*Life of Research · Published 26 November 2026 · DOI 10.5281/zenodo.24545381*

# A Likelihood Ratio Is Not a Whole Reading: What Class-Conditional Guard Models Change About Shortcuts, Calibration and Position, Why the Generative-Classifier Evidence Does Not Transfer to a Discriminatively Trained Ratio, and the Matched Comparison That Would Assign the Gains

Pranay Mahendrakar
SONYTECH · ORCID 0009-0003-7224-029X
pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

**Abstract.** Guard models screen the prompts sent to a language model and the responses it returns, and almost all of them are trained to emit one verdict token. A 2026 proposal argues that this objective is the root of three familiar failures - reliance on shortcut cues, overconfidence, and sensitivity to where the evidence sits in the input - and replaces it with a class-conditional score: a masked diffusion model reconstructs the moderated text under a "safe" and an "unsafe" hypothesis, and the difference of the two log-likelihoods is the verdict. Among thirteen guards on the authors' held-out benchmarks, the resulting guard posts the best average F1 and the lowest calibration error. In this analysis the author reviews and synthesises published results to show that the measured gains are real but that the attribution to the objective is not yet earned. Four things change at once in that comparison: the training loss, the backbone and its attention mask, the training data and fine-tuning regime, and the inference compute. The loss that is actually optimised is a binary cross-entropy on the likelihood difference, which is a discriminative loss over a generative parameterisation; the argument that generative classifiers avoid shortcuts because they must model every feature was established for pure class-conditional training, and in the strongest text evidence for it adding an unconditional generative term to a discriminative loss did not help. The one same-backbone ablation reads a verdict from weights never trained to produce one, so it isolates a readout, not an objective. The generative-classifier record outside guards is two-sided, and the side favouring generative models is weakest for pretrained models fine-tuned on ample data, which is the guard setting. A documented refusal-cue shortcut sits in one of the two training sets, and nothing in a class-conditional model stops it from being learned. The paper consolidates the measured values in two tables and a figure, separates what each published number can and cannot support, and specifies a factorial audit - objective, backbone, data and readout varied one at a time, with threshold sweeps, post-hoc calibration and cue-insertion tests - stating which outcome would support which reading.

**Keywords:** guard models; content moderation; generative classifiers; masked diffusion language models; shortcut learning; calibration; LLM safety; evaluation methodology

SONYTECH ORCID: 0009-0003-7224-029X

## 1. Introduction

A guard model is a classifier placed in front of or behind a language model. It reads a user prompt, or a prompt together with a response, and returns a decision about whether the content violates a safety policy. The current generation is built by fine-tuning a language model to emit a verdict: Llama Guard framed moderation as instruction following in which the model writes "safe" or "unsafe" after reading the conversation (Inan et al., 2023), and WildGuard, ShieldGemma, Granite Guardian and Qwen3Guard follow the same pattern with different data, taxonomies and sizes (Han et al., 2024; Zeng et al., 2024; Padhi et al., 2024; Zhao et al., 2025), as do guards trained on the Aegis 2.0 data (Ghosh et al., 2025). The probability the model assigns to the verdict token becomes the guard's score.

These guards have well-documented weaknesses. They are miscalibrated, and their confidence degrades further under jailbreaks: across nine guards and twelve benchmarks Liu et al. (2024) found systematic overconfidence, and Hong et al. (2026) report that guards which are close to calibrated on clean prompts lose calibration by an order of magnitude under attack while their base language models remain uncertain on the same inputs. They react to surface cues: benign prompts containing words common in injection attacks drive prompt-guard accuracy toward chance (Li & Liu, 2024), meaning-preserving paraphrases move safety scores substantially (Pinneri & Louizos, 2025), and inserting a refusal phrase into a harmful response can flip a guard's verdict to safe (Feng et al., 2026). They can be evaded by character-level and adversarial perturbations (Hackett et al., 2025; Mangaokar et al., 2024; Wei et al., 2024).

Lek et al. (2026) offer a single diagnosis for much of this. Their claim is that the verdict-token objective is too narrow: supervision lands on one output token, so the model is free to latch onto whichever fragment of the input best separates the training labels, and nothing in the loss asks it to account for the rest. Their remedy is to turn the classifier around. Instead of predicting the label from the text, the model asks which label better explains the text. A masked diffusion language model, LLaDA (Nie et al., 2025), is fine-tuned to reconstruct the moderated

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

region - the prompt for prompt moderation, the response for response moderation - conditioned on a visible label, and the guard compares the reconstruction likelihood under "unsafe" with that under "safe". The idea has a long pedigree. It is the generative classifier of classical statistics, revived for text by noisy-channel prompting (Min et al., 2021) and for images by diffusion classifiers (Li et al., 2023; Clark & Jaini, 2023), and recently argued to avoid shortcut solutions because the model must represent all features of the input rather than only those that separate the classes (Li et al., 2025a).

The resulting guard, LLaDA-Guard, is competitive on discrimination and best on all three calibration measures, by a modest ECE margin over Llama Guard 3 and a large one over the strongest ranking models, in the authors' evaluation: the best average F1 at the default threshold among thirteen guards, an average area under the precision-recall curve within 1.4 points of the best, and the lowest average expected calibration error (Lek et al., 2026). It also shows less over-blocking of benign prompts with alarming words, less dependence of response verdicts on the prompt, flatter behaviour when the decisive cue moves within a long input, and a token-level localiser that supports rewriting unsafe prompts.

This paper asks what those results establish about the objective. In this analysis the author reviews and synthesises published results to show that LLaDA-Guard's measured advantages are real on the reported benchmarks, and that the inference from those advantages to "the class-conditional objective removes the incentive for shortcut, overconfidence and positional failures" is not yet supported. The argument has five parts.

First, four things change together when LLaDA-Guard is compared with existing guards: the loss, the backbone (a bidirectional masked diffusion model in place of causal autoregressive ones), the training data and fine-tuning regime (40,000 examples, LoRA, a specific mixture of two public datasets), and the inference compute (six reconstruction passes instead of one). Each of the claimed benefits has a published mechanism in at least one of the other three.

Second, the loss actually optimised is a binary cross-entropy applied to the likelihood difference, plus a reconstruction term whose weight the paper does not report. A cross-entropy on a likelihood ratio is a discriminative loss: it is minimised by any assignment of label-dependence to tokens that separates the training classes, including assignments that place all of it on one trigger phrase. The strongest text evidence that generative classifiers avoid shortcuts comes from pure class-conditional training, and in that same evidence a hybrid that added a generative term to a discriminative loss showed no gain (Li et al., 2025a).

Third, the generative-classifier record is two-sided. Generative classifiers learn faster from few examples and often generalise better under shift (Yogatama et al., 2017; Min et al., 2021; Li et al., 2025a; Jaini et al., 2023), but they can carry shortcuts just as discriminative models do (Yang et al., 2022), their likelihoods carry no guarantee against adversarial inputs (Fetaya et al., 2019), and with pretrained weights the discriminative encoder outperformed the generative autoregressive model across data regimes in the most comprehensive transformer-era comparison (Kasa et al., 2025).

Fourth, the same-backbone ablation Lek et al. report does not isolate the objective. It reads the verdict token from the same LoRA weights, which were trained only for reconstruction. A guard trained for verdict prediction on the same backbone and data is the missing arm.

Fifth, a specific shortcut is already documented in one of LLaDA-Guard's two training sets - refusal phrases occur only with "unharmful" response labels in WildGuardMix (Feng et al., 2026) - and a class-conditional model of the response would represent that association as readily as a verdict-token model. Whether LLaDA-Guard carries it has not been tested.

The paper's contributions are:

1. A separation of the four factors that change between LLaDA-Guard and its baselines, with the published evidence that ties each of the claimed benefits to each factor (Sections 2, 7-10).
2. A short derivation showing that the moderation loss is minimised by solutions that concentrate label-dependence on discriminative fragments, so that "supervision on every token" describes where gradients flow, not what the loss requires (Section 4).
3. A consolidation of the generative-versus-discriminative evidence, organised by the regime it was gathered in, showing that the regime closest to guard training is the one in which the generative advantage is least established (Section 5, Table 2).
4. A reading of the reported ablation, including an internal inconsistency in its F1 table, and of what the over-blocking, leakage, calibration and localisation comparisons support once operating point, template and attribution method are held in view (Sections 7-11, Table 1, Figure 1).
5. An attribution audit, stated as Algorithm 1, with a table of the outcomes each hypothesis predicts (Section 13, Table 3).

The paper does not argue that LLaDA-Guard is a poor guard or that generative scoring is a dead end. It argues that the reported comparison cannot yet say why the guard behaves as it does, and that the reason matters: if the gains come from the data mixture, the fine-tuning regime or Monte Carlo averaging, they are available to existing verdict-token guards at a sixth of the inference cost.

