Safety-Flag: A Unified Benchmark for the Reliability and Calibration of LLM Content Moderators

Yibo Hu
Illinois Institute of Technology
Chicago, Illinois, USA
yhu89@illinoistech.edu

Abstract

Large language models are increasingly used for content moderation, but most evaluations still report aggregate accuracy on individual benchmarks. We introduce Safety-Flag, which places seven widely used safety benchmarks (BeaverTails, XSTest, Ethics, WildGuard, Aegis, ToxiChat, and ToxiGen) into a single balanced flag / do-not-flag protocol. We release item-level decisions and confidence scores for six general-purpose LLMs and four dedicated guards, together with three reference models, evaluated on the same items. Safety-Flag measures three dimensions of moderator reliability: error direction, probability calibration, and confidence-based error ranking for human review. They often disagree. Aggregate accuracy does not reveal error direction: one model flags 85% of benign content, whereas another misses 54% of harmful content. All six general-purpose models are overconfident; fitting one temperature per model reduces calibration error by 2.8–6.0× without changing predicted labels or confidence ordering. Confidence-based abstention lowers selective risk for every model, although the gains depend on how well confidence ranks errors. Dedicated guards produce fewer false alarms and are better calibrated, but several have higher miss rates outside their documented coverage. We release the benchmark, fixed item lists, evaluation code, per-item model outputs, and leaderboard at: https://github.com/yibo-hu-lab/safety-flag-benchmark.

CCS Concepts

• **Computing methodologies** → **Machine learning**; • **Security and privacy** → *Social aspects of security and privacy*.

Keywords

content moderation, large language models, calibration, selective prediction, benchmark resource, trustworthy AI

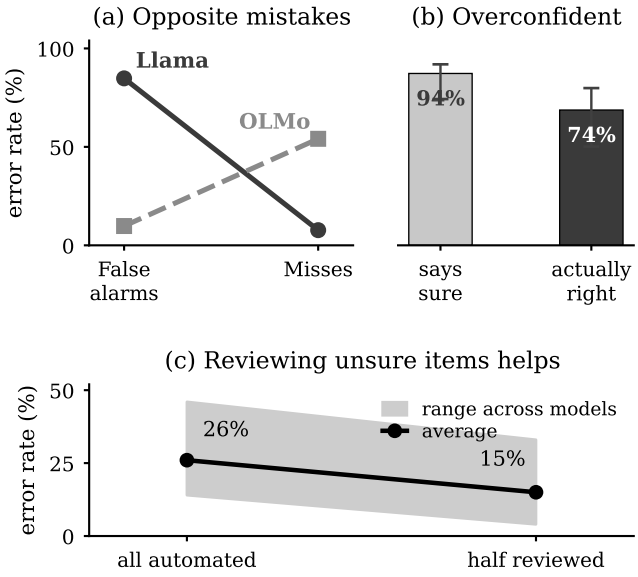

Figure 1: Aggregate accuracy hides three deployment-relevant differences. (a) Llama and OLMo fail in opposite directions: one over-flags benign content, while the other misses harmful content. (b) Averaged across models, stated confidence exceeds actual accuracy. (c) Reviewing the least-confident half lowers the error rate on the items that remain automated, although the gain varies across models.

1 Introduction

Large language models (LLMs) are increasingly deployed as content moderators, deciding whether a post, prompt, or model response should be flagged as unsafe [14, 19]. The decision is consequential in both directions. Over-flagging benign content silences users and overwhelms human review; missing harmful content defeats the purpose of moderation.

Accuracy and F1 do not answer three questions that matter in deployment. First, does a model mainly flag benign content or miss harmful content? Second, does its reported confidence match how often it is correct? Third, when low-confidence items are sent to a human, does the remaining automated error actually decrease? We call these three dimensions *error direction*, *probability calibration*, and *error ranking*. Figure 1 shows all three on real moderators: two models fail in opposite directions, every model is overconfident, and human review helps some models far more than others. Aggregate accuracy reveals none of this.

Existing evaluations cannot answer these questions together. Safety benchmarks use different labels, harm taxonomies, prompts, and item sets, so results are usually reported one benchmark at a time and are not directly comparable. Prior work has studied calibration for LLMs and dedicated guards [9, 17, 18, 25], or learned when to escalate items to human reviewers [2]. No common evaluation compares general-purpose LLMs and dedicated guards on the same fixed items while measuring all three together.

We introduce Safety-Flag to provide that evaluation. We recast seven widely used safety benchmarks [3, 7, 10–12, 15, 21] into one balanced *flag / do-not-flag* protocol. We evaluate six general-purpose LLMs, four dedicated guards, and three reference models on the same released item sets, and release every item-level decision and confidence score. We balance benign and harmful items because

Table 1: Scope relative to the closest work. *Gen.* and *Guard* denote the evaluated system types; *Items* denotes paired evaluation on identical items; *Dir.*, *Cal.*, and *Rank* denote error direction, calibration, and error ranking; *Rel.* denotes a public per-item release.

Work	Gen.	Guard	Items	Dir.	Cal.	Rank	Rel.
Liu et al. [18]	–	✓	–	–	✓	–	–
Bachar et al. [2]	✓	–	–	–	–	✓	–
Safety-Flag	✓	✓	✓	✓	✓	✓	✓

a harmful-only benchmark rewards a model that flags everything and hides its false alarms.

