ4**: Does the ensemble detector further improve detection performance by combining the HEXACO-based and conventional detectors?

### 4.1 Experimental Setup

We compare four detectors: OpenAI Moderation API (omni-moderation-latest, as of May 19, 2026), LlamaGuard (Dubey et al., 2024) (Meta, run locally), WildGuard (Han et al., 2024) (AllenAI), and the proposed HEXACO Detector; model details are in Appendix C.2.

We evaluate all methods on the AEGIS2 benchmark (Ghosh et al., 2025), covering diverse harm categories such as hate speech, violence, criminal planning, and sexual content. The train split (30,007 samples; 3,541 unsafe, 11,693 safe, and 14,773 null-label human-turn entries) is used for signal analysis (RQ1), Decision Model training, and ensemble estimation (RQ4). Null-label rows are retained as safe for Decision Model training and ensemble tuning but excluded from RQ1's per-factor analysis; see Limitations §6.1.3. The Decision Model is trained on this split; the HEXACO estimator itself is trained independently on synthetic personality-conditioned responses generated from RealToxicityPrompts (Gehman et al., 2020) (details in Appendix C.1), with no exposure to AEGIS2 labels. The test split (1,964 samples; 394 unsafe / 1,570 safe) is held out for all final evaluations (RQ2–RQ4).

We additionally evaluate on WildGuardMix test (Han et al., 2024) (1,725 samples; 284 unsafe) and XSTest (Röttger et al., 2024) (446 samples; 78 unsafe) for OOD generalization; dataset details are in Appendix C.3. For each benchmark, detectors are applied to the response/output text when available; when only prompt text is available (e.g., HarmBench used in §6.2.4), the result reflects prompt-side safety detection rather than response-level robustness.

The local HEXACO estimator uses Gemma-3-270M-IT (Gemma Team, 2025) fine-tuned with contrastive learning (Enc-E; see RQ2); LLM-based variants use OpenAI's gpt-5.4-nano.

For both the Decision Model (RQ2) and the WildGuard–HEXACO ensemble (RQ4), hyperparameters are selected on a 20% hash-based validation split of AEGIS2 train and applied unchanged to AEGIS2 test, WildGuardMix, and XSTest; full split-size and ensemble-row restrictions are documented in Appendix C.4.

### 4.2 RQ1: HEXACO Factors as Safety Signals

Whereas Wang et al. (2025) showed that prompting LLMs with low HEXACO factor settings elicits more inappropriate outputs, we test the converse hypothesis: that unsafe LLM responses themselves carry systematically distinct text-derived HEXACO scores.

Table 1: HEXACO factor statistics: safe vs. unsafe LLM responses (AEGIS2 train split, Enc-E estimator; null-label human-turn entries excluded). $\Delta$ = unsafe − safe mean. Benjamini–Hochberg false discovery rate (BH-FDR) correction (Benjamini and Hochberg, 1995) applied. *** $p < 0.001$, * $p < 0.05$. Score values reflect the local estimator's output distribution, not the absolute 1–5 HEXACO inventory scale. The relationship between this global comparison and the within-harm-category analysis of §RQ4 is discussed in Appendix D.2.

| Factor | Safe | Unsafe | $\Delta$ | $p_{\text{fdr}}$ | Cohen's $d$ |
|---|---|---|---|---|---|
| H (Honesty) | 3.16 | 3.09 | −0.07 | *** | −0.12 |
| E (Emotionality) | 2.95 | 3.00 | +0.05 | *** | +0.16 |
| X (Extraversion) | 3.07 | 2.90 | −0.17 | * | −0.34 |
| A (Agreeableness) | 2.93 | 2.60 | −0.33 | *** | −0.65 |
| C (Conscientiousness) | 2.91 | 3.02 | +0.11 | *** | +0.28 |
| O (Openness) | 2.80 | 2.69 | −0.11 | *** | −0.15 |

Table 1 shows the results. All six factors exhibit statistically significant differences after BH-FDR correction. The dominant signals are Agreeableness ($d = -0.65$) and Extraversion ($d = -0.34$), both depressed for unsafe responses; Conscientiousness shows a small positive effect ($d = +0.28$), and Honesty-Humility and Openness show small negative effects.

We then conduct a category-level analysis in which each unsafe category is compared against the entire safe population using the Enc-E estimator. This analysis reveals that Agreeableness is the most consistent discriminator across all harm categories in the train split, with large negative effects in threat, profanity, harassment, and violence categories. A complementary within-category comparison (detailed in Section 4.5) further shows Openness elevated and Conscientiousness depressed in unsafe compliance responses relative to safe refusals to the same prompt category.

***Finding (RQ1):*** *All six HEXACO factors discriminate safe from unsafe responses (BH-corrected); the dominant global signals are depressed Agreeableness ($d = -0.65$) and Extraversion ($d = -0.34$), and a within-category analysis (RQ4) further shows Openness elevated across all 24 harm categories (with Conscientiousness depressed in 17 of 24).*

### 4.3 RQ2: Decision Logic for HEXACO-based Detection

#### 4.3.1 Local Estimator Architecture Ablation

We first ablate the local HEXACO estimator architecture over five configurations (Table 2). Enc-B and Enc-C adopt a more expressive head but freeze the backbone, achieving substantially worse MAE ($> 1.0$). Unfreezing the backbone (Enc-A, Enc-E) consistently improves representation quality. Enc-E, which combines an MLP head with full fine-tuning and discriminative learning rates (backbone $2 \times 10^{-5}$, head $3 \times 10^{-4}$), achieves the best macro F1 (0.857) and MAE (0.384).

#### 4.3.2 Threshold-based and Learning-based Decision

For threshold-based detection, we apply the decision rule in Eq. 1. Factor-wise optimization (independent $\tau_i$ per factor, trained on AEGIS2 train) achieves F1=0.622 with precision 0.462 and recall 0.949 (Table 3).

As an alternative, we train two binary classifiers on the 10-dimensional feature vector described in Section 3.2.2. A 54-trial hyperparameter sweep identifies a 3-dimensional subset {O, A, X} ("oax") with $C = 0.01$ as the best LogReg configuration; we then set the decision threshold to $0.16$ via a recall-constrained sweep, prioritising the intended high-recall booster role over peak validation F1.

Table 3 shows that the oax LogReg achieves the highest F1 (0.651) with near-perfect recall (0.975, FN=10), while LightGBM favors precision (0.531) at lower recall. Although RQ1 identifies Agreeableness and Extraversion as the dominant individual signals (Table 1), the empirically optimal feature subset is {O, A, X}. The inclusion of O reflects empirical complementarity with A and X in the joint model rather than a strong standalone marginal effect ($d = -0.15$), while adding C did not improve validation F1 despite its individual significance ($d = +0.28$), suggesting its signal is subsumed by the O–A–X combination. The key trade-off is precision (0.489): oax LogReg generates more false positives than the default $\tau = 1.0$ rule, though fewer than the factor-wise optimal variant (401 vs. 436), making it suitable for recall-critical deployments.

#### 4.3.3 Validation-Set Tuning

Before evaluating on the test set, the Learning-based Decision Model threshold is optimized on a held-out validation set (20% of AEGIS2 train, 5,123 samples; 741 unsafe) using the same class-balanced training set. Table 4 shows the best threshold and validation metrics for both classifiers. The oax LogReg ($C = 0.01$) achieves val F1=0.399 at $\tau^* = 0.16$; the corresponding test result is F1=0.651, a substantial increase attributable to a distributional shift between the two pools (Appendix D.3). Although LightGBM achieves higher validation F1 (0.544 vs. 0.399), its lower recall (0.692 vs. 0.984) makes it unsuitable for the high-recall booster role; we therefore adopt the oax Logistic Regression as the main HEXACO Detector and report LightGBM as a precision-oriented alternative.

***Finding (RQ2):*** *The oax Logistic Regression on {O, A, X} ($C$=0.01, $\tau$=0.16) achieves F1=0.651 / Recall=0.975 and is selected as the high-recall HEXACO Detector for subsequent experiments.*

### 4.4 RQ3: Comparison with Conventional Detectors

As shown in Table 5, WildGuard achieves the highest F1 (0.804), while the HEXACO Detector reaches a lower F1 (0.651) but exhibits a markedly different profile: near-perfect recall (0.975, +0.193 over WildGuard), revealing a complementary error pattern that motivates ensemble combination in RQ4. Supplementary results for LLM-based baselines (HEXACO+GPT, LLM-Judge) are provided in Appendix E.2.