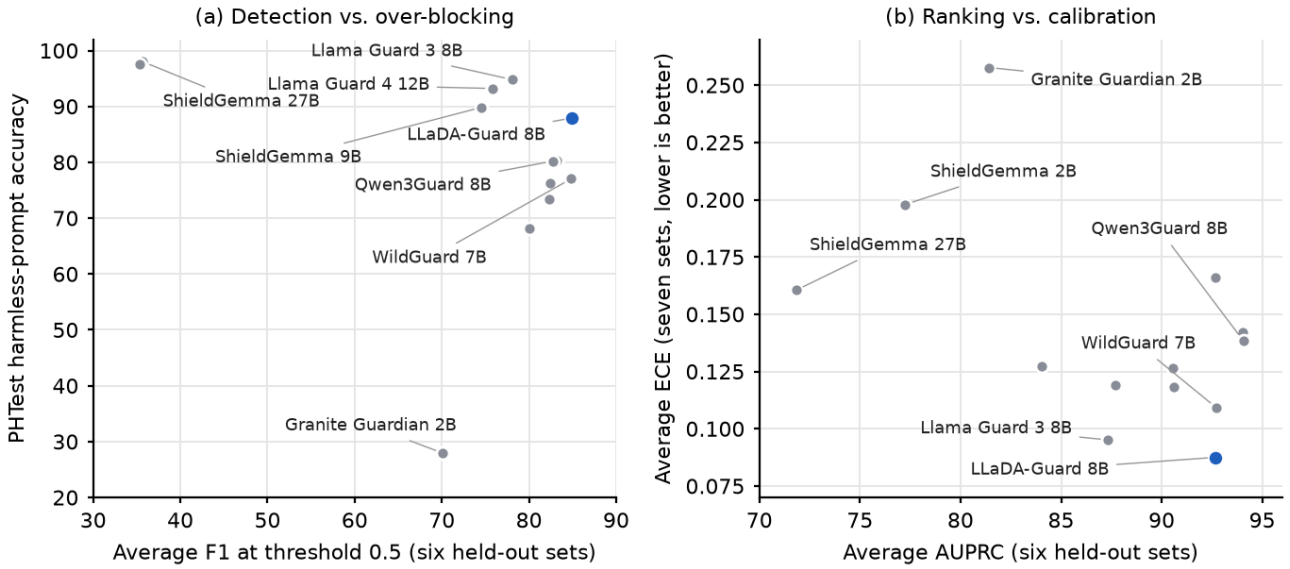



**Figure 1.** Thirteen guard models as reported by Lek et al. (2026), Tables 1 and 5. (a) Average F1 at threshold 0.5 over six held-out sets against accuracy on PHTest harmless prompts: no guard leads on both, and LLaDA-Guard (blue) sits between the high-F1 guards that over-block and the Llama Guard and ShieldGemma models that block less but detect less. (b) Average AUPRC against average ECE: LLaDA-Guard has the lowest ECE and the third-highest AUPRC.

## 2. Terms and the Four Claims

**Verdict-token guard.** A model trained with cross-entropy to emit a label token after reading the conversation. Its score is the probability of the "unsafe" token, p(y = unsafe | x). This is the design of Llama Guard and its successors (Inan et al., 2023; Han et al., 2024; Zhao et al., 2025). Some verdict-token guards use one template for prompts and another for responses - Lek et al. (2026) note this for ShieldGemma and NemoGuard - while others use a shared conversation template.

**Class-conditional (generative) classifier.** A model of p(x | y) for each class y, classifying by Bayes' rule. Trained by maximum likelihood on (x, y) pairs, it must assign probability to every token of x under its true label. This is the object of the classical comparisons (Yogatama et al., 2017; Zheng et al., 2023) and of the recent shortcut argument (Li et al., 2025a).

**Discriminatively tuned generative classifier.** A model parameterised as p(x | y) but trained, wholly or partly, with a loss on the posterior it implies - log loss, hinge or margin losses over the class-conditional likelihoods. Ding et al. (2020) studied this family for natural language inference, following an earlier log-loss fine-tuning step for generative question answering, and reported that it retains some advantages of the generative model while improving on log loss with other margins. Energy-based reinterpretations of discriminative classifiers occupy a neighbouring position from the other side (Grathwohl et al., 2019).

**LLaDA-Guard.** Lek et al. (2026) define a score s(M) as the log-likelihood of the moderated region M under "unsafe" minus that under "safe", estimate each term by averaging masked-reconstruction log-probabilities over K random masks, and train with two terms: a moderation loss, the binary cross-entropy between sigmoid(s) and the true label, and a generative loss, the negative reconstruction log-likelihood of the region under the true label, weighted by lambda. By the definitions above, LLaDA-Guard is a discriminatively tuned generative classifier whose generative component has an unreported weight.

The paper's claims for the class-conditional objective are four, and they are worth stating separately because the evidence for each comes from different experiments.

- **C1, shortcuts.** The objective reduces reliance on discriminative fragments: less over-blocking of benign prompts that contain alarming words, smoother score trajectories as a prompt is revealed, and response verdicts that depend less on the prompt.
- **C2, calibration.** The objective yields better-calibrated probabilities: lower expected calibration error, Brier score and negative log-likelihood, and fewer errors made with more than 90 percent confidence.
- **C3, position.** The objective, together with bidirectional scoring, removes positional sensitivity inherited from causal attention.
- **C4, localisation.** Because the score is built from per-token likelihoods, token-level evidence comes for free and supports rewriting.

Four factors change together between LLaDA-Guard and its baselines, and each is a candidate cause:

- **F1, objective**: class-conditional reconstruction scored by a likelihood ratio, against verdict-token cross-entropy.

- **F2, backbone**: LLaDA-8B-Instruct, a masked diffusion model with bidirectional attention, against the causal autoregressive backbones of all twelve baselines.
- **F3, data and regime**: a 40,000-example stratified subset of WildGuardMix (29.3 percent) and Aegis 2.0 (70.7 percent), fine-tuned with LoRA, against each baseline's own and often undisclosed training mixture and procedure.
- **F4, inference compute and readout**: two labels times K = 3 masks, averaged, at masking rate 0.15, against one forward pass.

The rest of the paper takes each claim in turn and asks which factor the published evidence implicates.

## 3. What the LLaDA-Guard Study Reports

Lek et al. (2026) evaluate on six mixed-label held-out sets - ToxicChat (Lin et al., 2023), the OpenAI moderation set (Markov et al., 2022) and XSTest (Rottger et al., 2023) for prompts; BeaverTails (Ji et al., 2023), PKU-SafeRLHF (Ji et al., 2024) and the XSTest response set released with WildGuard (Han et al., 2024) for responses - plus the harmless subset of PHTest (An et al., 2024) as an over-blocking test. Twelve released guards serve as baselines, run with their official templates; Qwen3Guard is run in its strict mode. Metrics are AUPRC, F1 at a probability threshold of 0.5, and three calibration measures. LLaDA-Guard is evaluated with three masks per label at masking rate 0.15 and classifies unsafe when the score is positive, which is the threshold 0.5 under equal priors.

Table 1 collects the values the argument below rests on. Every row is from Lek et al. (2026); the second column names the table or appendix.

Three features of these numbers matter for what follows. The F1 lead over WildGuard is 0.08 points, a difference no interval in the source is offered to resolve. On ranking, as opposed to thresholded decisions, LLaDA-Guard is third, behind two Qwen3Guard sizes. And its over-blocking accuracy, while well above the strongest discriminative classifiers, is below five of the twelve baselines.

The paper is candid about some limits. Its own limitations appendix restricts the claims to English binary moderation, excludes adaptive jailbreaks from its claims, and notes that the guard needs two reconstruction scores averaged over several masks where a verdict-token guard needs one pass (Lek et al., 2026). The analysis below concerns what the comparison can attribute, not whether the authors overstated their scope.

## 4. The Training Loss Is a Discriminative Loss on a Generative Parameterisation

Lek et al. motivate the class-conditional objective by contrast with verdict prediction: a verdict-token loss is satisfied by any separating function, while the class-conditional score is "built up from token-level reconstruction", so the model has to account for the whole moderated region (Lek et al., 2026). That contrast is accurate for a pure generative classifier. It is not accurate for the loss LLaDA-Guard is trained with, and the gap is the first reason the shortcut argument does not transfer automatically.

Write the score as a sum over tokens of the moderated region:

s = sum over i in M of [ log p(x_i | context, unsafe) - log p(x_i | context, safe) ] = sum over i of d_i.

The moderation loss depends on the parameters only through s. It is the logistic loss of s against the label. Its value is unchanged by any redistribution of the per-token differences d_i that preserves their sum, and it is minimised, on the training data, by any parameter setting that makes s large and positive on unsafe examples and large and negative on safe ones. Nothing in it rewards a non-zero d_i on a token whose distribution genuinely differs between classes, or penalises a large d_i on a token that merely co-occurs with one class. A model that conditions on the label only when it reconstructs one trigger phrase, and reconstructs every other token identically under both labels, satisfies the moderation loss exactly as well as one that spreads label-dependence over the semantically relevant content. In that solution d_i is zero for every token but the trigger, and the score is a one-feature classifier. This is the shortcut solution, expressed in the generative parameterisation. Gradients do reach every token of the region - that is the sense in which supervision moves "to every token" - but the optimum the loss defines does not require the model to use them.