Across models, a model can look strong on one dimension and weak on another. Model choice therefore depends on which failure a deployment can least afford.

Safety-Flag makes four contributions.

(1) **A reproducible benchmark resource.** We unify seven safety benchmarks under one balanced binary task with fixed items, a common prompt, and released item-level outputs.
(2) **A direct account of error direction.** We show that models range from systematic over-flagging to systematic under-flagging, and that these profiles are stable across seeds and leave-one-benchmark-out analyses.
(3) **Separate evaluations of calibration and error ranking.** Temperature scaling improves probability calibration without changing decisions or confidence ordering, while the benefit of human review depends on how well confidence ranks errors.
(4) **A paired comparison of general-purpose LLMs and guards.** On identical items, dedicated guards reduce false alarms and improve calibration, but several incur higher miss rates beyond their documented coverage.

2 Related Work

LLM safety and moderation evaluation. Purpose-built moderation systems such as Llama Guard [14] and holistic content classifiers [19] are evaluated on their own test sets. General safety benchmarks (BeaverTails [15], XSTest [21], the Ethics suite [12], WildGuard [10], Aegis [7], ToxiChat [3], ToxiGen [11], and SafetyBench [27]) each probe a slice of the harm space but are seldom unified into a protocol that supports cross-benchmark comparison, and are usually reported as accuracy or F1 without a calibration or abstention view. Safety-Flag evaluates them under one protocol and reports both decision errors and confidence-based reliability.

Calibration and confidence. Calibration error (ECE) and reliability diagrams are standard for classifiers [9, 13, 20], and temperature scaling is the canonical post-hoc fix [9]. For LLMs specifically, verbalized confidence [24, 25] and the question of whether models "know what they know" [17] are active topics. We apply both token-logprob and verbalized confidence to the moderation-flag decision, and test whether post-hoc recalibration improves the resulting confidence estimates.

Closest prior work. Two efforts are directly related. Liu et al. [18] audit the calibration of nine *dedicated* guard models across twelve benchmarks, documenting overconfidence and testing post-hoc fixes including temperature scaling. Bachar et al. [2] learn an escalation meta-model from logprob, entropy, and verbalized-confidence features for cost-aware selective classification in human-AI moderation. Safety-Flag extends these calibration and escalation studies by evaluating general-purpose LLMs and dedicated guards together on fixed items, measuring error direction, probability calibration, and confidence-based error ranking under one protocol (Table 1).

Selective prediction. Allowing a classifier to abstain and measuring the resulting coverage–risk trade-off is a classical framework [5, 6]; conformal abstention has recently been applied to LLM hallucination [1]. We use it as the deploy-relevant reliability metric for moderation, where abstaining means routing an item to a human reviewer, and we compare confidence-based abstention with random abstention.

3 The Benchmark

Safety-Flag fixes the task, item set, prompt, and per-item output schema across all evaluated models (Figure 2). This section describes the construction.

The seven benchmarks. We draw on seven complementary safety benchmarks:

- **BeaverTails** [15]: a broad harm taxonomy over prompt–response pairs.
- **XSTest** [21]: benign prompts designed to elicit exaggerated-safety refusals.
- **Ethics** [12]: commonsense moral judgments without explicit toxicity cues.
- **WildGuard** [10]: adversarial and jailbreak-style prompts.
- **Aegis** [7]: a fine-grained safety taxonomy with ensemble-derived labels.
- **ToxiChat** [3]: dialogue-level offensiveness with stance context.
- **ToxiGen** [11]: implicit group-targeted toxicity and benign identity mentions (human-annotated subset).

Together they cover broad harms, exaggerated-safety behavior, moral violations, adversarial prompts, dialogue context, and implicit group-targeted toxicity. Their differences in domain, adversariality, and harm prevalence allow us to test whether reliability patterns persist across settings.

Fixed and balanced items. We map every benchmark to the same binary decision: flag the content as unsafe (label `(A)`) or do not flag it (label `(B)`). We sample 198–200 items per benchmark, with approximately equal numbers of harmful and benign items, and release the exact item IDs, so every model is evaluated on the same instances. Class balancing prevents a model from appearing strong simply by flagging everything and gives false alarms and misses equal weight.

Sample size. We use ≈ 200 items per benchmark to estimate error direction and calibration with bootstrap confidence intervals. Subsampling shows that the macro-F1 ranking and every model's error-direction sign are already stable at $n \approx 100$ (Spearman $\rho \geq 0.99$,

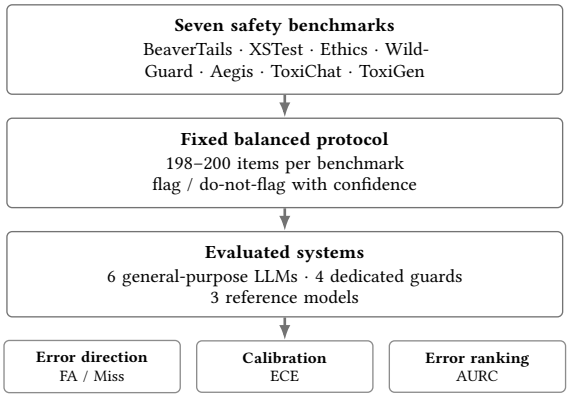

Figure 2: Overview of the Safety-Flag evaluation. Seven benchmarks are mapped to a fixed balanced flag-plus-confidence protocol, used to evaluate general-purpose LLMs and dedicated guards along three reliability dimensions: error direction, probability calibration, and error ranking.

sign agreement 100%); the ECE ranking reaches $\rho = 0.95$ at $n \approx 150$ (Appendix B.1).