***Finding (RQ3):*** *The HEXACO Detector lags WildGuard in F1 (−0.152) but substantially exceeds it in recall (+0.193), revealing a complementary detection profile that motivates ensemble combination.*

### 4.5 RQ4: WildGuard–HEXACO Ensemble

**Multi-Detector Aggregation Baseline.** Applying the multi-detector weighted aggregation framework of Section 3.3 to all four detectors, the strongest baseline (majority vote, $\geq 2/4$ detectors) attains the highest AEGIS F1 (+0.024 over WildGuard) but degrades on XSTest, and no other variant maintains WildGuard's performance across all three benchmarks (full results in Appendix E.1); we therefore explore a more targeted alternative below.

**Selective Complementation.** Pre-analyses (Appendix D.4) show that HEXACO achieves high

Table 2: HEXACO local estimator ablation across five head architectures, evaluated on the synthetic held-out test split (10% of RealToxicityPrompts-derived data, hashed by source-prompt ID; see Appendix C.1). AEGIS2 test is not used for this architecture selection. *Linear* = single linear head; *Enhanc.* = three-layer enhanced head; *MLP* = two-layer projection with GELU activation (full Enc-E spec in Appendix C.1). *Frozen* = backbone weights fixed (head-only LR shown). MAE = mean absolute error against 1–5 reference scores; Macro F1 = score-bin classification F1 averaged over factors.

| Run | Head | Frozen | LR (back./head) | MAE↓ | Macro F1↑ |
|---|---|---|---|---|---|
| Enc-A | Linear | No | 2e-5 / 2e-5 | 0.451 | 0.839 |
| Enc-B | Enhanc. | Yes | — / 3e-4 | 1.082 | 0.563 |
| Enc-C | Enhanc. | Yes | — / 3e-4 | 1.088 | 0.564 |
| Enc-D | MLP | Yes | — / 3e-4 | 0.844 | 0.682 |
| **Enc-E** | **MLP** | **No** | **2e-5 / 3e-4** | **0.384** | **0.857** |

Table 3: HEXACO Detector performance on AEGIS2 test (Enc-E). *Default* $\tau = 1.0$ applies Eq. 1 with a single global threshold; *Factor-wise optimal* uses independent $\tau_i$ per factor, tuned on AEGIS2 train. *oax LogReg* uses only {O, A, X} features with L2 strength $C = 0.01$ and probability threshold $\tau = 0.16$. Learning-based classifiers are trained on a 1:1 class-balanced subset (safe samples undersampled).

| Strategy | Config | Prec. | Recall | F1 | Acc. |
|---|---|---|---|---|---|
| **Threshold-based** | Default $\tau = 1.0$ | 0.578 | 0.178 | 0.272 | 0.809 |
| | Factor-wise optimal | 0.462 | 0.949 | 0.622 | 0.768 |
| **Learning-based** | oax LogReg ($C$=0.01) | 0.489 | 0.975 | **0.651** | 0.791 |
| | LightGBM (sweep best) | 0.531 | 0.657 | 0.587 | 0.815 |

*rescue efficiency* against WildGuard FNs in a handful of harm categories and exhibits an *open but irresponsible* within-category profile (Openness elevated in 24/24 categories, Conscientiousness depressed in 17/24). Building on these observations, we adopt *selective complementation*: we retain WildGuard's positive decisions and override only its negative (safe) decisions via $y_{\text{ensemble}} = y_{\text{WG}} \vee \rho(\mathbf{t})$, with four override predicates $\rho$ spanning the design space of label requirement and HEXACO usage (formal setup in Appendix D.4).

**Integration Strategies.** The four strategies are summarised in Table 6; their validation-tuned hyperparameters are listed in Appendix E.3.

**WG+Text (Pure Text k-NN; HEXACO-free baseline).** The override fires when the unsafe ratio in the text-embedding $K$-neighborhood exceeds $r$:

$$\rho_{\text{T}}(\mathbf{t}) = \mathbf{1}\left[\frac{1}{K}\sum_{k=1}^{K}\mathbf{1}[\text{neighbor}_k \text{ is unsafe}] \geq r\right]. \tag{6}$$

This HEXACO-free baseline attains AEGIS F1=0.844 (+0.040 over WildGuard).

**WG+HDist (Hybrid k-NN; HEXACO in distance, label-free[1]).** We augment the text embedding with HEXACO factor scores and search in the resulting hybrid space:

$$\mathbf{h}_{\text{hybrid}} = \left[\mathbf{e}(\mathbf{t})/\|\mathbf{e}(\mathbf{t})\|,\ \ (\mathbf{f}(\mathbf{t})-1)/4\cdot\alpha/(1-\alpha)\right], \tag{7}$$

where $\mathbf{f}(\mathbf{t})$ is the HEXACO factor vector and $\alpha \in (0, 1)$ controls the relative HEXACO weight (Appendix D.5 for the affine rescaling and value-range note). The override fires under the same neighbor-unsafe-ratio rule (Eq. 6) but applied to neighbors retrieved in the hybrid space. Under validation-tuned hyperparameters, this configuration attains AEGIS F1=0.841 (+0.037) and WG-Mix F1=0.445 (+0.023); WG+HDist scores 0.003 below the WG+Text baseline and a McNemar test (McNemar, 1947) does not detect a significant difference at this sample size ($p = 0.423$).

**WG+HGate (k-NN Category Proxy + HEXACO AND; label-free).** Instead of injecting HEXACO into the distance, we use the same text k-NN to *proxy* the sample's harm category and apply HEXACO as a confidence threshold. Let $\mathcal{C}_{\text{hi}}$ denote a set of categories where HEXACO most effectively complements WildGuard, derived from a held-out rescue-efficiency analysis on the AEGIS2 train validation split (Appendix D.5). The override fires when the high-efficiency-category proportion in the $K$-neighborhood exceeds $r_{\text{cat}}$ and HEXACO

[1]Throughout this section, *label-free* indicates that an integration requires no ground-truth harm-category label for the test sample at inference; training-pool labels remain available for $k$-NN retrieval.

Table 4: Validation-set threshold tuning for the Learning-based Decision Model (Enc-E). Training set: 1:1 under-sampled; validation set: 5,123 samples (741 unsafe).

| Model | Config | $\tau^*$ | val Prec. | val Recall | val F1 | test F1 |
|---|---|---|---|---|---|---|
| oax Logistic Regression | $C = 0.01$, features {O,A,X} | 0.16 | 0.250 | 0.984 | 0.399 | 0.651 |
| LightGBM (sweep best) | $K = 600$, $\lambda = 1.0$, sub=0.7 | 0.72 | 0.448 | 0.692 | 0.544 | 0.587 |

Table 5: Detection results on AEGIS2 test split (1,964 samples; 394 unsafe / 1,570 safe) using the best decision model established in RQ2; HEXACO Detector uses oax LogReg ($C$=0.01, Enc-E).

| Model | Prec. | Recall | F1 | Acc. |
|---|---|---|---|---|
| OpenAI Moderation | 0.859 | 0.510 | 0.640 | 0.885 |
| LlamaGuard | 0.855 | 0.525 | 0.651 | 0.887 |
| WildGuard | 0.828 | 0.782 | 0.804 | 0.924 |
| **HEXACO Detector** (oax LogReg) | 0.489 | **0.975** | 0.651 | 0.791 |

is sufficiently confident:

$$\rho_{\mathrm{G}}(\mathbf{t}) = \mathbf{1}\left[\frac{1}{K}\sum_{k=1}^{K}\mathbf{1}[c(\text{neighbor}_k) \in \mathcal{C}_{\mathrm{hi}}] \geq r_{\mathrm{cat}}\right] \cdot \mathbf{1}[p_{\mathrm{hex}}(\mathbf{t}) \geq \beta]\,. \tag{8}$$

Under validation-tuned hyperparameters, this configuration converges to essentially WildGuard's behaviour on all three datasets; the strict val-tuned $r_{\mathrm{cat}}$ causes the override to fire rarely, so the HEXACO AND condition does not provide an independent gain once $\mathcal{C}_{\mathrm{hi}}$ is derived without test-set leakage.