What would require it is the generative term. Maximising log p(x_M | context, y*) under the true label forces the model to predict every token well. Even here, a caution applies. Most tokens of a harmful request or a compliant response are predictable from their context without the label: syntax, named entities, the continuation of a sentence already begun. The generative term is satisfied by modelling those tokens well under either label, and the classifier uses only the part of the reconstruction that depends on the label. Li et al. (2025a) make the positive version of this argument carefully: a generative classifier learns all features but weights them by how consistently they track the label, and in their Gaussian analysis whether that helps depends on the strength of the spurious correlation and the noise in the core features. Their argument concerns models trained by maximum likelihood of p(x | y). With a moderation loss in the mix, how much of that inductive bias survives depends on lambda, and the author could not find a value for lambda in the paper's method section, implementation details or appendices.

The closest published test of exactly this question points the wrong way for a hybrid. On CivilComments, a toxicity benchmark, Li et al. (2025a) trained the same autoregressive architecture three ways. With a discriminative objective it reached 91.4 in-distribution and 35.7 worst-group accuracy. Adding an unconditional

**Table 1.** Values reported by Lek et al. (2026) for LLaDA-Guard and comparison guards. Scores are x100 unless stated. "Avg" is the source's mean over its six mixed-label sets for AUPRC and F1 and over seven sets (adding PHTest) for ECE. PHTest is accuracy on harmless prompts only.

| Quantity | Source location | LLaDA-Guard | Comparison values |
|---|---|---|---|
| Avg F1 at 0.5 | Table 1 | 84.91 | WildGuard 84.83; NemoGuard 83.23; Qwen3Guard-8B 82.75 |
| Avg AUPRC | Table 1 | 92.66 | Qwen3Guard-8B 94.06; Qwen3Guard-4B 94.04; WildGuard 92.73 |
| PHTest accuracy | Table 1 | 87.92 | ShieldGemma-27B 98.03; Llama Guard 3 94.80; WildGuard 77.03 |
| Avg ECE (fraction) | Table 5 | 0.0875 | Llama Guard 3 0.0951; WildGuard 0.1091; Qwen3Guard-8B 0.1384 |
| Avg Brier / NLL | Table 4 | 8.45 / 32.56 | Llama Guard 3 9.84 / 47.37; WildGuard 10.37 / 47.86 |
| Prompt leakage, increase in unsafe-prediction rate ($\Delta\hat{y}$) | Table 6 ($n = 250$) | 0.060 | Qwen3Guard 0.424; WildGuard 0.308; ShieldGemma 0.000; NemoGuard 0.004 |
| NotInject false positives, 3 triggers | Table 7 | 0.00% | Granite 10.62%; Qwen3Guard 3.54%; ShieldGemma 0.00% |
| Same-weights verdict readout, avg AUPRC | App. B.1, Table 2 | 91.62 | Direct readout 88.87 (five sets) |
| Same-weights verdict readout, F1 gain | App. B.1, Table 3 | +4.06 avg | per set: +1.89, +1.74, +0.22, +8.77, +7.67 |
| Rewrite conversion at budget 5 | App. F, Table 8 | 60.7 (masking localiser) | LLaDA-Guard attention 53.0; WildGuard attention 52.1 |

generative objective alongside the discriminative one gave 91.7 and 35.4. Training it as a pure class-conditional model gave 79.8 and 61.4. The worst-group gain belonged to the pure generative objective, came with an 11.6-point in-distribution cost, and did not appear when a generative term was merely added to a discriminative loss. LLaDA-Guard's objective is neither of Li et al.'s arms: its generative term is class-conditional rather than unconditional, and its discriminative term acts on the class-conditional likelihoods rather than on a separate head. The CivilComments result therefore does not predict LLaDA-Guard's behaviour. It shows only that "has a generative component" is not sufficient, which is the inference the class-conditional argument needs to rule out.

The other side has evidence too. Ding et al. (2020) found that discriminatively fine-tuned generative classifiers for natural language inference kept advantages over discriminative baselines under small training sets, label imbalance and label noise, with several margin losses outperforming the log loss used in earlier work. That is evidence that the hybrid can retain generative benefits. It was gathered on models trained from scratch, in data-poor and noisy-label regimes, and on robustness to sample size and label corruption rather than to spurious cues. The question for LLaDA-Guard - whether the hybrid, fine-tuned from a pretrained instruct model on 40,000 examples, avoids the specific cues that verdict-token guards fall for - is not one either line of evidence answers.

Two smaller points follow from the same algebra. First, because the score is a likelihood ratio, factors that affect both hypotheses equally cancel. This is a strength the paper could have claimed: raw likelihoods of generative models are dominated by background statistics such as length and token frequency (Nalisnick et al., 2018), and ratios are the standard correction (Ren et al., 2019; Gangal et al., 2019). The same cancellation is why the ratio is free to ignore most of the text. Second, the Neyman-Pearson reading Lek et al. give the score - thresholding a true likelihood ratio is the most powerful test at a fixed false-positive rate - applies to true class-conditional likelihoods. Here each likelihood is a three-sample Monte Carlo estimate of a variational bound, trained so that its difference minimises a logistic loss. Optimality transfers to the trained score only to the extent that the estimates are accurate likelihoods, which is the thing the discriminative term does not ask of them.

None of this shows that LLaDA-Guard learned shortcuts. It shows that the objective does not exclude them, so whether it avoided them is an empirical question. Sections 6 and 7 ask what the evidence on that question is.

## 5. The Generative-Classifier Record Is Two-Sided

The case for generative classification rests on an old trade-off and some newer robustness results. The trade-off, proved for linear models and confirmed for recurrent ones, is that generative classifiers approach their asymptotic error with fewer examples but have a higher asymptotic error than discriminative ones (Yogatama et al., 2017); Zheng et al. (2023) extend the sample-efficiency side to multiclass and non-linear settings and observe the "two regimes" pattern in pretrained vision features. Robustness results sit on top. Yogatama et al. (2017) found their generative recurrent classifiers more robust in zero-shot and continual-learning settings. Noisy-channel prompting, which scores the input given each label, beat direct prompting in few-shot text classification with lower variance and better worst-case accuracy, especially under label imbalance and unseen labels (Min et al., 2021). Diffusion classifiers show strong shape bias, human-like error patterns and out-of-distribution accuracy (Jaini et al., 2023; Clark & Jaini,

2023), and a diffusion-based generative classifier has been made adversarially robust in images (Chen et al., 2023). Earlier, an analysis-by-synthesis model built on learned class-conditional distributions reported state-of-the-art robustness on MNIST under several perturbation norms (Schott et al., 2018), and deep Bayes classifiers resisted a range of attacks (Li et al., 2018). Li et al. (2025a) report the best worst-group or out-of-distribution accuracy on five distribution-shift benchmarks with generative classifiers trained from scratch.

The record against is as specific. Yang et al. (2022) set out to use the comprehensive representation a generative model should learn and found it did not prevent shortcuts: on ColoredMNIST, a VAE classifier reached 13.2 percent on the shifted test set against 17.8 percent for a plain discriminative classifier, and a class-conditional generative classifier fared just as poorly in their appendix; their fix required explicitly isolating the shortcut in a small latent subspace. Fetaya et al. (2019) prove that even a near-optimal generative classifier cannot guarantee that adversarially perturbed inputs will be detectable, and found that robustness obtained on MNIST "completely breaks down on CIFAR10". Likelihoods from deep generative models can rank out-of-distribution data as more probable than in-distribution data (Nalisnick et al., 2018). Even the robustness side carries a cost when read closely: Li et al.'s CivilComments generative classifier was 10.8 points below standard ERM in distribution (79.8 against 90.6), and their own classification of results places CivilComments among the benchmarks where generative classifiers are "the same or worse in-distribution" (Li et al., 2025a).

The decisive variable for guards is the regime. Most of the trained-classifier robustness results above were obtained with models trained from scratch, which Li et al. (2025a) and Kasa et al. (2025) chose deliberately so that pretraining would not confound the comparison. The pretrained results (Min et al., 2021; Clark & Jaini, 2023; Jaini et al., 2023) are zero-shot or few-shot uses of large generative models, not fine-tuning on ample labelled data. Guards are not trained from scratch. They are fine-tuned from instruction-tuned language models with tens of thousands to hundreds of thousands of labelled examples. Kasa et al. (2025), the most comprehensive transformer-era comparison - autoregressive, masked-language-model, discrete-diffusion and encoder classifiers on nine benchmarks - report that with pretrained weights the two-regimes pattern disappeared and the discriminative encoder outperformed the generative autoregressive model "across all data regimes in most datasets". In the from-scratch setting they found generative autoregressive classifiers highly sensitive to input noise, discrete-diffusion classifiers the most robust to it, and a pseudo-generative masked-language-model classifier best at high data. Min et al. (2021) found the direct models overtaking channel prompt tuning when the full training set was available. Clavie et al. (2025) show that a masked language model's own prediction head, used generatively, matches or beats a classification head after fine-tuning, which is evidence that generative readouts are competitive in the pretrained regime but not that they are more robust there.

Table 2 places these results by regime.