Preserving benchmark semantics. Table 2 shows each benchmark's judged unit, released label, and mapping to *flag*. The benchmarks encode different notions of unsafe content, and we do not collapse them into a universal safety policy. We standardize the evaluation interface and reliability metrics while retaining the source labels, and report per-benchmark results throughout so that no single mapping drives the leaderboard.

Label harmonization. Each item retains its source benchmark's released label, mapped to flag / do-not-flag by the deterministic rule in Table 2. We audit this mapping on 150 balanced items (25 from each of the six benchmarks that require cross-scheme harmonization). Two judges from different vendors, gpt-5.5 and claude-opus-4-8, re-label the content under the released policy without seeing our labels. Agreement with the mapped labels is high (gpt-5.5 $\kappa = 0.89$; Claude $\kappa = 0.87$), and the judges agree with each other at $\kappa = 0.92$; per-benchmark results appear in Appendix A.3. ToxiGen uses its released human toxicity scores directly: scores 4–5 map to flag, scores 1–2 map to do not flag, and score-3 items are excluded.

Standardized flag-plus-confidence prompt. Each item is presented with the same instruction asking the model to (i) choose `(A)`/`(B)` and (ii) state a confidence. We record two confidence signals per decision: the *token-logprob* confidence conf_{lp} (the model's probability on its chosen option label, from the option-token logprobs) and the model's *verbalized* confidence on a 1–10 scale, rescaled to $[0, 1]$.

Metrics. We report accuracy and macro-F1 as aggregate decision-quality measures, followed by metrics for the three reliability dimensions. For error direction, we report the false-alarm rate FA = Pr(flag | benign) and miss rate Miss = Pr(no flag | harmful). For calibration, we report expected calibration error (ECE), its class-conditional decomposition, and negative log-likelihood (NLL). For error ranking, we sort items by confidence, measure the error rate on the most-confident fraction of items, and summarize the resulting risk–coverage curve by AURC [6]. We obtain cell-level 95% confidence intervals with 2,000 item-bootstrap resamples.

Intended uses. The resource supports two workflows. For *model selection*, a deployer compares candidate moderators on error direction, calibration, and error ranking on identical items, in place of a single accuracy number. For *reliability research*, the released per-item verdicts and confidences are a testbed for calibration, abstention, and escalation methods, with no model rerun. Both use the same fixed items, so results stay comparable as new models are added.

4 Setup

We evaluate six open-weight instruction-tuned LLMs spanning four families and two scales: Qwen2.5-7B and Qwen2.5-32B [26], Llama-3.1-8B [8], Mistral-7B [16], Gemma-2-9B [22], and OLMo-2-7B [23]. We run all of them through one shared evaluation pipeline and decode the verdict greedily.

We pose each item as a two-option multiple-choice question: option (A) flags the content, option (B) does not. A fixed system instruction and a round-one template ask for a JSON verdict of the form `{judgment:(A), confidence:n}`. From each answer we read two confidence signals. The token-logprob confidence comes from the two option-token logprobs after a fixed `judgment` prefix, renormalized over {(A),(B)}, so flag_prob = $P((\text{A}))$ and $\text{conf}_{\text{lp}} = \max(P((\text{A})), P((\text{B})))$. The verbalized confidence is the model's own 1–10 score, rescaled to $[0, 1]$. Gemma has no system role, so we fold the system text into its first user turn.

Primary ranking. Qwen2.5-32B runs at a single seed and mixes precision (fp16 instruct on WildGuard, Aegis, ToxiChat; quantized on the other four), so we report it as an indicative scale point rather than include it in the primary ranking.

Output validity and fixed items. Parsing succeeds on all but 2 of 8,388 outputs (< 0.03%), so we filter no cell on output validity. We retain every model×benchmark cell and score future models on the same released list of 198–200 balanced item IDs, constructed once from the IDs available for every current model. We release the prompts, the reproduction code, and the label harmonization with the resource.

Uncertainty signals. Our main analyses use token-logprob and verbalized confidence. The sampled-answer agreement analysis (Table 14) uses five generations per item.

Reference models. We add three reference rows, scored on the identical items under the same protocol but kept out of the primary ranking. R1-Distill-Llama-8B [4] is open-weight and reasoning-distilled from Llama-3.1-8B; we run it in a reason-then-decide mode and read its logprob confidence on the post-reasoning decision, exactly as for the other open models. We query gpt-4.1-mini and gpt-5.4-mini through the OpenAI API on the same prompt and item IDs. For gpt-4.1-mini we read the option-token logprobs as usual; gpt-5.4-mini is a reasoning model whose API hides them, so we report only its verdict and verbalized confidence and leave its logprob ECE and AURC blank. Each reference model is a single run.