**WG+HOracle (GT Category + HEXACO Gate; oracle upper bound).** WG+HOracle replaces the $k$-NN-based category proxy in Eq. 8 with the ground-truth harm category $c^*(\mathbf{t})$, using a test-derived $\mathcal{C}_{\mathrm{hi}}^{\mathrm{test}}$ (Appendix D.5):

$$\rho_{\mathrm{O}}(\mathbf{t}) = \mathbf{1}[\,c^*(\mathbf{t}) \in \mathcal{C}_{\mathrm{hi}}\,] \cdot \mathbf{1}[\,p_{\mathrm{hex}}(\mathbf{t}) \geq \beta\,]\,. \tag{9}$$

This establishes an upper bound on the value of category information, attaining AEGIS F1=0.859, the strongest result; it should be interpreted as an oracle upper bound, since $\mathcal{C}_{\mathrm{hi}}^{\mathrm{test}}$ was identified on the test split (see §6.3.2).

***Finding (RQ4):*** *Under val-tuned hyperparameters, WG+Text and WG+HDist achieve the largest AEGIS gains (+0.040, +0.037) but lose 0.023–0.038 on XSTest; WG+HDist sits 0.003 below the HEXACO-free WG+Text baseline and the difference is not detected as significant by a McNemar test.*

Table 6: RQ4: WildGuard–HEXACO integration strategies on AEGIS2 test and OOD datasets. Hyperparameters for the three label-free integrations are tuned on a held-out AEGIS2 train validation split (see Section 4.1); WG+HGate's high-efficiency category set $\mathcal{C}_{\mathrm{hi}}$ is derived from the same val split (Appendix Table 16). F1 (AEGIS) gain over WildGuard shown in parentheses. †WG+HOracle is an oracle upper bound requiring ground-truth category labels at inference and uses the test-derived $\mathcal{C}_{\mathrm{hi}}^{\mathrm{test}}$ (Appendix Table 15; see Limitations §6.3.2). Per-dataset Precision, Recall, Accuracy, and confusion-matrix entries are reported in Appendix Table 14.

| Method | F1 (AEGIS) | F1 (WGMix) | F1 (XSTest) |
|---|---|---|---|
| WildGuard | 0.804 | 0.422 | **0.523** |
| WG+Text | **0.844** (+0.040) | **0.448** | 0.500 |
| WG+HDist | 0.841 (+0.037) | 0.445 | 0.485 |
| WG+HGate | 0.806 (+0.002) | 0.419 | 0.517 |
| WG+HOracle† | 0.859 (+0.055) | 0.422 | 0.523 |

## 5 Conclusion

We proposed a HEXACO-based framework for unsafe LLM response detection, pairing a local Gemma-3-270M-IT personality estimator with an oax Logistic Regression classifier and a hybrid-k-NN ensemble with WildGuard. Experiments on AEGIS2 confirm that HEXACO scores carry consistent safety signal: Agreeableness and Extraversion are globally depressed for unsafe responses, and Openness is consistently elevated within all 24 harm categories. The HEXACO Detector (oax LogReg, Recall=0.975, F1=0.651) functions as a high-recall booster. Among four WildGuard–HEXACO integrations, WG+Text and WG+HDist improve AEGIS and WildGuardMix F1 over WildGuard but reduce XSTest, with WG+HDist reaching 0.841, which is statistically indistinguishable from the HEXACO-free WG+Text baseline (0.844, McNemar $p = 0.423$); WG+HGate behaves almost identically to WildGuard once its high-efficiency category set is derived on training data. These results position HEXACO as a compact, interpretable substitute for high-dimensional text retrieval rather than an independently dominant signal, and cross-distribution robustness remains an open challenge.

# 6 Limitations

## 6.1 HEXACO Detector Calibration and Decision Model

### 6.1.1 Estimator Score Distribution Drift

The HEXACO estimator's score distribution depends on architecture and training configuration and is not anchored to the 1–5 personality scale in an absolute sense. Between Enc-A (linear head) and Enc-E (MLP head), the global score distributions shifted substantially (Openness factor, population-level: $\Delta_o = -0.37$ for Enc-A vs. $\Delta_o = -0.11$ for Enc-E, both null-label-excluded values), and the default threshold $\tau = 1.0$ yielded recall=0.000 with Enc-A and recall=0.178 with Enc-E. Note that the within-category Openness signal ($\Delta_o > +0.50$ in 24/24 harm categories; see Section 4.5) is robust to this global drift, but absolute thresholds tuned on one estimator do not transfer to another. Any production deployment must therefore recalibrate thresholds whenever the estimator is updated, imposing ongoing operational overhead.

### 6.1.2 Decision Model Generalization and Recall–Precision Trade-off

The Decision Model parameters are optimized on AEGIS2, and their generalizability to other distributions has not been validated. This concern applies particularly to non-English or domain-specific datasets. The oax Logistic Regression achieves near-perfect recall (0.975) at the cost of a relatively large number of false positives (TP=384, FP=401, FN=10, TN=1,169 on the AEGIS2 test split; derivable from Precision=0.489 and Recall=0.975 in Table 5), a consequence of F1-maximizing threshold selection on an imbalanced validation set. Practical deployments may instead require precision-constrained optimization, for example, maximizing recall subject to precision $\geq 0.55$ or using $F_\beta$ with $\beta < 1$.

### 6.1.3 Inconsistent Treatment of Null-Label Rows

The AEGIS2 train split includes 14,773 null-label human-turn rows that the label pipeline normalises as safe; we exclude these from RQ1's per-factor analysis but retain them for Decision Model training and ensemble tuning, inflating the validation false-positive tail (Appendix D.3). A cleaner protocol would exclude null-label rows from all stages or treat empty-response detection as a separate task; we defer this re-experiment to future work.

## 6.2 Generalization and Robustness

### 6.2.1 Adversarial Polite Inputs

The proposed method rests on the observation that unsafe responses exhibit systematically distinct HEXACO profiles: Agreeableness and Extraversion are depressed at the global level, while within harm categories, Openness is elevated and Conscientiousness is depressed relative to safe refusals. RQ1 provides empirical support for this on AEGIS2. However, adversarially crafted inappropriate content written in a polite and agreeable tone may not trigger the expected pattern and could therefore evade detection. Robustness to adversarial inputs is an important open problem.

### 6.2.2 Hyperparameter Selection Protocol

For the three label-free integrations (WG+Text, WG+HDist, WG+HGate), hyperparameters $(K, r)$, $(\alpha, K, r)$, and $(K, r_{\text{cat}}, \beta)$ are tuned on a held-out 20% hash-based validation split of the full 30,007-row AEGIS2 train set (the same split used for the Decision Model in RQ2). The split is computed first; each experiment then restricts the validation pool to the rows with the required valid predictions: the Decision Model uses the 5,123 rows reported in Table 4, while ensemble tuning further restricts to the $n = 1{,}997$ rows that have both non-empty WildGuard predictions (the 2,161-row valid-WG subset; Appendix Table 16) and cached text embeddings (164 valid-WG rows are excluded due to an upstream batching constraint). We maximise F1 on this validation set and apply the resulting single configuration unchanged to AEGIS2 test, WildGuardMix, and XSTest. This protocol avoids the test-set tuning we used in earlier drafts. While per-dataset hyperparameter tuning (a separate sweep per test split) could yield higher OOD F1 on each split individually, that protocol corresponds to test-set leakage on the OOD splits and we therefore report only the held-out single-config values in Table 6. A residual concern is that the validation split is drawn from the AEGIS2 train distribution; hyperparameters that are optimal on AEGIS2 may not be optimal on OOD distributions whose category and stylistic mix differ.

### 6.2.3 Out-of-Domain Generalization

The four-detector majority-vote ensemble (Appendix E.1) achieves AEGIS F1=0.828 but exhibits a sharp distribution-dependent profile: WG-

Mix F1=0.449 (approximately 46% relative drop from AEGIS, still slightly above WildGuard) and XSTest F1=0.444 ($-0.079$ vs. WildGuard), revealing AEGIS-specific overfitting that is not robust across OOD distributions. For the WildGuard–HEXACO integrations of Table 6, the val-tuned protocol (§6.2.2) yields the following OOD picture: WG+Text and WG+HDist retain a small positive margin on WildGuardMix ($+0.026$ and $+0.023$) but degrade on XSTest ($-0.023$ and $-0.038$); WG+HGate, with its honestly train-derived $\mathcal{C}_{\text{hi}}$ (Appendix Table 16) and strict val-tuned $r_{\text{cat}} = 0.9$, applies its override rarely and converges to essentially WildGuard's behaviour on all three datasets (within $\pm 0.006$ F1). WG+HOracle reduces to WildGuard on both OOD splits because their taxonomies do not overlap with the test-derived high-efficiency category set $\mathcal{C}_{\text{hi}}^{\text{test}}$. The standalone oax Logistic Regression decision model degrades substantially on OOD data (WGMix F1=0.293, XSTest F1=0.290 vs. AEGIS F1=0.651; Appendix Table 11), which motivates its use within an ensemble rather than as a standalone detector. Overall, hyperparameter optima identified on AEGIS do not transfer uniformly across distributions, leaving cross-distribution robustness an open challenge for the proposed framework.