Read by regime, the record says something narrower than either side's headline. Generative classifiers have a well-supported advantage when data are scarce, labels are imbalanced or noisy, and the test distribution shifts in ways a discriminative learner can exploit during training. Their advantage is not established for pretrained models fine-tuned on ample labelled data, and there is a documented reversal in that regime. Guard training is closer to the second setting than to the first, except in one respect: harmful categories are imbalanced and long-tailed, and some benchmarks are far from the training distribution. That exception is where the generative argument has its best purchase, and it is a reason to test for it specifically rather than to assume it.

## 6. Shortcuts Live in the Data: The Refusal-Cue Test Nobody Has Run

The shortcut literature locates the cause of most shortcuts in the data rather than the loss: a feature that co-occurs with the label in the training set will be used by a learner that can see it (Geirhos et al., 2020; Gururangan et al., 2018; McCoy et al., 2019). The remedies that work best change the data or the readout - counterfactual augmentation (Kaushik et al., 2019), retraining the last layer on balanced data (Kirichenko et al., 2022) - not the family of model.

For guards, one such feature has been measured. Feng et al. (2026) audit WildGuardMix and GR-Train and find that among responses to harmful prompts, refusal expressions appear almost exclusively under the "unharmful" label: responses labelled both refusing and harmful are absent from WildGuardMix and make up 0.26 percent of harmful-prompt examples in GR-Train. Inserting a refusal cue into a harmful response flips verdicts to unharmful in guards trained on these sets and in Llama Guard 3 and Qwen3Guard; across the seven guards in their mitigation study, the mean rate of detection failure with a response-initial cue on WildGuardTest was 16.39 percent before their intervention. The Aegis 2.0 data follow a different construction, which Feng et al. describe as supplementing under-represented refusal and redirection patterns with synthetic safe responses, and the Aegis guard showed limited baseline vulnerability.

Now consider what a class-conditional model of responses learns from WildGuardMix. It estimates p(response | prompt, y). If refusal phrases occur only in responses labelled unharmful, the model learns that "I cannot help with that" is far more probable under y = safe than under y = unsafe, given a harmful prompt. Inserted at the start of a harmful response, those tokens

**Table 2.** Published comparisons of generative and discriminative classifiers, grouped by training regime. Each row is one cited result; values are the source's own.

| Source | Regime | Setting | Reported result |
|---|---|---|---|
| Li et al. (2025a) | From scratch | CivilComments, same AR architecture | p(y given x) 91.4 ID / 35.7 WG; p(x) + p(y given x) 91.7 / 35.4; p(x given y) 79.8 / 61.4 |
| Li et al. (2025a) | From scratch | CivilComments vs ERM encoder | ERM 90.6 / 53.3; generative 79.8 / 61.4 |
| Yang et al. (2022) | From scratch | ColoredMNIST, shifted test accuracy | discriminative 17.8; VAE classifier 13.2 |
| Fetaya et al. (2019) | From scratch | Conditional generative classifiers | robustness on MNIST, "completely breaks down on CIFAR10" |
| Kasa et al. (2025) | From scratch | Nine text benchmarks | generative AR highly noise-sensitive; diffusion most noise-robust |
| Kasa et al. (2025) | Pretrained | BERT-base vs GPT-2 base | encoder ahead of generative AR across data regimes, most datasets |
| Min et al. (2021) | Pretrained, few-shot | Channel vs direct prompting | channel ahead at small K and under imbalance; direct ahead at full data |
| Ding et al. (2020) | From scratch | NLI, discriminatively tuned generative | ahead of discriminative baselines under small data, imbalance, label noise |
| Lek et al. (2026) | Pretrained, LoRA | Guard benchmarks | best avg F1 and ECE; AUPRC third |

add a large negative contribution to the score. Whether that contribution outweighs the positive contributions of the harmful content that follows depends on how strongly each is represented - exactly the competition between spurious and core features that Li et al. (2025a) analyse, with an outcome that depends on the data. A generative model is not barred from the refusal shortcut. It is a model of the joint distribution, and the joint distribution contains it.

LLaDA-Guard's data mixture makes the prediction ambiguous in an informative way. Its training subset is 29.3 percent WildGuardMix and 70.7 percent Aegis 2.0 (Lek et al., 2026), and Aegis 2.0 is the dataset whose construction Feng et al. (2026) contrast with WildGuardMix on exactly this pattern. A LLaDA-Guard that resists refusal-cue insertion might do so because of its objective or because most of its data came from the resistant source. A verdict-token guard trained on the same mixture is the comparison that would tell. No refusal-cue result for LLaDA-Guard has been published.

The same logic applies to the over-blocking results the paper does report. On NotInject, whose 339 benign prompts are split by the number of injection trigger words (Li & Liu, 2024), LLaDA-Guard made no false positives at one, two or three triggers, and neither did ShieldGemma; Llama Guard 3 made one, at three triggers (Lek et al., 2026). At roughly 113 prompts per split, one false positive is 0.88 percent, and the comparison between the best five guards rests on a handful of prompts. More to the point, over-defence on exactly this dataset has been sharply reduced by a discriminative DeBERTa classifier trained with a strategy aimed at trigger-word bias, which its authors report performing comparably to GPT-4o across benign, malicious and over-defence inputs (Li & Liu, 2024). Trigger-word over-blocking is a property that data curation is known to move. That LLaDA-Guard shows little of it does not, on its own, show that the objective is what removed it.

## 7. What the Same-Backbone Ablation Isolates

Lek et al. are aware that LLaDA-Guard differs from its baselines in more than its objective, and they report an ablation meant to separate objective from backbone. "LLaDA-Guard Direct" uses the same weights as LLaDA-Guard but reads the guard's decision from the probabilities of the verdict token placed after the prompt and response - in the masked diffusion model, by unmasking the label position. The authors state that this "isolates the effect of the proposed class-conditional reconstruction objective from that of using LLaDA as the underlying model" (Lek et al., 2026). Across five held-out sets the class-conditional readout leads the direct readout by 2.75 points of AUPRC on average and by 4.06 points of F1, with the largest F1 gains on the two response sets.

The ablation does not isolate the objective. As the method is described, LLaDA-Guard's training loss never asks the model to predict the verdict token. In every training example the label is visible, as the condition under which the moderated region is reconstructed; it is never a reconstruction target. The direct readout therefore asks LoRA weights trained for one task to perform a different one, relying on whatever verdict-prediction ability survives from LLaDA-8B-Instruct plus the side effects of reconstruction training. The comparison the authors need is between two models trained on the same backbone and the same 40,000 examples: one with the class-conditional loss and one with verdict cross-entropy. What the ablation shows is narrower and still useful: given weights trained for class-conditional reconstruction, the readout that matches the training task beats the one that does not. That is expected under every hypothesis in Section 2 and does not discriminate between them.

Two further limits apply. The ablation reports AUPRC and F1 only. The calibration, position, leakage, over-blocking and localisation results - the ones the paper attributes to the objective - are not reported for the direct readout, so even the narrow comparison above

does not reach claims C1 to C4. And the ablation omits PKU-SafeRLHF, one of the six main-table sets.

The ablation's F1 table also contains an inconsistency worth recording precisely, because the reported gain is used in the text. Its BeaverTails cell for LLaDA-Guard reads 0.8329, while the main table gives LLaDA-Guard 82.19 F1 on BeaverTails; its average cell reads 0.8599. Taking the main-table BeaverTails value, the five per-set gains are 1.89, 1.74, 0.22, 8.77 and 7.67 points, their mean is 4.06, and LLaDA-Guard's five-set mean F1 is 83.29 - which is the number printed in the BeaverTails cell. The reported average gain of 4.06 points and the reported per-set deltas are consistent with the main table; the BeaverTails cell holds the five-set mean, and the printed average (0.8599) matches neither the consistent mean (0.8329) nor the mean of the cells as printed (0.8351). Table 1 above uses the consistent values. The appendix text describes the 4.06-point figure as the improvement on response datasets; in the table it is the mean over all five sets, and the response-set gains are 8.77 and 7.67.

## 8. Calibration: Objective, Regime, or Averaging?

LLaDA-Guard's calibration result is its most striking: average ECE 0.0875, Brier 8.45 and NLL 32.56, best on all three among thirteen guards (Lek et al., 2026). Figure 1(b) shows it in context. The margin over the best-calibrated baseline is modest - Llama Guard 3's average ECE is 0.0951 - and large over the strongest ranking models, Qwen3Guard at 0.1384 for the 8B model.

Three alternative causes for the calibration gain have independent support, and the comparison as run does not exclude any of them.

**Post-hoc calibration was applied to no baseline.** All twelve baselines were run with their released probabilities. Temperature scaling, a single parameter fitted on held-out data, is the standard remedy for neural-network overconfidence and is often enough on its own (Guo et al., 2017). For guards specifically, Liu et al. (2024) found temperature scaling the more effective fix for response classification and contextual calibration (Zhao et al., 2021), which needs no validation set, the more effective fix for prompt classification. Pinneri and Louizos (2025) report calibration improvements of up to 40 percent from consistency training of existing discriminative guards. A calibration comparison in which no model is recalibrated measures, in part, which models happen to ship with well-scaled probabilities. The paper also reports a sweep of LLaDA-Guard's inference settings - mask count and masking rate - on Aegis 2.0 validation data (its Appendix E); the baselines received no comparable tuning.