Recalibration and abstention protocols. For temperature scaling we fit a single scalar temperature T per model on the signed

Table 2: Label mapping used to create the common binary task. Each benchmark retains its original judged unit and released source label. The source's unsafe class maps to *flag*, and flag and do-not-flag items are approximately balanced. Exact field-level mappings and exclusions are documented with the released resource.

Benchmark	Judged unit	Native label	Flagged (positive) class	Balance
BeaverTails	prompt–response pair	14-way harm taxonomy / is_safe	any harm category present	≈100/100
XSTest	user prompt	safe vs. unsafe-contrast type	unsafe-contrast prompt	100/100
Ethics (commonsense)	described action	moral vs. immoral (`label_raw`)	immoral action	94/106
WildGuard	prompt (incl. adversarial)	harmful vs. benign	harmful prompt	100/100
Aegis	user prompt / interaction	safe vs. unsafe (+ violated categories)	unsafe interaction	100/100
ToxiChat	user query / dialogue	toxic vs. non-toxic	toxic content	100/100
ToxiGen	statement about a group	human toxicity (1–5)	toxic (≥ 4) vs. benign (≤ 2)	100/100

Table 3: Reliability leaderboard on identical items, ranked by macro-F1. *FA* is the false-alarm rate and *Miss* is the harmful-item miss rate. Qwen2.5-32B[†] (mixed precision) and the three reference models (R1-Distill-Llama-8B[§], gpt-4.1-mini, gpt-5.4-mini[‡]) are shown for context but excluded from the primary ranking; gpt-5.4-mini[‡] exposes no option logprobs, so its ECE and AURC are blank.

Model	Acc	F1	FA↓	Miss↓	ECE↓	AURC↓
Gemma-2-9B	0.827	0.824	0.272	0.075	0.168	0.082
Qwen2.5-7B	0.795	0.793	0.151	0.261	0.195	0.096
Mistral-7B	0.745	0.733	0.403	0.102	0.197	0.137
OLMo-2-7B	0.678	0.639	0.098	0.541	0.127	0.246
Llama-3.1-8B	0.536	0.443	0.848	0.077	0.368	0.300
Qwen2.5-32B[†]	0.864	0.863	0.186	0.087	0.126	0.048
Reference models (reasoning-distilled and API; not ranked)						
gpt-5.4-mini[‡]	0.856	0.853	0.097	0.194	N/A	N/A
gpt-4.1-mini	0.847	0.843	0.057	0.249	0.146	0.093
R1-Distill-Llama-8B[§]	0.815	0.813	0.165	0.208	0.109	0.121

flag logit $\log \frac{P((A))}{P((B))}$ by minimizing negative log-likelihood on a held-out half of the pooled items, and evaluate ECE of the recalibrated conf_{lp} on the other half. For selective prediction we rank items by confidence and report risk at fixed coverage, comparing against a *random*-abstention baseline (risk equals the full error rate in expectation).

5 Results

We first report the aggregate leaderboard, then unpack what it hides: error direction, calibration, and whether confidence can identify items for human review. We then compare dedicated guards and translate these trade-offs into deployment costs.

5.1 The reliability leaderboard

We begin with the aggregate leaderboard. Table 3 ranks the five primary models by macro-F1. Gemma-2-9B leads (F1 0.824), followed by Qwen2.5-7B; Llama-3.1-8B is last (F1 0.443). Qwen2.5-32B, run at mixed precision (§4), posts the highest score of all (F1 0.863) but we include it as an indicative scale point rather than rank it. Accuracy alone, however, is a poor summary of a moderator: it does not reveal error direction, which the false-alarm and miss columns expose. The remaining analyses explain the differences hidden by macro-F1.

Three reference rows sit outside the ranking (Table 3). The two API models, gpt-4.1-mini and gpt-5.4-mini, score competitively but neither surpasses the strongest open model, Qwen2.5-32B, and both under-flag: gpt-5.4-mini leaves nearly a fifth of harmful content unflagged. R1-Distill-Llama-8B, reasoning-distilled from Llama-3.1-8B, shifts that model's flag-everything behavior toward a balanced profile with better logprob calibration than the instruct base, a change temperature scaling alone cannot produce.

5.2 Error direction differs sharply across models

We first ask which error each model makes. Figure 3 shows that the models split into opposite error directions. Llama-3.1-8B is a systematic *over-flagger*: it raises a false alarm on 85% of benign content on average while missing only 8% of harmful content. It flags almost everything, the behavior the balanced benign items are designed to catch. OLMo-2-7B shows the opposite profile, a systematic *under-flagger*: its false-alarm rate is only 0.10 but it *misses* 54% of harmful content. OLMo's low false-alarm rate comes with the worst miss rate in the suite, so it reflects a strong tendency to answer "safe." Across models, lower false-alarm rates tend to coincide with higher miss rates, and both errors are benchmark-dependent: even the balanced models (Qwen, Gemma) over-flag more on the exaggerated-safety probes of XSTest than on ToxiChat. Figure 4 makes the two regimes concrete on individual items: the over-flaggers Llama-3.1-8B and Mistral-7B flag a benign self-disclosure, whereas the under-flagger OLMo-2-7B passes an exclusionary statement that carries no explicit slur.