#### 6.2.4 Sanity Check on Unsafe-Prompt Benchmarks

As a sanity check on adversarial inputs, we run detectors on two unsafe-prompt benchmarks: HarmBench standard (Mazeika et al., 2024) (200 unsafe prompts) and the adversarial subset of WildGuardMix test (Han et al., 2024) (796 samples; 131 unsafe / 665 safe, filtered by the `adversarial` flag). Since HarmBench provides only prompts, these numbers reflect prompt-side detection rather than the response-level robustness that the rest of the paper targets. Table 7 summarizes the results.

On HarmBench, WildGuard and the val-tuned WG+HGate both miss the same four cybercrime/illegal prompts, while the HEXACO Detector captures all 200, providing additional evidence of complementary recall when used as a standalone high-recall booster. On WildGuardMix adversarial, all three methods degrade substantially in F1 compared to AEGIS2, but the $F_2$ ranking reverses: the HEXACO Detector attains the highest $F_2$=0.501 with a miss rate of 3.1%, versus WildGuard's $F_2$=0.444 with a 46.6% miss rate. WG+HGate behaves essentially identically to WildGuard on this adversarial subset (zero rescued true positives against two added false positives), consistent with its near-WildGuard behaviour observed on AEGIS2, WildGuardMix and XSTest in Table 6: under the val-tuned $r_{\text{cat}} = 0.9$, the selective-complementation override rarely fires under adversarial-prompt distributions. These mixed results indicate a deployment-dependent recommendation: the HEXACO Detector's near-perfect recall is the appropriate operating point for safety-critical applications where missing a harmful response is unacceptable, whereas WildGuard remains the strongest single detector on AEGIS2 for balanced traffic; a more comprehensive evaluation covering attack-success rate on LLM-generated responses, gradient-based attacks, and roleplay jailbreaks is left for future work.

#### 6.2.5 Multilingual and Cross-Cultural Applicability

All evaluations are conducted on English-language datasets (AEGIS2, WildGuardMix, XSTest). The HEXACO model originates from English-language personality inventories, and the LLM-as-a-Judge prompts used for factor scoring are written in English. The validity of the HEXACO-based detection approach for non-English text (e.g., Japanese, Chinese, Arabic) and across culturally diverse safety norms has not been validated. Future work should evaluate the method on multilingual benchmarks.

### 6.3 Statistical and Methodological Caveats

#### 6.3.1 Small-Sample Categories and Statistical Significance

The AEGIS2 test split contains 394 unsafe samples across 24 harm categories, yielding an average of approximately 16 unsafe samples per category. Several rescue efficiency scores are $\infty$ (e.g., Guns/IW, Suicide/SH: 3 rescued, 0 added FPs), which reflects the expectation under small $n$ rather than a robust empirical finding. No bootstrap confidence intervals or McNemar tests are reported for per-category comparisons. Results for low-frequency categories should be interpreted with caution.

#### 6.3.2 WG+HOracle Test-Set Leakage and Category-Classifier Dependency

The oracle WG+HOracle (AEGIS F1=0.859) requires ground-truth category labels at inference time and so cannot be deployed as-is. Two limitations apply. First, the set of *high-efficiency*

Table 7: Jailbreak robustness on HarmBench (200 unsafe prompts) and WildGuardMix adversarial subset (131 unsafe / 665 safe). Detection = recall on HarmBench (all-unsafe set); F1, Recall, and $F_2$ are reported on WGMix adversarial. $F_2$ weights recall over precision and is more appropriate when miss-cost dominates false-positive-cost. WG+HGate uses the val-tuned configuration of Table 6.

| Method | HarmBench Det. | WGMix F1 | WGMix Recall | WGMix $F_2$ |
|---|---|---|---|---|
| WildGuard | 0.980 (196/200) | **0.354** | 0.534 | 0.444 |
| HEXACO Detector | **1.000** (200/200) | 0.291 | **0.970** | **0.501** |
| WG+HGate | 0.980 (196/200) | 0.352 | 0.534 | 0.443 |

categories used in WG+HOracle's override rule ({Guns/IW, Suicide/SH, Controlled Substances, Profanity, Harassment}) was identified using the rescue-efficiency analysis on the AEGIS2 test split itself (Appendix Table 15), introducing test-set leakage; the reported AEGIS F1=0.859 should therefore be interpreted as an oracle upper bound, not a deployable result. For comparison, we separately report a held-out, train-derived rescue-efficiency analysis (Appendix Table 16), which is the basis for WG+HGate's $\mathcal{C}_{\text{hi}}$; on that scale no primary category exceeds efficiency 1.0, indicating that the dramatic test-derived efficiencies ($\infty$, 4.00, 3.00, 2.00) were small-sample artifacts. Second, WG+HOracle presupposes the availability of an accurate category classifier at inference time; in a deployed system the realised gain would depend directly on the classifier's accuracy, and a misclassified category could either suppress a valid override or trigger a spurious one. WG+HOracle is included in our comparison to quantify the headroom that category information could unlock; it motivates future work on training a category classifier and rederiving the high-efficiency category set on the training split, rather than serving as a final deployment recommendation.

## 6.4 Deployment Cost and Reproducibility

### 6.4.1 LLM-as-a-Judge Latency and Cost

The HEXACO+GPT variant relies on LLM-as-a-Judge (Zheng et al., 2023), which introduces inference latency and API cost proportional to the number of factors estimated per sample. The local HEXACO estimator addresses this by eliminating the external API dependency, but its estimation quality may differ from that of the LLM-as-a-Judge approach. A systematic comparison of the two estimation strategies in terms of accuracy, latency, and robustness is left for future work.

### 6.4.2 HEXACO Detector Inference Cost

The local HEXACO Detector (Enc-E) requires six separate model instances, one per HEXACO factor, each using Gemma-3-270M-IT (Gemma Team, 2025) as backbone (total ≈1.62B parameters). While the six models can be run in parallel, the memory and orchestration overhead is non-trivial compared with a single-model deployment such as WildGuard (Mistral-7B, ≈7B parameters, one inference call). The paper does not provide latency or throughput benchmarks, and actual deployment cost depends on hardware and batching strategy. A future direction is to share a single backbone across all six factor heads, which could reduce total parameters to a single 270M model with six lightweight regression heads.

### 6.4.3 Proprietary API Version Drift

The OpenAI Moderation baseline uses `omni-moderation-latest` (as of May 19, 2026), and HEXACO+GPT factor scoring uses `gpt-5.4-nano`. These APIs are subject to provider-side updates that may alter outputs without notice, making exact numerical reproduction impossible after the API versions change. Replication studies should pin specific API versions or use the local HEXACO estimator variant.

## 6.5 Potential Risks

While the technical limitations above quantify performance trade-offs, several deployment-side risks warrant explicit acknowledgement. First, the high-recall operating point (precision $\approx$ 0.49) leads to substantial over-blocking, with approximately 25% of safe samples flagged. Such over-blocking can suppress legitimate interactions and erode user trust. Second, treating depressed Agreeableness and Extraversion as safety signals risks systematic disadvantage to users whose communication style is naturally direct or blunt, including neurodiverse users and speakers from cultures that value directness; personality features are correlations with unsafe content rather than morally relevant traits, and should not be used to deny service in isolation. Third, the locally fine-tuned HEXACO estimator is

a dual-use artifact: while developed here for safety classification, the same model could be repurposed for personality profiling in employee monitoring, targeted manipulation, or social scoring, and downstream releases should restrict permissible uses accordingly. Adversarial polite-but-harmful inputs and the English-only, Western-derived nature of HEXACO (as discussed above) further imply that this detector should not serve as a standalone safety arbiter, particularly in safety-critical or multilingual production settings.