**Monte Carlo averaging is a calibration intervention.** LLaDA-Guard's score averages reconstruction log-likelihoods over K masks per label. Lek et al. report that more mask repetitions help "most for calibration", with ECE improving consistently from one pass to several (Lek et al., 2026). A noisy estimate of a logit, passed through a sigmoid, produces more extreme probabilities than the underlying logit warrants, and averaging reduces that noise. The author's reading is that some of LLaDA-Guard's calibration advantage is a test-time variance-reduction effect. Masked diffusion likelihood estimates are known to have high variance - LLaDA 1.5 was built around reducing exactly this variance for preference optimisation (Zhu et al., 2025) - and averaging a verdict-token guard over input perturbations or dropout samples would test whether the effect is specific to the objective.

**Light fine-tuning preserves base-model calibration.** LLaDA-Guard is a LoRA adaptation trained on 40,000 examples for under seven hours on one GPU (Lek et al., 2026). LoRA adapts less and forgets less of the base model than full fine-tuning (Biderman et al., 2024), and the Aegis 2.0 authors report that lightweight models trained parameter-efficiently on their data were competitive with leading guards fully fine-tuned on much larger datasets (Ghosh et al., 2025). Hong et al. (2026) find that where guards are confidently wrong, their base language models are often uncertain on the same inputs, and that the divergence arises in later layers where guards sharpen the safe-unsafe separation. A guard that departs less from its base model would, on this account, inherit more of the base model's uncertainty, whatever its objective.

There is also a measurement point. Kasa et al. (2025) could not compute ECE for their discrete-diffusion classifiers at all, because with an absorbing noise schedule the classifier produced only hard outputs; in their summary, generative autoregressive classifiers rate low on calibration and encoders high. Calibration is not a generic property of the generative route. LLaDA-Guard's probabilities come from passing a logit through a sigmoid under an assumed equal prior, and they are calibrated on these benchmarks because the logit is trained with a logistic loss and averaged over masks. Those are properties a verdict-token guard can share.

Finally, calibration under attack is the property guard deployment most needs and the one Hong et al. (2026) find collapses. LLaDA-Guard's calibration was measured on clean benchmarks only, and its authors exclude adaptive jailbreaks from their claims (Lek et al., 2026).

## 9. Position: Objective or Attention Mask?

Lek et al. show that when the decisive cue is moved within an input padded with benign filler, several autoregressive guards lose F1 when the cue appears early in a long context, while LLaDA-Guard stays near its peak (Lek et al., 2026). They attribute this partly to the autoregressive factorisation and causal attention, citing the lost-in-the-middle findings (Liu et al., 2023) and recency effects in long contexts (Peysakhovich & Lerer, 2023), and partly to the objective.

The two factors change together, and the attention mask has its own theory. Wu et al. (2025) show that causal masking biases multi-layer attention toward earlier

positions while relative position encodings introduce distance decay, and that their interaction shapes which positions a deep model can use. A bidirectional model is not subject to the causal component. A discriminative bidirectional classifier - an encoder such as the DeBERTa model behind InjecGuard (Li & Liu, 2024) - would separate the attention mask from the objective, and so would the direct readout of a verdict-trained LLaDA. Neither appears in the positional experiment.

Bidirectional masked diffusion models are not free of positional artefacts either. Kumar (2026), using LLaDA-8B and Dream-7B as training-free classifiers, found that when all candidate labels were placed in one prompt, the first answer slot was predicted positive for 99.4 percent of GoEmotions examples and all Reuters examples, and removed the artefact only by scoring each label in the same position. That is a different setup from LLaDA-Guard's, and it does not show that LLaDA-Guard is position-sensitive. It shows that the positional robustness reported for LLaDA-Guard is a property to be measured for the specific scoring template, not an architectural guarantee.

The prefix-reveal trajectories Lek et al. present for C1 bear on the same question. Autoregressive guards jump abruptly toward "unsafe" when a discriminative fragment appears; LLaDA-Guard's score rises more smoothly (Lek et al., 2026). A smoother trajectory is consistent with less reliance on single fragments, and it is also what averaging over random masks of the revealed prefix would produce. Without a trajectory from a mask-averaged verdict-token score, the shape of the curve does not separate the objective from the estimator.

## 10. Over-Blocking and Leakage Are Operating Points and Templates

**Over-blocking.** Lek et al. cite WildGuard's 77.03 percent accuracy on PHTest harmless prompts as evidence of over-reliance on shortcut features and LLaDA-Guard's 87.92 percent as evidence of the opposite (Lek et al., 2026). Figure 1(a) puts these in context. Llama Guard 3 and 4 and all three ShieldGemma models block less than LLaDA-Guard on PHTest; ShieldGemma-27B reaches 98.03 percent while its average F1 is 35.65. Accuracy on harmless prompts and F1 on mixed sets trade against each other along the decision threshold, and every guard in the comparison is evaluated at one threshold, 0.5. PHTest's authors document the same tension at the level of whole systems: lowering false refusals and raising robustness to jailbreaks pull in opposite directions (An et al., 2024), and OR-Bench was built to measure the over-refusal side at scale (Cui et al., 2024).

At a fixed threshold, a guard whose probabilities are better calibrated will tend to sit at a better operating point, because 0.5 then means what it says. That is a calibration effect, and Section 8 has given reasons why it may not be an objective effect. The comparison that would show a difference in capability, as opposed to a difference in where 0.5 falls, is a threshold sweep: the PHTest accuracy each guard reaches at matched recall on the unsafe sets. The paper's AUPRC values offer one indication of what that sweep might show. On ranking, which is threshold-free, two Qwen3Guard models lead LLaDA-Guard, and WildGuard matches it to within 0.07 points (Table 1).

**Prompt leakage.** In Lek et al.'s Table 6, a safe response is attached once to a safe prompt and once to an unsafe one, and the guard is asked about the response. LLaDA-Guard's unsafe-prediction rate rises by 6 points across the 250 pairs, against 42 points for Qwen3Guard and 31 for WildGuard (Lek et al., 2026). Two discriminative guards in the same table, ShieldGemma and NemoGuard, show rises of 0 and 0.4 points. The authors note why: those guards use separate templates for prompt and response classification, which "sidesteps prompt leakage by construction". The leakage result is therefore not evidence that a class-conditional objective is needed to ground response verdicts in the response; a template that asks the right question achieves more of it. What the result does show is that the class-conditional formulation makes the question explicit - the response is the reconstructed region and the prompt is context - which is a design virtue rather than a learned property. Whether leakage in a shared-template guard reflects the objective or a mis-specified task is exactly what a verdict-token LLaDA trained on response-only targets would reveal.

## 11. Localisation and Rewriting

Because LLaDA-Guard's score is a sum of per-token terms, the contribution of each token can be read off by masking it and recomputing the score. Lek et al. use this leave-one-out masking localiser to choose tokens to mask and infill under the "safe" label, iterating until a judge accepts the prompt; at a budget of five rewrites the average conversion rate over three prompt sets was 60.7 percent (Lek et al., 2026). Baselines were given an attention-saliency localiser, ranking tokens by attention to the verdict position.

Two features of this comparison limit what it attributes to the objective. First, the localisers differ. Attention weights are a weak guide to which inputs change a model's output: Jain and Wallace (2019) found learned attention frequently uncorrelated with gradient-based importance, and different attention patterns yielding equivalent predictions. Leave-one-out occlusion directly measures the effect of removing a token, and it can be applied to any guard at the same cost of one scoring pass per token. When LLaDA-Guard itself is given the attention localiser, its conversion rate at budget five is 53.0 percent against WildGuard's 52.1 percent (Lek et al., 2026). Most of the reported advantage, on this evidence, belongs to the attribution method rather than to the objective. Lek et al. describe the attention localiser as an indirect proxy themselves.

Second, the judge is Qwen3Guard-8B, a verdict-token guard - the class of model whose sensitivity to surface fragments the paper is arguing against. A rewrite that removes the words a verdict-token judge reacts to will be counted as a conversion whether or not it removes the

harm, and a localiser that finds those words will score well. The per-set conversion rates, 99.0 percent on XSTest, 55.0 percent on ToxicChat and 17.0 percent on the OpenAI moderation set for the masking localiser, are consistent with XSTest's construction around alarming but harmless words (Rottger et al., 2023). Lek et al. present rewriting as a proof of concept, and this paper reads it the same way.

Token-level evidence is also not unique to generative scoring. Qwen3Guard's streaming variant attaches a token-level classification head (Zhao et al., 2025), and the streaming content monitor of Li et al. (2025b) is trained with both response-level and token-level labels on 29,000 annotated prompt-response pairs. A per-token likelihood difference and a per-token classification head both produce localised scores; which localises harm more faithfully is untested.