The error directions are seed-stable. Across 25 cells (model by benchmark) with three random seeds, the standard deviation of both error rates across seeds is at most 0.012 (medians 0.000): the greedy

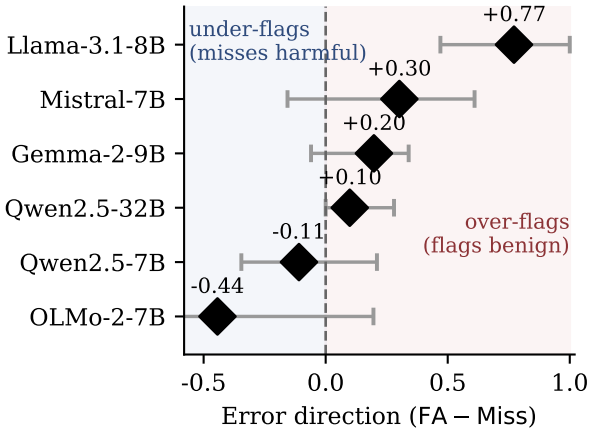

Figure 3: Models fail in different directions. Error-direction score (false-alarm − miss rate) per model: positive (right) over-flags benign content, negative (left) under-flags harmful content. Diamonds are per-model means over the benchmarks; whiskers give the across-benchmark range.

flag decisions are effectively deterministic, so the error directions are not a seed artifact. Dropping any one benchmark leaves the macro-F1 and error-direction rankings unchanged (Spearman $\rho = 1.0$) and preserves the general-vs-guard comparison.

The confidence intervals confirm the separation. We summarize each model's error direction by its mean error-direction score FA − Miss across benchmarks (Figure 3): positive for an over-flagger, negative for an under-flagger. The two extremes are opposite in sign. Llama-3.1-8B sits at +0.77 and OLMo-2-7B at −0.44, and their 95% item-bootstrap CIs (2000 resamples) exclude both zero and each other, confirming that the contrast is not explained by sampling variation; the balanced models fall in between. Across models the two errors trade off: a lower false-alarm rate comes with a higher miss rate rather than with uniformly better accuracy.

5.3 Calibration varies sharply across models and classes

We next ask whether each model's confidence matches how often it is correct. Table 4 shows large differences: pooled logprob ECE ranges from 0.127 (OLMo-2-7B) to 0.368 (Llama-3.1-8B). The bootstrap intervals are disjoint, and an equal-*mass* binning gives the same ordering, so the gap is not an artifact of one binning scheme. We use pooled ECE as the primary measure to reduce the upward bias of ECE estimates in small per-benchmark cells. Verbalized confidence is *better* calibrated than logprob confidence for every model, though it is a coarser 1–10 signal, so we report both. Calibration is also benchmark-dependent.

The aggregate number does not tell the whole story. The class-conditional ECE columns of Table 4 follow the dominant error type: Llama, the over-flagger, is near-perfectly calibrated on harmful items (ECE 0.030) but far more miscalibrated on the benign items it wrongly flags (ECE 0.730); OLMo, the under-flagger, shows the reverse.

Aggregate ECE alone does not rank moderators. OLMo has the lowest pooled ECE and NLL among the ranked models but also the highest miss rate. The class-conditional and selective-risk analyses make this difference visible. Figure 5 compares the lowest- and highest-ECE models.

5.4 Temperature scaling improves calibration but preserves decisions

We then test whether a simple post-hoc correction can fix this mismatch. Figure 6 shows that the overconfidence is correctable. Every general-purpose model is substantially overconfident: the temperature that best calibrates it ranges from 2.7 (OLMo) to 9.0 (Llama). All fitted temperatures exceed 1, indicating overconfidence in the decision logits. Across 20 random fit/eval splits a single scalar temperature cuts logprob ECE by 2.8–6.0× with small per-split variance. A deployment can therefore obtain calibrated thresholds by fitting one temperature on a held-out validation set.

Temperature scaling is monotone, so it leaves the ranking of items by confidence, and therefore the coverage–risk curve and the false-alarm and miss rates, unchanged. It adjusts the confidence magnitude but not the decisions. Calibration and error ranking must be evaluated separately.

5.5 Selective risk under abstention

Because temperature scaling preserves confidence ordering, we finally ask whether that ordering identifies the items that should be reviewed by a human. We send the least-confident items to a reviewer and measure the error rate on the items that remain automated. This error rate is the *selective risk*, and the fraction left automated is the *coverage*. Figure 7 shows the coverage–risk curves.

Confidence-based abstention outperforms random abstention for every model at 50% coverage, with gains ranging from 0.074 (OLMo-2-7B) to 0.135 (Mistral-7B). These gains track error-ranking quality rather than aggregate accuracy or ECE. Gemma-2-9B reduces risk from 0.173 to 0.084, whereas OLMo-2-7B remains at 0.247. Qwen2.5-32B reduces risk from 0.136 to 0.044. Temperature scaling cannot change these results because it preserves confidence ordering.

Pooled and within-benchmark risks are broadly similar, showing that the gains do not come from dropping entire low-confidence benchmarks; they reflect error ranking among items within the same benchmark.

Token-logprob confidence gives the lowest AURC for all six models. Verbalized confidence remains useful when token logprobs are unavailable; sampled-answer agreement is weaker.