## References


Michael C. Ashton and Kibeom Lee. 2007. Empirical, theoretical, and practical advantages of the hexaco model of personality structure. *Personality and Social Psychology Review*, 11(2):150–166.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. 2025. AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. In *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)*, pages 5992–6026, Albuquerque, New Mexico. Association for Computational Linguistics.

Lewis R. Goldberg. 1990. An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, and Madian Khabsa. 2023. Llama guard: LLM-based input-output safeguard for human-AI conversations. *Preprint*, arXiv:2312.06674.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. SALAD-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3923–3954, Bangkok, Thailand. Association for Computational Linguistics.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4694–4702, Singapore. Association for Computational Linguistics.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*

*and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *Preprint*, arXiv:2402.04249.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Traian Rebedea, Razvan Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. 2023. NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 431–445, Singapore. Association for Computational Linguistics.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Shuo Wang, Renhao Li, Xi Chen, Yulin Yuan, Min Yang, and Derek F. Wong. 2025. Exploring the impact of personality traits on LLM bias and toxicity. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4125–4143, Suzhou, China. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. 2024. ShieldGemma: Generative AI content moderation based on Gemma. *Preprint*, arXiv:2407.21772.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

# A HEXACO Factors, Facets, and Negative Descriptors

Table 8 lists the facets and negative descriptors for each HEXACO factor, used as reference terms in the LLM-as-a-Judge prompts for personality estimation.

# B Prompt for Personality Estimation

Figure 2 shows the template provided to the LLM for factor-score estimation in HEXACO-based inappropriate-text detection. In the template, the placeholders "`[[Factor Name]]`", "`[[Facet Description]]`", "`[[Positive Descriptor List]]`", and "`[[Negative Descriptor List]]`" are filled in with the corresponding information for each of the six HEXACO factors. We instantiate one prompt per factor from this template and query the LLM to obtain each factor score.

```
Please determine how well the given
    input sentence matches the HEXACO
    factor
"[[Factor Name]]".

[Procedure]
1. Read the input sentence and evaluate
    it according to the facets of
   [[Factor Name]].
   [[Facet Description]]
2. If positive expressions are
    observed, increase the score:
   [[Positive Descriptor List]]
3. If negative expressions are
    observed, decrease the score:
   [[Negative Descriptor List]]
4. Output a score from 1 to 5. A higher
    score means better alignment with
   [[Factor Name]]. If neither positive
       nor negative expressions are
       observed,
   output 3.

[Output format]
Return the result in the following JSON
    format:

{
  "category": "[[Factor Name]]",
  "score": <1~5>
}
```

Figure 2: Prompt template for HEXACO factor estimation. Placeholders are filled in with factor-specific descriptors at inference time, yielding one prompt per HEXACO factor.

# C Experimental Setup Details

## C.1 HEXACO Estimator Training Data

The local HEXACO estimator (Enc-E) is trained on synthetic HEXACO-conditioned responses generated independently of AEGIS2. We use the *challenge* subset of RealToxicityPrompts (Gehman et al., 2020) (1,199 unique source prompts) as conditioning context. For each source prompt $p$, we query `gpt-5.4-nano` (temperature $= 0$, `max_tokens`$= 256$, seed $= 42$) under each of 6 HEXACO factors $\times$ {high, low} polarity persona descriptions, plus one unconditioned "base" persona, yielding 13 response records per source prompt. This produces $1{,}199 \times 13 = 15{,}587$ generated training records. We partition these records by source prompt into 80% train, 10% dev, and 10% test using a deterministic hash on the source-prompt identifier; the dev split is used for early stopping and intermediate diagnostics, and the held-out test split is used for the head-architecture ablation reported in main-text Table 2. The estimator is then fine-tuned with a contrastive rank-margin objective (margin 1.5, rank loss weight 0.5, regression loss weight 1.0) so that, for the same source prompt, the response generated under the high-polarity persona for factor $i$ receives a higher predicted score on $M_i$ than the response generated under the low-polarity persona for $i$. The Enc-E variant adopted in all subsequent experiments uses an MLP head over the Gemma-3-270M-IT (Gemma Team, 2025) backbone: LayerNorm $\rightarrow$ Linear($640 \rightarrow 128$) $\rightarrow$ GELU $\rightarrow$ Dropout $\rightarrow$ Linear($128 \rightarrow 1$), with the backbone and head jointly fine-tuned for 50 epochs at discriminative learning rates ($2 \times 10^{-5}$ for the backbone, $3 \times 10^{-4}$ for the head). Importantly, this training data contains no AEGIS2 examples and no safety labels, and AEGIS2 test is not used for any estimator-side model selection: the Enc-E architecture is chosen on the synthetic held-out test split described above, and AEGIS2 in RQ1–RQ4 is reserved for the downstream safety-detection evaluation as a strict zero-shot transfer.

## C.2 Detection Models

Table 9 lists all detection models evaluated in this work. OpenAI Moderation refers to OpenAI's Moderation API (omni-moderation-latest), and Llama-Guard and WildGuard are LLM-based classifiers run locally. HEXACO+GPT applies the LLM-as-a-Judge prompting strategy to estimate HEXACO factors before classification. LLM-Judge directly

Table 8: HEXACO factors, facets, and negative descriptors (Ashton and Lee, 2007) used in LLM-as-a-Judge prompts.

| Factor | Facets | Negative Descriptors |
|---|---|---|
| Honesty-Humility (H) | Sincerity, fairness, greed-avoidance, modesty | Cunning, deceptive, greedy, pretentious, hypocritical, boastful, arrogant, conceited, self-centered |
| Emotionality (E) | Fearfulness, anxiety, dependence, sentimentality | Tough, fearless, unemotional, highly independent, overly assertive, unfriendly, insensitive |
| Extraversion (X) | Social self-esteem, social boldness, sociability, liveliness | Shy, passive, withdrawn, introverted, quiet, reserved, inhibited, gloomy |
| Agreeableness (A) | Forgiveness, gentleness, flexibility, patience | Bad-tempered, quick to fight, stubborn, irritable, short-tempered, obstinate, blunt |
| Conscientiousness (C) | Organization, diligence, perfectionism, prudence | Careless, neglectful, reckless, lazy, irresponsible, absent-minded, sloppy |
| Openness (O) | Aesthetic appreciation, inquisitiveness, creativity, unconventionality | Superficial, simplistic, unimaginative, conventional, closed-minded |

Table 9: Detection models used in experiments.

| Name | Model / Service | Parameter scale |
|---|---|---|
| OpenAI Moderation | OpenAI / omni-moderation-latest (as of May 19, 2026) | API model (not disclosed) |
| LlamaGuard (Dubey et al., 2024) | Meta / Llama-3.1-8B class (local) | 8B |
| WildGuard (Han et al., 2024) | AllenAI / Mistral-7B fine-tune (local) | 7B |
| HEXACO+GPT | OpenAI / gpt-5.4-nano | API model (not disclosed) |
| LLM-Judge | OpenAI / gpt-5.4-nano | API model (not disclosed) |
| **HEXACO Detector** | Gemma-3-270M-IT (Gemma Team, 2025) $\times$6 (Enc-E) + oax LogReg (local) | $\approx$1.62B + small head |

queries gpt-5.4-nano for a binary safety judgment without explicit HEXACO reasoning. The HEXACO Detector uses a locally fine-tuned text encoder (Enc-E: MLP head, 50 epochs, discriminative learning rates; see RQ2 in the main paper) trained with contrastive learning.

### C.3 Datasets

Table 10 summarizes all datasets. The AEGIS2 train split is used for HEXACO signal analysis (RQ1), Learning-based Decision Model training with 1:1 class balancing, and ensemble weight estimation. The AEGIS2 test split is strictly held out for all RQ2–RQ4 evaluations. WildGuardMix (Han et al., 2024) contains adversarially curated prompts spanning diverse harm types; XSTest (Röttger et al., 2024) focuses on over-refusal cases and safe prompts that models commonly misclassify. Both OOD datasets are used exclusively for generalization evaluation (RQ4).