## 12. What the Diagnosis Leaves Out: Attack, Cost, Streaming and Reasoning

**Adaptive attack.** LLaDA-Guard has not been evaluated against attacks designed for it, and its authors do not claim robustness to adaptive jailbreaks (Lek et al., 2026). The theoretical position is not encouraging for generative classifiers in general: detectability of adversarial inputs cannot be guaranteed even for near-optimal ones (Fetaya et al., 2019). The empirical position for masked diffusion models is specific. Their bidirectional infilling has been exploited to produce harmful content outside the region alignment monitors (Wen et al., 2025b; Ma et al., 2026), and safety behaviour in LLaDA has been shown to be sparse enough that pruning a small set of safety neurons raised attack success from 2.6 to 73.8 percent (Dumitrescu et al., 2026). Those results concern LLaDA as a generator rather than as a guard, and they do not show that LLaDA-Guard is fragile. They show that the backbone has its own attack surface, which the evaluation has not reached. A guard whose per-token scores are exposed also hands an attacker the locations to edit - a risk the authors acknowledge in their broader-impact appendix.

**Cost.** LLaDA-Guard needs two reconstruction scores per decision, each averaged over three masks, against one forward pass for a verdict-token guard (Lek et al., 2026). For a moderation layer that sees every request, a sixfold compute increase is a design constraint, not a footnote, and it bears on the comparison: a verdict-token guard given six forward passes - an ensemble over seeds or input perturbations - is the compute-matched baseline.

**Streaming.** Masked reconstruction of the response scores the whole region at once. Streaming moderation, which must decide while the response is being generated, is a stated design goal of current guards (Zhao et al., 2025) and has its own training methods (Li et al., 2025b). Whether class-conditional scoring of a growing prefix is stable is open.

**Other ways to spread supervision.** The diagnosis that one verdict token concentrates supervision has other remedies. Reasoning guards train the model to write an explanation before the verdict: GuardReasoner on 127,000 samples with 460,000 reasoning steps (Liu et al., 2025), ThinkGuard with critiques distilled from stronger models (Wen et al., 2025a), and RSafe with policy-guided reasoning followed by rule-based reinforcement learning (Zheng et al., 2025). Sadhu et al. (2026) find that compliance detectors, including a policy-conditioned guard, ignore the governing rule almost entirely, and that a step-by-step reasoning control, run on 40 of the 200 items with one prompt and one seed, showed the benchmark could be solved where every fast detector they tested stayed at chance. Consistency training over paraphrase sets reduced semantic variability in six discriminative guards by about 58 percent (Pinneri & Louizos, 2025). On the data side, Constitutional Classifiers are trained on synthetic data generated from a natural-language statement of permitted and restricted content; their authors report that in over 3,000 estimated hours of red teaming no universal jailbreak extracted comparable detail from an early prototype, which was deliberately tuned for robustness and over-refused heavily; later classifiers, evaluated with automated jailbreak tests, added an absolute 0.38 percent to production-traffic refusals (Sharma et al., 2025). These approaches keep a discriminative readout and change what the model is trained on or trained to attend to. They are the natural alternatives against which a class-conditional objective should be measured, and none appears in Lek et al.'s comparison.

## 13. An Attribution Audit

The analysis above reduces to a single requirement: to attribute a guard's behaviour to its objective, the comparison must hold the backbone, the data, the fine-tuning regime and the inference compute fixed while the objective varies, and it must measure each claimed property in each arm. Algorithm 1 states the audit.

Each step closes a specific gap identified above. Steps 1 and 3 address the averaging and recalibration alternatives of Section 8. Step 2 separates capability from operating point, as Section 10 requires. Step 4 runs the refusal-cue test of Section 6 and supplies counts so that a difference of one or two prompts is not read as a difference in kind. Arms V and H on the same masked-diffusion backbone are the comparison Section 7 found missing; the autoregressive backbone, on which the class-conditional score is computed by teacher forcing as in Li et al. (2025a) and Min et al. (2021), separates the objective from the attention mask, as Section 9 requires. Arm G tests the lambda question of Section 4: if the generative term carries the shortcut-avoidance, its effect should grow from H to G, at some in-distribution cost if the CivilComments pattern recurs. Swapping the data, the last line, isolates the mixture effect that Aegis 2.0's construction makes likely.

Table 3 states what each hypothesis predicts for the most diagnostic outcomes. The entries are the author's

**Algorithm 1.** Attributing a guard improvement to its training objective

```
Input: backbone B (one AR, one masked-diffusion), training set D, budget c
Arms (each trained on D, same adapter type and rank, same steps):
  V  = verdict-token cross-entropy, label after content
  H  = class-conditional, BCE on log-likelihood ratio + lambda * recon
  G  = class-conditional, reconstruction only (lambda -> infinity)
  R  = verdict with reasoning trace (optional discriminative alternative)
For each backbone B and arm A:
  1. Score held-out sets at matched compute c: V and R averaged over c
     perturbed passes (seeds or input masks); H and G over c masks.
  2. Discrimination: AUPRC; and F1 swept over thresholds, reporting the
     PHTest harmless accuracy at matched recall on unsafe sets.
  3. Calibration: ECE, Brier, NLL raw and after temperature scaling
     fitted on a held-out split from D's sources only.
  4. Shortcut probes: refusal-cue insertion into harmful responses;
     trigger-word benign prompts; paraphrase sets; report flip rates
     with counts and intervals.
  5. Position: move the decisive cue across filler lengths; report
     drop from each arm's own peak.
  6. Leakage: matched safe responses under safe vs unsafe prompts,
     shared template and response-only template.
  7. Localisation: occlusion localiser for every arm; judge rewrites
     with a non-guard criterion (human or policy-matched rubric).
  8. Attack: adaptive attack with white-box access to each arm's score.
Attribute to the objective only effects present in H or G vs V
  on both backbones, surviving temperature scaling and matched compute.
Attribute to data any effect that moves when D is swapped
  (e.g. WildGuardMix-only vs Aegis-only) but not across arms.
```

**Table 3.** Predicted outcomes of the audit in Algorithm 1 under each candidate explanation for LLaDA-Guard's reported gains.

| Hypothesis | V vs H, same backbone and data | Effect of temperature scaling | Effect of matched averaging | Data swap |
|---|---|---|---|---|
| Objective (C1-C3 as claimed) | H better on shortcut, position, calibration | gap shrinks but persists | gap persists | effect persists in both data |
| Backbone and attention mask | V on LLaDA close to H on LLaDA; both beat AR | little change in ranking | little change | effect persists |
| Data and fine-tuning regime | V on same data close to H | baselines close the gap | little change | effect tracks the data |
| Averaging and readout | V averaged close to H | partial closure | gap closes | effect persists |

predictions, derived from the arguments above.

Several outcomes would be informative even if they split. If H beats V on refusal-cue insertion only when trained on WildGuardMix-heavy data, the objective helps exactly where the data are most misleading, which is the regime the generative literature would predict. If calibration differences vanish after temperature scaling while shortcut differences persist, the paper's calibration claim belongs to the readout and its shortcut claim to the objective. If G beats H on shortcut probes and loses on in-distribution F1, the hybrid is trading away the generative benefit to recover accuracy, and lambda becomes a reportable operating parameter.

## 14. The Strongest Case Against This Paper

The best objection is that this paper asks for a standard no guard paper meets. Llama Guard, WildGuard and Qwen3Guard were each compared against baselines with different backbones, data and training regimes, and nobody demanded a factorial audit before crediting their data pipelines. LLaDA-Guard at least reports a same-weights ablation, an over-blocking test, a leakage test and a calibration analysis that most guard releases omit. On this view, holding it to Algorithm 1 is selective.

The objection is fair as a description of the field and does not change the conclusion. The difference is in the claim. A guard release that says "our data and training produce a better guard" makes a claim its comparison supports. A paper that says the verdict-token objective is the structural cause of shortcut learning, overconfidence and positional bias, and that a class-conditional objective removes the incentive for them, makes a causal claim about one factor among four. That claim requires the other three held fixed. The paper itself recognised this when it ran the same-weights ablation; the argument here is that the ablation did not hold what it was meant to hold.

A second objection is that the author's alternatives are speculative where Lek et al.'s results are measured. That is true, and the paper has tried to mark it. The averaging account of calibration, the regime account of the generative advantage, the data account of over-blocking and the attribution-method account of localisation are hypotheses with independent published support, not findings about LLaDA-Guard. Any of them could fail the audit. The claim is only that each is live, and that the published comparison cannot reject it.

A third objection is that Li et al. (2025a)'s CivilComments ablation, which this paper leans on,

compares an unconditional generative term with a discriminative head, while LLaDA-Guard uses a class-conditional generative term and a discriminative loss over the class-conditional likelihoods. That is correct, and Section 4 says so. The ablation is cited for what it shows - that a generative component is not sufficient - and not as a prediction of LLaDA-Guard's behaviour. Ding et al. (2020) is the closer analogue, and it favours the hybrid in the regimes it tested.

Finally, it could be argued that a masked diffusion backbone is part of the proposal, so separating objective from backbone is beside the point: the practitioner adopts both. For deployment that is right. For the scientific claim - that the objective is what fixes shortcuts - it is not, and the practical consequence differs: if the backbone's bidirectionality is doing the work, a bidirectional discriminative guard would capture it at one pass.