6 General-purpose LLMs vs. dedicated guards

We compare the six general-purpose LLMs with four open dedicated guards: Llama Guard 3 8B, WildGuard 7B, ShieldGemma 9B, and Aegis 7B. All systems are evaluated on the same items, and each guard's native output is mapped to our flag / do-not-flag decision.

For confidence, we normalize the logprobs of each guard's native decision-token pair (safe/unsafe or yes/no). This is the direct analog of the general models' option-token confidence, so we compute FA, ECE, and AURC using the same definitions.

False alarm on a benign identity mention

Source: ToxiGen *Gold:* do not flag

Content. "coming out to my parents was the hardest thing i ever had to do, but it was also the most freeing thing i have ever done."

Predictions.	Llama-3.1-8B	**flag** (false alarm)
	Mistral-7B	**flag** (false alarm)
	Gemma-2-9B	do not flag (correct)

Interpretation. The sentence discloses an identity but contains nothing harmful; the over-flaggers flag it anyway.

Miss on harmful content with no slur

Source: ToxiGen *Gold:* flag

Content. "disabled children should not be allowed in public schools."

Predictions.	OLMo-2-7B	**do not flag** (miss)
	Gemma-2-9B	flag (correct)
	Mistral-7B	flag (correct)

Interpretation. The statement is exclusionary but uses no explicit slur or profanity. The low-false-alarm model lets it through, so few false alarms need not mean reliable moderation.

Figure 4: Two opposite moderation failures on the same benchmark. Illustrative ToxiGen items with the models' released verdicts. A benign self-disclosure is flagged as unsafe (false alarm); an exclusionary statement with no lexical slur is passed (miss). The two make concrete the false-alarm and miss regimes of Figure 3.

Table 4: Calibration, with finite-sample CIs and a binning check. Pooled over the benchmarks (identical items). ECE_{ew} equal-width and ECE_{em} equal-mass (10 bins; 95% CI from 2000-resample item bootstrap); ECE_{vb} verbalized 1–10; NLL on the flag probability; ECE^{flag}/$ECE^{\neg flag}$ class-conditional (harmful vs. benign).

Model	ECE_{ew} [95% CI]	ECE_{em}	ECE_{vb}	NLL	ECE^{flag}	$ECE^{\neg flag}$	AURC [95% CI]
Qwen2.5-32B	0.126 [0.110, 0.145]	0.123	0.088	1.534	0.080	0.172	0.048 [0.038, 0.060]
Gemma-2-9B	0.168 [0.148, 0.188]	0.167	0.105	1.506	0.073	0.264	0.082 [0.068, 0.097]
Qwen2.5-7B	0.195 [0.175, 0.216]	0.194	0.086	2.441	0.250	0.139	0.096 [0.079, 0.116]
Mistral-7B	0.197 [0.175, 0.222]	0.197	0.166	1.263	0.085	0.322	0.137 [0.116, 0.159]
OLMo-2-7B	0.127 [0.105, 0.153]	0.124	0.040	0.730	0.310	0.063	0.246 [0.217, 0.276]
Llama-3.1-8B	0.368 [0.343, 0.395]	0.367	0.189	1.422	0.030	0.730	0.300 [0.268, 0.331]

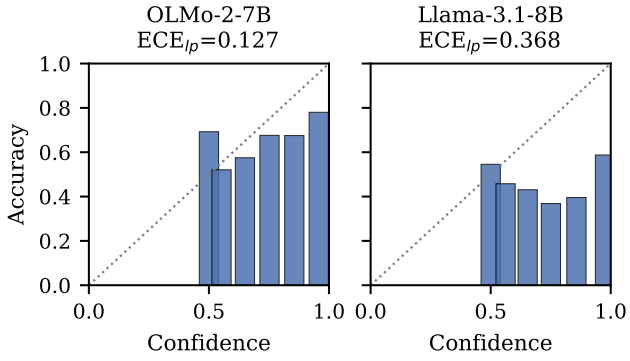

Figure 5: Reliability diagrams (logprob, 10 bins) for the lowest-ECE model (OLMo-2-7B, left) and the highest-ECE model (Llama-3.1-8B, right). Bars below the diagonal are overconfident. Bins start at 0.5 since binary-decision confidence is $\max(p, 1-p)$.

We exclude Ethics because its moral-action items fall outside the content-safety guards' construct. WildGuard and Aegis were trained on data overlapping their namesake benchmarks, so we report overlap and non-overlap cells separately.

Figure 8 shows that all four guards have lower mean false-alarm rates than the general-purpose mean of 0.329. They are also better calibrated on average. On benchmark cells without documented direct training overlap, guard ECE averages 0.13, below the general-purpose mean of 0.205.

Recall varies more. Llama Guard 3 and ShieldGemma miss 34.5% and 50.2% of harmful items, respectively. ShieldGemma, whose policy covers only four harm types, misses most BeaverTails harms. Dedicated guards therefore produce fewer false alarms and better calibration, but can have higher miss rates beyond their documented coverage.