### C.4 Validation Splits for Decision Model and Ensemble Tuning

The 20% hash-based validation split is applied to all 30,007 AEGIS2 training samples; each downstream experiment then restricts the resulting validation pool to the rows with the predictions it requires. For the Decision Model (RQ2), the validation subset comprises the 5,123 rows reported in Table 4 (741 unsafe; null-label human-turn entries are labelled safe by the normalisation pipeline and retained when the Decision Model produces a valid prediction). For ensemble tuning (RQ4), this validation subset is further restricted to the $n = 1{,}997$ rows that have both non-empty WildGuard predictions (the 2,161-row valid-WG subset used for the train-derived rescue-efficiency analysis in Table 16) and cached text embeddings (287 unsafe; 164 valid-WG rows lack cached embeddings due to an upstream batching constraint; see Limitations §6.2.2). Ensemble hyperparameters $(K, r)$ for WG+Text, $(\alpha, K, r)$ for WG+HDist, and $(K, r_{\text{cat}}, \beta)$ for WG+HGate are selected to maximise F1 on this validation subset and applied unchanged to AEGIS2 test, WildGuardMix, and XSTest.

## D Supplementary Methodological Notes

### D.1 Threshold Direction Heterogeneity

The threshold-based decision rule (Eq. 1) imposes a monotone "low score $\Rightarrow$ unsafe" assumption. The empirical HEXACO signal partially supports this: unsafe responses show lower Agreeableness and Extraversion (consistent with the rule), but also

Table 10: Datasets used in experiments. AEGIS2 train (30,007) includes 15,234 rows with explicit `response_label` (3,541 unsafe / 11,693 safe) and 14,773 human-turn entries (`response_label=null`, no LLM output), which are treated as safe (`normalized_label=False`) by the label normalization pipeline.

| Dataset | Total | Unsafe | Unsafe% | Use |
|---|---|---|---|---|
| AEGIS2 train | 30,007 | 3,541 | 11.8% | Train (RQ1, RQ2, RQ4) |
| AEGIS2 test | 1,964 | 394 | 20.1% | Eval (RQ2–RQ4) |
| WildGuardMix test | 1,725 | 284 | 16.5% | OOD eval (RQ4) |
| XSTest | 446 | 78 | 17.5% | OOD eval (RQ4) |

higher Conscientiousness globally and elevated Openness in the within-harm-category comparison. This directional restriction caps the achievable threshold-based F1 and motivates the unconstrained Learning-based Decision Model used in subsequent experiments.

### D.2 Within-Category vs. Global HEXACO Interpretation

In Table 1 the global Openness difference is slightly negative ($\Delta_o = -0.11$), whereas the within-harm-category comparison reported in §RQ4 shows $\Delta_o > +0.50$ in 24 of 24 categories. The two analyses ask different questions: the global comparison aggregates over heterogeneous safe content (refusals, neutral chat, null-label human turns), whereas the within-category comparison contrasts unsafe compliance against safe refusals *within* a fixed harm topic. The apparent sign reversal is therefore a Simpson's-paradox-style effect rather than an inconsistency.

### D.3 Validation–Test F1 Discrepancy in RQ2

The validation–test F1 gap for the oax LogReg (val F1 $= 0.399 \rightarrow$ test F1 $= 0.651$ at $\tau^* = 0.16$; Table 4) reflects a distributional difference between the two pools. The AEGIS2 train split (and therefore the validation subset) includes 14,773 null-label human-turn rows that the normalisation pipeline labels safe but which carry no LLM response, producing HEXACO scores with a wider false-positive tail. The AEGIS2 test split contains no such null-label rows, so the same threshold yields substantially higher precision ($0.250 \rightarrow 0.489$) at the same recall ($\approx 0.98$) and the F1 jumps accordingly.

### D.4 RQ4 Pre-analyses and Selective Complementation Setup

**Rescue efficiency.** To quantify complementarity between HEXACO and WildGuard, we measure *rescue efficiency*: HEXACO-rescued TPs per added FP, category-wise (full per-category breakdown in Table 15). Two categories (Guns/IW + Criminal Planning, Suicide/Self-Harm) show $\infty$ efficiency, and three additional categories (Controlled Substances, Profanity, Harassment) achieve efficiency $\geq 2.0$ on the test set, meaning each added FP recovers at least two missed unsafe samples.

**Within-category factor profile.** A complementary analysis contrasts unsafe compliance against safe refusals to the same prompt category. It reveals two consistent patterns: Openness is elevated in all 24 harm categories ($\Delta_o > +0.50$ in 24/24), and Conscientiousness is depressed in most categories ($\Delta_c < -0.30$ in 17/24), suggesting an *open but irresponsible* compliance profile.

**Formal setup.** We retain WildGuard's positive decisions ($y_{\text{WG}} = M_{\text{WG}}(\mathbf{t}) = 1$) and override only its negative (safe) decisions ($y_{\text{WG}} = 0$) via $y_{\text{ensemble}} = y_{\text{WG}} \vee \rho(\mathbf{t})$ for an override predicate $\rho$. We consider four predicates spanning the design space of label requirement (none, $k$-NN-based proxy, or ground-truth category) and HEXACO usage (none, in distance, or as AND-gate), where the HEXACO-based unsafe probability $p_{\text{hex}}(\mathbf{t})$ is obtained from the oax Logistic Regression decision model (Section 3.2.2) and the $K$ nearest neighbors are retrieved from the AEGIS2 training pool using `multilingual-e5-large` embeddings (Wang et al., 2024).

### D.5 RQ4 Integration Strategy Implementation Details

**WG+HDist (hybrid k-NN).** The hybrid embedding (Eq. 7) uses $\mathbf{f}(\mathbf{t}) \in \mathbb{R}^6$, the raw HEXACO factor vector trained against 1–5 targets but not strictly clipped (see Table 1); the affine map $(\mathbf{f}(\mathbf{t}) - 1)/4$ is applied as a coarse rescaling toward $[0, 1]$. The McNemar test reported in §RQ4 yields $p = 0.423$; we stop short of claiming formal equivalence between WG+HDist and WG+Text, which would require

a pre-registered margin and either an equivalence test or paired-bootstrap confidence intervals.

**WG+HGate (k-NN category proxy + HEXACO AND).** The high-efficiency category set $\mathcal{C}_{\text{hi}}$ is derived from a held-out rescue-efficiency analysis on the AEGIS2 train validation split (Table 16); we select the top-5 primary categories by efficiency: {Guns/IW, Sexual, Profanity, Controlled/Regulated Substances, Criminal Planning/Confessions}. The validation-tuned hyperparameters are $K = 30$, $r_{\text{cat}} = 0.9$, and $\beta = 0.1$, yielding AEGIS F1=0.806 ($+0.002$), WGMix F1=0.419 ($-0.003$), and XSTest F1=0.517 ($-0.006$). The strict $r_{\text{cat}} = 0.9$ causes the override to fire rarely, so WG+HGate effectively reduces to WildGuard on all three datasets.

**WG+HOracle (GT category + HEXACO gate).** WG+HOracle has the same AND structure as WG+HGate (Eq. 8) but replaces the $k$-NN-based category proxy with the ground-truth harm category $c^*(\mathbf{t})$; the ensemble combines it with WildGuard via $y_{\text{ensemble}} = y_{\text{WG}} \vee \rho_{\text{O}}(\mathbf{t})$. The override fires when $c^*$ lies in the test-derived high-efficiency set $\mathcal{C}_{\text{hi}}^{\text{test}}$ = {Guns/IW + Criminal Planning, Suicide/Self-Harm, Controlled Substances, Profanity, Harassment} (Table 15) and HEXACO is sufficiently confident.

# E Supplementary Experimental Results

## E.1 Four-Detector Multi-Detector Aggregation Results (RQ4)

Table 11 reports the multi-detector weighted aggregation results when all four detectors (OpenAI Moderation, LlamaGuard, WildGuard, and the HEXACO Detector) are combined. We sweep weighting hyperparameters (k-NN: $S \in \{1, 5, 10, 20\}$; clustering: $K \in \{5, 10, 20, 50\}$) with weighting metric $m \in$ {F1, Recall, Precision} on the AEGIS2 train split, and report the best variant per ensemble strategy. For completeness, we additionally include a majority-vote baseline ($\geq 2/4$ detectors flag unsafe), an OR-vote baseline ($\geq 1/4$), and a Gate meta-learner that uses HEXACO factor scores as additional features.

Among the three weighting strategies introduced in Section 3.3, the embedding-based clustering variant reaches the highest AEGIS2 F1 (0.759), and the k-NN variant generalizes the best to OOD (WildGuardMix F1=0.608 and XSTest F1=0.720). The simpler majority-vote baseline attains the strongest in-domain F1 (0.828) but degrades sharply on OOD data (WGMix 0.449), revealing AEGIS-specific overfitting of the broad consensus signal. None of these multi-detector aggregations improve substantially over WildGuard alone (AEGIS F1=0.804), which motivates the focused selective-complementation approach reported in the main paper (Section 4.5).