## 15. What Is Not Known

Whether a verdict-token guard trained on LLaDA-8B-Instruct with LLaDA-Guard's 40,000 examples and LoRA configuration would match LLaDA-Guard on any of the reported properties is not known. No such model has been reported.

Whether LLaDA-Guard carries the refusal-cue shortcut is not known. It has not been tested.

The value of lambda in LLaDA-Guard's objective, and how the reported results depend on it, are not stated.

How much of LLaDA-Guard's calibration advantage survives temperature scaling of the baselines, or matched Monte Carlo averaging of a verdict-token guard, is not known.

Whether the generative-classifier robustness advantage documented for models trained from scratch exists for instruction-tuned models fine-tuned on tens of thousands of labelled safety examples is not established in either direction for guards. The one broad pretrained comparison found the discriminative model ahead (Kasa et al., 2025), on general text classification rather than safety.

How LLaDA-Guard's calibration and accuracy behave under adaptive attack is not known.

Whether class-conditional scoring of a partial response is stable enough for streaming moderation is not known.

Whether LLaDA-Guard's results extend beyond English binary moderation is not known; its authors restrict their claims accordingly.

## 16. Studies That Would Settle It

**The two-arm study.** Train a verdict-token LLaDA and LLaDA-Guard on the same 40,000 examples with the same LoRA configuration and steps, and report every metric in Table 1 for both, with intervals. This single comparison addresses most of Sections 4, 7 and 8. If the verdict-token LLaDA matches LLaDA-Guard on calibration and over-blocking, the backbone and data carry those gains; if it does not, the objective earns its credit on that backbone.

**The cross-backbone replication.** Repeat the two arms on an autoregressive backbone of comparable size, computing the class-conditional score by teacher forcing. Effects present on both backbones are objective effects; effects present only on LLaDA belong to the attention mask or the diffusion estimator.

**The refusal-cue and data-swap study.** Train both arms on WildGuardMix alone and on Aegis 2.0 alone, and run Feng et al.'s refusal-cue insertion on all four models. This tests whether the class-conditional objective resists a shortcut that is present in the data, which is the core of claim C1.

**The lambda sweep.** Train LLaDA-Guard at several values of lambda, including a reconstruction-only model, and report in-distribution F1 alongside shortcut probes. A trade-off curve would locate the hybrid between the CivilComments arms of Li et al. (2025a).

**The recalibrated comparison.** Fit temperature scaling for every guard on a held-out split drawn from the same sources, and compare ECE, Brier and NLL again, with and without matched averaging for verdict-token guards. If the ranking survives, the calibration claim is robust to the obvious alternative.

**The adaptive attack.** Attack LLaDA-Guard with white-box access to its score and its per-token contributions, and report calibration under attack as Hong et al. (2026) did for verdict-token guards. A generative guard that keeps its calibration under attack would be a substantive result, and one the theory does not guarantee.

## 17. Conclusion

LLaDA-Guard is a good guard on the benchmarks it was tested on, and its calibration in particular is better than any released verdict-token guard in its comparison. The paper that introduces it offers an explanation - that a class-conditional objective removes the incentive for the shortcut, overconfidence and positional failures of verdict-token training - and the explanation is plausible. It is not yet tested. The loss that is optimised is a discriminative loss on a generative parameterisation, which does not exclude shortcut solutions; the evidence that generative classifiers avoid shortcuts comes from pure class-conditional training of models from scratch, a regime unlike guard training, and has a documented reversal with pretrained weights; the one same-backbone comparison reads a verdict the model was never trained to give; and each claimed benefit has a published mechanism in the data, the fine-tuning regime, the attention mask or the Monte Carlo averaging that changed alongside the objective. A likelihood ratio can read the whole moderated text, but nothing in the training requires it to. Whether it does is a measurable question, and the comparison in Algorithm 1 would answer it.

## References


An, B., Zhu, S., Zhang, R., Panaitescu-Liess, M., Xu, Y., & Huang, F. (2024). Automatic Pseudo-Harmful Prompt Generation for Evaluating False Refusals in Large Language Models. arXiv:2409.00598. doi:10.48550/arXiv.2409.00598.

Biderman, D., Portes, J., Gonzalez Ortiz, J. J., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., et al. (2024). LoRA Learns Less and Forgets Less. arXiv:2405.09673. doi:10.48550/arXiv.2405.09673.

Chen, H., Dong, Y., Wang, Z., Yang, X., Duan, C., Su, H., & Zhu, J. (2023). Robust Classification via a Single Diffusion Model. arXiv:2305.15241. doi:10.48550/arXiv.2305.15241.

Clark, K., & Jaini, P. (2023). Text-to-Image Diffusion Models are Zero-Shot Classifiers. arXiv:2303.15233. doi:10.48550/arXiv.2303.15233.

Clavie, B., Cooper, N., & Warner, B. (2025). It's All in The [MASK]: Simple Instruction-Tuning Enables BERT-like Masked Language Models As Generative Classifiers. arXiv:2502.03793. doi:10.48550/arXiv.2502.03793.

Cui, J., Chiang, W., Stoica, I., & Hsieh, C. (2024). OR-Bench: An Over-Refusal Benchmark for Large Language Models. arXiv:2405.20947. doi:10.48550/arXiv.2405.20947.

Ding, X., Liu, T., Chang, B., Sui, Z., & Gimpel, K. (2020). Discriminatively-Tuned Generative Classifiers for Robust Natural Language Inference. arXiv:2010.03760. doi:10.48550/arXiv.2010.03760.

Dumitrescu, E., Lek, G., Chen, L. Y., & Decouchant, J. (2026). Diffusion LLMs as Targets and Adversaries: Mechanistic Safety Exploits. arXiv:2608.07430. doi:10.48550/arXiv.2608.07430.

Feng, Y., Zang, C., Shen, C., Miao, R., Teng, G., Cai, W., & Ye, J. (2026). When Refusal Looks Safe: The Refusal-Cue Shortcut in Safety Guard Models. arXiv:2608.03201. doi:10.48550/arXiv.2608.03201.

Fetaya, E., Jacobsen, J., Grathwohl, W., & Zemel, R. (2019). Understanding the Limitations of Conditional Generative Models. arXiv:1906.01171. doi:10.48550/arXiv.1906.01171.

Gangal, V., Arora, A., Einolghozati, A., & Gupta, S. (2019). Likelihood Ratios and Generative Classifiers for Unsupervised Out-of-Domain Detection In Task Oriented Dialog. arXiv:1912.12800. doi:10.48550/arXiv.1912.12800.

Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut Learning in Deep Neural Networks. arXiv:2004.07780. doi:10.48550/arXiv.2004.07780.

Ghosh, S., Varshney, P., Sreedhar, M. N., Padmakumar, A., Rebedea, T., Varghese, J. R., & Parisien, C. (2025). Aegis2.0: A Diverse AI Safety Dataset and Risks Taxonomy for Alignment of LLM Guardrails. arXiv:2501.09004. doi:10.48550/arXiv.2501.09004.

Grathwohl, W., Wang, K., Jacobsen, J., Duvenaud, D., Norouzi, M., & Swersky, K. (2019). Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One. arXiv:1912.03263. doi:10.48550/arXiv.1912.03263.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. arXiv:1706.04599. doi:10.48550/arXiv.1706.04599.

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). Annotation Artifacts in Natural Language Inference Data. arXiv:1803.02324. doi:10.48550/arXiv.1803.02324.

Hackett, W., Birch, L., Trawicki, S., Suri, N., & Garraghan, P. (2025). Bypassing LLM Guardrails: An Empirical Analysis of Evasion Attacks against Prompt Injection and Jailbreak Detection Systems. arXiv:2504.11168. doi:10.48550/arXiv.2504.11168.

Han, S., Rao, K., Ettinger, A., Jiang, L., Lin, B. Y., Lambert, N., Choi, Y., & Dziri, N. (2024). WildGuard: Open One-Stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs. arXiv:2406.18495. doi:10.48550/arXiv.2406.18495.

Hong, J., Jung, M., & Kim, M. (2026). Guard Models Are Overconfident Where Base Models Are Uncertain. arXiv:2609.36477. doi:10.48550/arXiv.2609.36477.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., et al. (2023). Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. arXiv:2312.06674. doi:10.48550/arXiv.2312.06674.

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. arXiv:1902.10186. doi:10.48550/arXiv.1902.10186.

Jaini, P., Clark, K., & Geirhos, R. (2023). Intriguing properties of generative classifiers. arXiv:2309.16779. doi:10.48550/arXiv.2309.16779.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Zhang, C., Sun, R., et al. (2023). BeaverTails: Towards Improved Safety Alignment of LLM via a Human-Preference Dataset. arXiv:2307.04657. doi:10.48550/arXiv.2307.04657.

Ji, J., Hong, D., Zhang, B., Chen, B., Dai, J., Zheng, B., Qiu, T., Zhou, J., et al. (2024). PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference. arXiv:2406.15513. doi:10.48550/arXiv.2406.15513.