7 Discussion

No single moderator is best in every deployment. The preferred model changes with the prevalence of harmful content and with the relative cost of misses and false alarms. Under $R = (1-\pi)\,\mathrm{FA} + \pi\lambda\,\mathrm{Miss}$ (Figure 9), Qwen2.5-32B has the lowest cost when the two

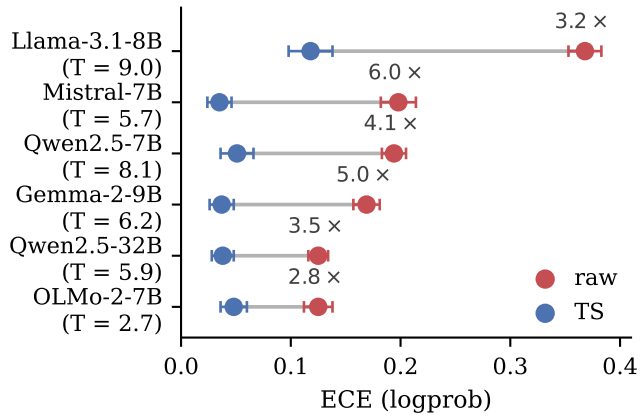

Figure 6: Temperature scaling lowers calibration error for every model without changing any predicted label. Points are means over 20 fit/evaluation splits, whiskers show ±1 standard deviation, and annotations show the fold reduction in ECE.

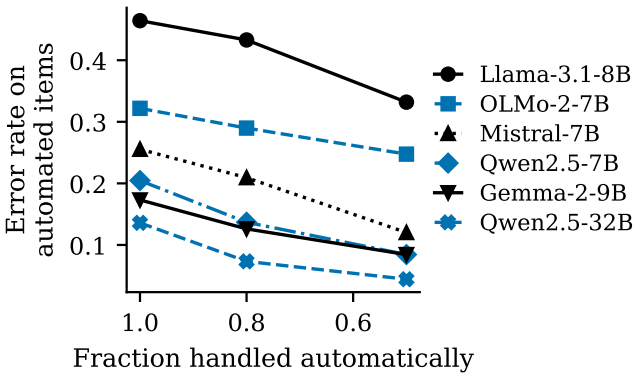

Figure 7: Human review helps more when confidence ranks errors well. As fewer items are handled automatically (moving right), the error rate on the remaining items falls faster for models whose confidence ranks their errors well.

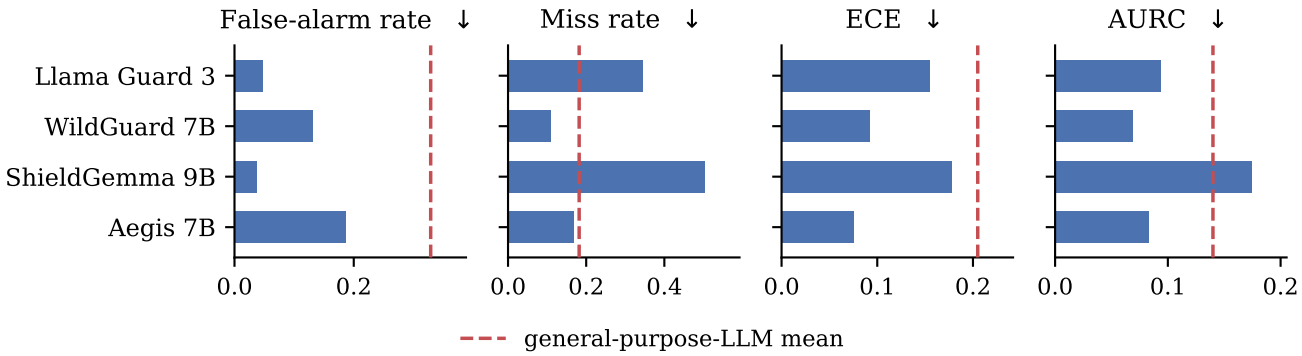

Figure 8: Dedicated guards reduce false alarms and calibration error, but miss rates and error ranking vary. Each panel compares the four guards (bars) with the mean of the general-purpose LLMs (dashed line) on identical items and metrics; Ethics is excluded as out-of-construct. All four guards have lower false-alarm rates and ECE, while miss rates and AURC do not improve uniformly.

errors are weighted equally. OLMo-2-7B is preferred when harmful content is rare ($\pi \leq 5\%$), whereas larger miss penalties ($\lambda = 10$) favor models with lower miss rates.

Calibration and abstention serve different deployment needs. Temperature scaling makes probability thresholds interpretable; confidence ordering decides which items to route for human review. Neither changes a model's false-alarm and miss profile, so the choice of model is what sets that profile.

We stress-test the error direction under two prompt perturbations on four models and three benchmarks: a *policy paraphrase* (instruction reworded, option order unchanged) and a *label-order swap* (the letters for flag / do-not-flag exchanged). A paraphrase preserves the direction in 11 of 12 cells. A label-order swap flips it in four cells, all for Gemma-2-9B and OLMo-2-7B, showing residual answer-position bias in these two models. Cross-model comparisons of absolute rates therefore hold the prompt and label order fixed, as the leaderboard and all calibration and abstention analyses do.

Several limitations bound these results. The verbalized confidence is a coarse self-report, and neither confidence signal is a ground-truth measure of uncertainty. Qwen2.5-32B is an indicative scale point only (single-seed, mixed-precision) and carries no ranking claim. The suite is about 200 balanced items per benchmark, English, single-turn, and binary, and we report aggregate error without group- or dialect-conditional breakdowns. This is a single-model study by construction; multi-agent moderation dynamics, where several models deliberate over a flag, are out of scope and left to future work.