Table 11: Four-detector ensemble results on AEGIS2 test, WildGuardMix test, and XSTest. All ensembles aggregate four weak detectors (OpenAI Moderation, LlamaGuard, WildGuard, HEXACO Detector). Best variants reported per ensemble strategy: k-NN ($S$=20, weight=F1); Clustering ($K$=5, weight=Precision); Gate (HEXACO factor features + meta-LogReg, $C$=0.1); Majority Vote ($\geq 2/4$); OR Vote ($\geq 1/4$). Single-detector results are shown for reference.

| Method | AEGIS F1 | WGMix F1 | XSTest F1 |
|---|---|---|---|
| OpenAI Moderation | 0.640 | 0.483 | 0.653 |
| LlamaGuard | 0.651 | 0.485 | 0.620 |
| WildGuard | 0.804 | 0.422 | 0.523 |
| HEXACO Detector | 0.651 | 0.293 | 0.290 |
| Embedding-based k-NN (best) | 0.746 | **0.608** | **0.720** |
| Embedding-based Clustering (best) | 0.759 | 0.580 | 0.638 |
| Gate meta-learner (HEXACO) | 0.763 | 0.555 | 0.650 |
| Majority Vote ($\geq 2/4$) | **0.828** | 0.449 | 0.444 |
| OR Vote ($\geq 1/4$) | 0.659 | 0.286 | 0.302 |

## E.2 LLM-based Baselines and Detailed Results (RQ3)

Table 12 presents results for LLM-based detection approaches, and Table 13 provides full confusion matrices for all RQ3 models.

HEXACO+GPT uses an LLM-as-a-Judge prompt to estimate HEXACO factors and then applies a fixed detection threshold. LLM-Judge directly queries gpt-5.4-nano for a safety classification. Both approaches achieve near-perfect recall (>0.97) at the cost of very low precision (≈0.20–0.22).

Table 12: LLM-based baselines on AEGIS2 test split (default settings).

| Model | Prec. | Recall | F1 | Acc. |
|---|---|---|---|---|
| HEXACO+GPT | 0.204 | 0.990 | 0.339 | 0.220 |
| LLM-Judge | 0.220 | 0.977 | 0.360 | 0.302 |

## E.3 WildGuard–HEXACO Integration Detailed Metrics (RQ4)

Table 14 reports the Precision, Recall, F1, Accuracy, and confusion-matrix entries underlying the F1 values shown in main-text Table 6. Hyperparameters $(K, r)$ for WG+Text, $(\alpha, K, r)$ for

Table 13: Full RQ3 results with confusion matrices on AEGIS2 test split (1,964 samples; 394 unsafe / 1,570 safe). All models at default settings; HEXACO at $\tau$=1.0 (pre-optimization, for reference).

| Model | Prec. | Recall | F1 | Acc. | TP | FP | FN | TN |
|---|---|---|---|---|---|---|---|---|
| OpenAI Moderation | 0.859 | 0.510 | 0.640 | 0.885 | 201 | 33 | 193 | 1537 |
| LlamaGuard | 0.855 | 0.525 | 0.651 | 0.887 | 207 | 35 | 187 | 1535 |
| WildGuard | 0.828 | 0.782 | 0.804 | 0.924 | 308 | 64 | 86 | 1506 |
| HEXACO+GPT | 0.204 | 0.990 | 0.339 | 0.220 | 390 | 1527 | 4 | 43 |
| LLM-Judge | 0.220 | 0.977 | 0.360 | 0.302 | 385 | 1362 | 9 | 208 |
| HEXACO Detector ($\tau$=1.0) | 0.578 | 0.178 | 0.272 | 0.809 | 70 | 51 | 324 | 1519 |

WG+HDist, and $(K, r_{\text{cat}}, \beta)$ for WG+HGate are selected on a held-out AEGIS2 train validation split (Section 4.1; Limitations §6.2.2) and applied unchanged to AEGIS2 test, WildGuardMix, and XSTest. WG+HGate uses $\mathcal{C}_{\text{hi}}$ derived from the same validation split (Table 16). WG+HOracle requires ground-truth category labels in the test-derived high-efficiency set $\mathcal{C}_{\text{hi}}^{\text{test}}$ (Table 15); the WildGuardMix and XSTest taxonomies do not overlap with $\mathcal{C}_{\text{hi}}^{\text{test}}$, so WG+HOracle reduces to WildGuard on those datasets (denoted "≡WG"). On AEGIS, WG+Text and WG+HDist trade $-0.021$ and $-0.032$ Precision for $+0.101$ and $+0.109$ Recall to gain F1 by $+0.040$ and $+0.037$; WG+HGate's strict val-tuned $r_{\text{cat}} = 0.9$ causes the override to fire rarely and the method effectively replicates WildGuard's confusion matrix on all three datasets. WG+HOracle uses the test-derived (leakage) category set and serves as an upper bound only.

**Validation-Tuned Hyperparameters.** The validation-tuned hyperparameters used in main-text Table 6 are: WG+Text ($K = 30$, $r = 0.5$); WG+HDist ($\alpha = 0.2$, $K = 30$, $r = 0.45$); WG+HGate ($K = 30$, $r_{\text{cat}} = 0.9$, $\beta = 0.1$). WG+HOracle is not val-tuned: $\beta = 0.16$ is inherited from the RQ2 Decision Model threshold, and the high-efficiency category set $\mathcal{C}_{\text{hi}}^{\text{test}} =$ {Guns/IW + Criminal Planning, Suicide/Self-Harm, Controlled Substances, Profanity, Harassment} is identified via the test-split rescue-efficiency analysis (Table 15).

### E.4 Full Rescue Efficiency Table (RQ4)

Table 15 shows rescue efficiency for all AEGIS2 test categories with at least 8 unsafe samples (12 of 24 harm categories meet this threshold; categories with fewer unsafe samples are excluded for statistical stability, see §6.3.1). Five categories achieve efficiency $\geq 2.0$, meaning that each added FP recovers at least two missed unsafe samples: Guns/IW, Suicide/SH, Controlled Substances, Profanity, and Harassment. PII/Privacy and Violence–Criminal Planning have efficiency $< 1.0$.

### E.5 Train-Derived Rescue Efficiency (used by WG+HGate)

Table 15 (above) is computed on the AEGIS2 test split and therefore carries test-set leakage when its high-efficiency categories are used to construct WG+HOracle's $\mathcal{C}_{\text{hi}}^{\text{test}}$. To define WG+HGate's $\mathcal{C}_{\text{hi}}$ without leakage, we re-compute the same rescue efficiency on the AEGIS2 train validation split (valid-WG subset, $n = 2{,}161$, held-out from the oax LogReg fit; see Section 4.1), aggregating by primary AEGIS2 category via substring matching against the multi-label tags so that each sample contributes to every primary category present in its label. The $n = 2{,}161$ pool used here is a superset of the $n = 1{,}997$ validation subset used for ensemble hyperparameter tuning (Section 4.1): rescue-efficiency computation requires only (HEXACO factors, WildGuard prediction, label, category), while ensemble tuning additionally requires the text-embedding cache entry, which is missing for 164 train rows. Table 16 reports the result. Unlike the small-$n$ test analysis, no primary category exceeds efficiency 1.0 at train scale; the original $\geq 2.0$ threshold was a small-sample artifact. WG+HGate's $\mathcal{C}_{\text{hi}}$ is the top-5 by efficiency (with $n_{\text{unsafe}} \geq 8$): {Guns and Illegal Weapons, Sexual, Profanity, Controlled/Regulated Substances, Criminal Planning/Confessions}. Three of these five also appear in the test-derived set used by WG+HOracle (Guns/IW, Profanity, Controlled/Regulated); the train-derived set adds Sexual and Criminal Planning, while the test-derived set adds Suicide and Self Harm and Harassment (ranks 8 and 12 on train).