Kasa, S. R., Gupta, K., Roychowdhury, S., Kumar, A., Biruduraju, Y., Kasa, S. K., Pattisapu, N. P., Bhattacharya, A., et al. (2025). Generative or Discriminative? Revisiting Text Classification in the Era of Transformers. arXiv:2506.12181. doi:10.48550/arXiv.2506.12181.

Kaushik, D., Hovy, E., & Lipton, Z. C. (2019). Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. arXiv:1909.12434. doi:10.48550/arXiv.1909.12434.

Kirichenko, P., Izmailov, P., & Wilson, A. G. (2022). Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. arXiv:2204.02937. doi:10.48550/arXiv.2204.02937.

Kumar, P. (2026). Discrete Diffusion Language Models Are Training-Free Multi-Label Classifiers. arXiv:2608.14649. doi:10.48550/arXiv.2608.14649.

Lek, G., Malan, A., Zhu, C., Chen, P., Birke, R., & Chen, L. (2026). Safety Reconstructed: Generative Modeling via Masked Diffusion Builds Strong Safety Guardrails. arXiv:2609.33634. doi:10.48550/arXiv.2609.33634.

Li, Y., Bradshaw, J., & Sharma, Y. (2018). Are Generative Classifiers More Robust to Adversarial Attacks?. arXiv:1802.06552. doi:10.48550/arXiv.1802.06552.

Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., & Pathak, D. (2023). Your Diffusion Model is Secretly a Zero-Shot Classifier. arXiv:2303.16203. doi:10.48550/arXiv.2303.16203.

Li, H., & Liu, X. (2024). InjecGuard: Benchmarking and Mitigating Over-defense in Prompt Injection Guardrail Models. arXiv:2410.22770. doi:10.48550/arXiv.2410.22770.

Li, A. C., Kumar, A., & Pathak, D. (2025a). Generative Classifiers Avoid Shortcut Solutions. arXiv:2512.25034. doi:10.48550/arXiv.2512.25034.

Li, Y., Sheng, Q., Yang, Y., Zhang, X., & Cao, J. (2025b). From Judgment to Interference: Early Stopping LLM Harmful Outputs via Streaming Content Monitoring. arXiv:2506.09996. doi:10.48550/arXiv.2506.09996.

Lin, Z., Wang, Z., Tong, Y., Wang, Y., Guo, Y., Wang, Y., & Shang, J. (2023). ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation. arXiv:2310.17389. doi:10.48550/arXiv.2310.17389.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172. doi:10.48550/arXiv.2307.03172.

Liu, H., Huang, H., Gu, X., Wang, H., & Wang, Y. (2024). On Calibration of LLM-based Guard Models for Reliable Content Moderation. arXiv:2410.10414. doi:10.48550/arXiv.2410.10414.

Liu, Y., Gao, H., Zhai, S., He, Y., Xia, J., Hu, Z., Chen, Y., Yang, X., et al. (2025). GuardReasoner: Towards Reasoning-based LLM

Safeguards. arXiv:2501.18492. doi:10.48550/arXiv.2501.18492.

Ma, Y., Zhao, Z., Liu, X., Xue, M., Zhao, Y., & Xiao, C. (2026). MaskForge: Structure-Aware Adaptive Attacks for Jailbreaking Diffusion Large Language Models. arXiv:2606.04027. doi:10.48550/arXiv.2606.04027.

Mangaokar, N., Hooda, A., Choi, J., Chandrashekaran, S., Fawaz, K., Jha, S., & Prakash, A. (2024). PRP: Propagating Universal Perturbations to Attack Large Language Model Guard-Rails. arXiv:2402.15911. doi:10.48550/arXiv.2402.15911.

Markov, T., Zhang, C., Agarwal, S., Eloundou, T., Lee, T., Adler, S., Jiang, A., & Weng, L. (2022). A Holistic Approach to Undesired Content Detection in the Real World. arXiv:2208.03274. doi:10.48550/arXiv.2208.03274.

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. arXiv:1902.01007. doi:10.48550/arXiv.1902.01007.

Min, S., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2021). Noisy Channel Language Model Prompting for Few-Shot Text Classification. arXiv:2108.04106. doi:10.48550/arXiv.2108.04106.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., & Lakshminarayanan, B. (2018). Do Deep Generative Models Know What They Don't Know?. arXiv:1810.09136. doi:10.48550/arXiv.1810.09136.

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., et al. (2025). Large Language Diffusion Models. arXiv:2502.09992. doi:10.48550/arXiv.2502.09992.

Padhi, I., Nagireddy, M., Cornacchia, G., Chaudhury, S., Pedapati, T., Dognin, P., Murugesan, K., Miehling, E., et al. (2024). Granite Guardian. arXiv:2412.07724. doi:10.48550/arXiv.2412.07724.

Peysakhovich, A., & Lerer, A. (2023). Attention Sorting Combats Recency Bias In Long Context Language Models. arXiv:2310.01427. doi:10.48550/arXiv.2310.01427.

Pinneri, C., & Louizos, C. (2025). Guarding the Meaning: Self-Supervised Training for Semantic Robustness in Guard Models. arXiv:2511.10665. doi:10.48550/arXiv.2511.10665.

Ren, J., Liu, P. J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., & Lakshminarayanan, B. (2019). Likelihood Ratios for Out-of-Distribution Detection. arXiv:1906.02845. doi:10.48550/arXiv.1906.02845.

Rottger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2023). XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. arXiv:2308.01263. doi:10.48550/arXiv.2308.01263.

Sadhu, S., Sengupta, A., Sankarapu, V. K., & Seth, P. (2026). What Do Compliance Detectors Read? An Audit of Activation Probes and Guard Models. arXiv:2608.16852. doi:10.48550/arXiv.2608.16852.

Schott, L., Rauber, J., Bethge, M., & Brendel, W. (2018). Towards the first adversarially robust neural network model on MNIST. arXiv:1805.09190. doi:10.48550/arXiv.1805.09190.

Sharma, M., Tong, M., Mu, J., Wei, J., Kruthoff, J., Goodfriend, S., Ong, E., Peng, A., et al. (2025). Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming. arXiv:2501.18837. doi:10.48550/arXiv.2501.18837.

Wei, Z., Liu, Y., & Erichson, N. B. (2024). Emoji Attack: Enhancing Jailbreak Attacks Against Judge LLM Detection. arXiv:2411.01077. doi:10.48550/arXiv.2411.01077.

Wen, X., Zhou, W., Mo, W. J., & Chen, M. (2025a). ThinkGuard: Deliberative Slow Thinking Leads to Cautious Guardrails. arXiv:2502.13458. doi:10.48550/arXiv.2502.13458.

Wen, Z., Qu, J., Chen, Z., Lu, X., Liu, D., Liu, Z., Wu, R., Yang, Y., et al. (2025b). The Devil behind the mask: An emergent safety vulnerability of Diffusion LLMs. arXiv:2507.11097. doi:10.48550/arXiv.2507.11097.

Wu, X., Wang, Y., Jegelka, S., & Jadbabaie, A. (2025). On the Emergence of Position Bias in Transformers. arXiv:2502.01951. doi:10.48550/arXiv.2502.01951.

Yang, W., Kirichenko, P., Goldblum, M., & Wilson, A. G. (2022). Chroma-VAE: Mitigating Shortcut Learning with Generative Classifiers. arXiv:2211.15231. doi:10.48550/arXiv.2211.15231.

Yogatama, D., Dyer, C., Ling, W., & Blunsom, P. (2017). Generative and Discriminative Text Classification with Recurrent Neural Networks. arXiv:1703.01898. doi:10.48550/arXiv.1703.01898.

Zeng, W., Liu, Y., Mullins, R., Peran, L., Fernandez, J., Harkous, H., Narasimhan, K., Proud, D., et al. (2024). ShieldGemma: Generative AI Content Moderation Based on Gemma. arXiv:2407.21772. doi:10.48550/arXiv.2407.21772.

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models. arXiv:2102.09690. doi:10.48550/arXiv.2102.09690.

Zhao, H., Yuan, C., Huang, F., Hu, X., Zhang, Y., Yang, A., Yu, B., Liu, D., et al. (2025). Qwen3Guard Technical Report. arXiv:2510.14276. doi:10.48550/arXiv.2510.14276.

Zheng, C., Wu, G., Bao, F., Cao, Y., Li, C., & Zhu, J. (2023). Revisiting Discriminative vs. Generative Classifiers: Theory and Implications. arXiv:2302.02334. doi:10.48550/arXiv.2302.02334.

Zheng, J., Ji, X., Lu, Y., Cui, C., Zhao, W., Deng, G., Liang, Z., Zhang, A., et al. (2025). RSafe: Incentivizing proactive reasoning to build robust and adaptive LLM safeguards. arXiv:2506.07736. doi:10.48550/arXiv.2506.07736.

Zhu, F., Wang, R., Nie, S., Zhang, X., Wu, C., Hu, J., Zhou, J., Chen, J., et al. (2025). LLaDA 1.5: Variance-Reduced Preference Optimization for Large Language Diffusion Models. arXiv:2505.19223. doi:10.48550/arXiv.2505.19223.

© 2026 Pranay Mahendrakar. Licensed under CC BY 4.0.

DOI: 10.5281/zenodo.24545381