Table 14: Per-dataset Precision, Recall, F1, Accuracy, and confusion-matrix entries for the four WG–HEXACO integration strategies (RQ4). Hyperparameters for WG+Text, WG+HDist, and WG+HGate are tuned on the AEGIS2 train validation split (Section 4.1) and applied unchanged to all three test sets. WG+HGate uses $\mathcal{C}_{\text{hi}}$ re-derived from train-val rescue efficiency (Table 16); WG+HOracle uses the original test-derived $\mathcal{C}_{\text{hi}}^{\text{test}}$ (Table 15) with $\beta = 0.16$ and is not val-tuned. F1 values match main-text Table 6 exactly. "≡WG" indicates the override condition is never satisfied (no GT high-efficiency categories in the dataset taxonomy).

| Dataset | Method | Prec. | Recall | F1 | Acc. | TP | FP | FN | TN |
|---|---|---|---|---|---|---|---|---|---|
| AEGIS2 test | WildGuard | 0.828 | 0.782 | 0.804 | 0.924 | 308 | 64 | 86 | 1506 |
| | WG+Text | 0.807 | 0.883 | 0.844 | 0.934 | 348 | 83 | 46 | 1487 |
| | WG+HDist | 0.796 | 0.891 | 0.841 | 0.932 | 351 | 90 | 43 | 1480 |
| | WG+HGate | 0.827 | 0.787 | 0.806 | 0.924 | 310 | 65 | 84 | 1505 |
| | WG+HOracle† | 0.811 | 0.914 | 0.859 | 0.940 | 360 | 84 | 34 | 1486 |
| WGMix test | WildGuard | 0.313 | 0.644 | 0.422 | 0.709 | 183 | 401 | 101 | 1040 |
| | WG+Text | 0.318 | 0.757 | 0.448 | 0.692 | 215 | 462 | 69 | 979 |
| | WG+HDist | 0.314 | 0.761 | 0.445 | 0.688 | 216 | 471 | 68 | 970 |
| | WG+HGate | 0.309 | 0.648 | 0.419 | 0.704 | 184 | 411 | 100 | 1030 |
| | WG+HOracle† | ≡WG | | | | | | | |
| XSTest | WildGuard | 0.374 | 0.872 | 0.523 | 0.722 | 68 | 114 | 10 | 254 |
| | WG+Text | 0.343 | 0.923 | 0.500 | 0.677 | 72 | 138 | 6 | 230 |
| | WG+HDist | 0.329 | 0.923 | 0.485 | 0.657 | 72 | 147 | 6 | 221 |
| | WG+HGate | 0.365 | 0.885 | 0.517 | 0.711 | 69 | 120 | 9 | 248 |
| | WG+HOracle† | ≡WG | | | | | | | |

Table 15: HEXACO rescue efficiency on the AEGIS2 test split (WildGuard baseline), restricted to the 12 of 24 harm categories with ≥8 unsafe samples for statistical stability (see §6.3.1). "Rescued" = WildGuard FNs recovered by HEXACO; "+FP" = newly added false positives (category-scoped). ∞ efficiency reflects zero added FPs under small-$n$ category counts and should be interpreted with caution. Categories sorted by efficiency.

| Category | $n_{\text{unsafe}}$ | WG FN | Rescued | +FP | Efficiency |
|---|---|---|---|---|---|
| Guns/IW + Criminal Planning | 9 | 3 | 3 | 0 | ∞ |
| Suicide and Self Harm | 8 | 3 | 3 | 0 | ∞ |
| Controlled/Regulated | 17 | 8 | 8 | 2 | 4.00 |
| Profanity | 14 | 7 | 6 | 2 | 3.00 |
| Harassment | 17 | 4 | 4 | 2 | 2.00 |
| Criminal Planning | 62 | 9 | 9 | 5 | 1.80 |
| Criminal + Controlled | 14 | 3 | 3 | 2 | 1.50 |
| Violence | 9 | 4 | 4 | 3 | 1.33 |
| Hate/Identity Hate | 33 | 3 | 3 | 3 | 1.00 |
| Sexual | 21 | 3 | 3 | 3 | 1.00 |
| PII/Privacy | 16 | 3 | 3 | 4 | 0.75 |
| Violence + Criminal Planning | 10 | 1 | 1 | 2 | 0.50 |

Table 16: Train-derived HEXACO rescue efficiency (WildGuard baseline, AEGIS2 train val subset; valid-WG only, $n = 2{,}161$). Held-out oax LogReg predictions; aggregated by primary category via substring match. Categories sorted by efficiency. WG+HGate's $\mathcal{C}_{\text{hi}}$ uses the top-5 entries with $n_{\text{unsafe}} \geq 8$.

| Primary category | $n_{\text{unsafe}}$ | WG FN | Rescued | +FP | Efficiency |
|---|---|---|---|---|---|
| Guns and Illegal Weapons | 73 | 36 | 36 | 62 | 0.58 |
| Sexual | 100 | 57 | 55 | 108 | 0.51 |
| Profanity | 80 | 48 | 48 | 95 | 0.51 |
| Controlled/Regulated Substances | 66 | 38 | 38 | 79 | 0.48 |
| Criminal Planning/Confessions | 325 | 185 | 185 | 400 | 0.46 |
| Violence | 136 | 78 | 77 | 175 | 0.44 |
| Threat | 10 | 8 | 8 | 20 | 0.40 |
| Suicide and Self Harm | 16 | 7 | 7 | 18 | 0.39 |
| PII/Privacy | 80 | 41 | 41 | 106 | 0.39 |
| Hate/Identity Hate | 139 | 78 | 74 | 209 | 0.35 |
| Fraud/Deception | 18 | 13 | 13 | 37 | 0.35 |
| Harassment | 137 | 68 | 66 | 200 | 0.33 |
| Malware | 9 | 3 | 3 | 12 | 0.25 |
| Needs Caution | 53 | 28 | 27 | 124 | 0.22 |
| Political/Misinformation/Conspiracy | 14 | 9 | 9 | 42 | 0.21 |

### E.6 Detector Pairwise Agreement (RQ4)

Table 17 shows pairwise agreement statistics on the AEGIS2 test split. All six pairs have positive Phi (0.431–0.588), confirming that the four detectors agree more often than independence would predict. HEXACO and WildGuard share the highest TP Jaccard overlap (0.761) with Phi=+0.563; HEXACO achieves high recall (0.975) and covers most of WildGuard's positives, yet 41 HEXACO-unique true positives missed by all three conventional detectors and the rescue capability of Table 15 indicate a non-redundant complementary signal rather than full subsumption. Among the three WildGuard pairings, Phi is highest with OpenAI (+0.588), followed by LlamaGuard (+0.574) and HEXACO (+0.563), reflecting WildGuard's strong agreement with the other content-based detectors; the HEXACO pairings with LlamaGuard and OpenAI sit lower (+0.440, +0.431), consistent with HEXACO's distinct error profile (much higher recall, much higher false-positive count) relative to the conventional content-based detectors.

Table 17: Pairwise detector agreement on AEGIS2 test split. Phi: Matthews correlation coefficient (MCC) over all 1,964 samples (binary predictions). Jaccard: TP set overlap (unsafe samples only). Predictions: `flag_output` for OpenAI/LlamaGuard; text-parsed "Safety rating: yes" for WildGuard; oax LogReg ($C = 0.01$, $\tau = 0.16$) for HEXACO.

| Pair | Phi | Jaccard |
|---|---|---|
| HEXACO × WildGuard | $+0.563$ | 0.761 |
| LlamaGuard × WildGuard | $+0.574$ | 0.528 |
| OpenAI × WildGuard | $+0.588$ | 0.524 |
| HEXACO × LlamaGuard | $+0.440$ | 0.523 |
| HEXACO × OpenAI | $+0.431$ | 0.504 |
| LlamaGuard × OpenAI | $+0.460$ | 0.427 |

**HEXACO-Unique True Positives.** HEXACO detects 41 cases (10.4% of 394 unsafe test samples) missed by all three conventional detectors (OpenAI, LlamaGuard, WildGuard). These are concentrated in Controlled Substances (5 cases), Profanity (4), and multiple single-category cases, reflecting HEXACO's complementary strength in categories where the conventional supervised classifiers fall short.

## F Disclosure Statements

**Use of AI Assistants.** This manuscript was originally drafted in Japanese; AI language assistants were used for English translation and limited prose polishing. All technical content, analyses, and conclusions are the authors' own.

**License and Artifact Use.** All datasets (AEGIS2, WildGuardMix, XSTest, HarmBench, and RealToxicityPrompts) and models (LlamaGuard, WildGuard, Gemma-3-270M-IT, `multilingual-e5-large`, OpenAI Moderation API, and `gpt-5.4-nano`) are used in compliance with their respective licenses and intended research uses.