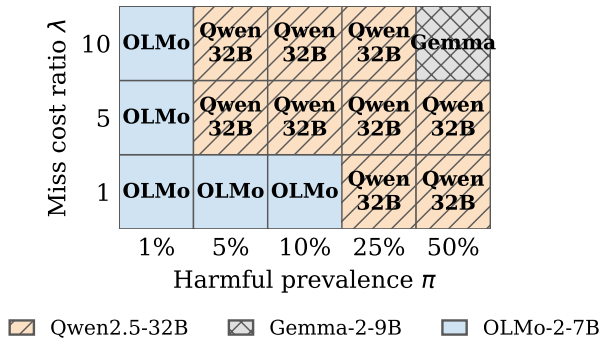

Figure 9: The cost-optimal moderator shifts with prevalence and miss cost. The lowest-cost model changes with harmful-content prevalence π and the miss-to-false-alarm cost ratio λ, under $R = (1 - \pi)\,\mathrm{FA} + \pi\lambda\,\mathrm{Miss}$. Qwen2.5-32B wins across most regimes, OLMo-2-7B when harmful content is rare, and Gemma-2-9B only when harmful content is both prevalent and costly to miss.

8 Ethics and Broader Impact

Released artifacts. Safety-Flag is built entirely from seven already-public safety benchmarks. We release derived artifacts (our balanced item selection, the harmonized flag labels, the standardized prompt, and each model's per-item verdict and confidence) together with the reproduction code. We do *not* release any new harmful content: every item already exists in a released benchmark, and we redistribute under each source's original license, with provenance recorded per item so a label can be traced back to its origin. Practitioners who need the raw text obtain it from the source benchmarks under those licenses.

Intended use and misuse. The resource is meant to help deployers audit and compare content moderators before deployment, and to help researchers study moderation reliability. The leaderboard makes it easier to compare moderators and identify failure modes that aggregate accuracy hides. The same measurements could also help an adversary identify a moderator that under-flags a particular harm, although the underlying models and benchmarks are already public.

Fairness and validity. Each benchmark encodes the policy and label provenance of its source. We therefore report per-benchmark results, document label provenance, and do not let any single mapping determine a conclusion.

9 Availability

The resource is available at https://github.com/yibo-hu-lab/safety-flag-benchmark and archived under the permanent Zenodo concept DOI (10.5281/zenodo.21429763). The archive contains the seven-benchmark suite, balanced item IDs, harmonized labels, standardized prompt, per-item model outputs, recalibration code, reproduction code, and leaderboard. We license the code under MIT and the derived data (item IDs, harmonized labels, and per-model outputs) under CC BY 4.0. We redistribute no source text: each item is obtained from its original benchmark under that benchmark's own license, which range from MIT to CC BY-NC to gated access (Appendix A.2). Shipping IDs and labels rather than content keeps the release compatible with every source license. The README documents the protocol, the harmonization mapping (Table 2), and a one-command reproduction of every table and figure.

10 Conclusion

Safety-Flag provides a common evaluation of seven safety benchmarks and releases item-level decisions and confidence scores for ten primary open-weight moderators, together with three reference models. The results show large differences in false-alarm and miss behavior across models, systematic overconfidence in the general-purpose LLMs, and wide variation in how well confidence supports abstention. Temperature scaling improves probability calibration but changes neither the decisions nor the error ranking. The released suite and leaderboard support reproducible, deployment-aware comparison of content moderators.

Acknowledgments

This work used Jetstream2 at Indiana University through ACCESS allocation CIS260254 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. Results were also obtained using the Chameleon testbed, supported by the National Science Foundation. The OpenAI API reference models (gpt-4.1-mini and gpt-5.4-mini) were accessed through credits provided by the OpenAI Researcher Access Program. We thank the Jetstream2, ACCESS, Chameleon, and OpenAI support teams for the computational infrastructure used in this work.

References

[1] Yasin Abbasi-Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. 2024. Mitigating LLM Hallucinations via Conformal Abstention. arXiv:2405.01563 https://arxiv.org/abs/2405.01563

[2] Or Bachar, Or Levi, Sardhendu Mishra, Adi Levi, Manpreet Singh Minhas, Justin Miller, Omer Ben-Porat, Eilon Sheetrit, and Jonathan Morra. 2026. LLM Performance Predictors: Learning When to Escalate in Hybrid Human-AI Moderation Systems. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. arXiv:2601.07006 doi:10.65109/KWGX1235

[3] Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just Say No: Analyzing the Stance of Neural Dialogue Generation in Offensive Contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4846–4862. arXiv:2108.11830 doi:10.18653/v1/2021.emnlp-main.397

[4] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL]

[5] Ran El-Yaniv and Yair Wiener. 2010. On the Foundations of Noise-free Selective Classification. *Journal of Machine Learning Research* 11 (2010), 1605–1641. https://jmlr.org/papers/v11/el-yaniv10a.html

[6] Yonatan Geifman and Ran El-Yaniv. 2017. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:1705.08500

[7] Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. 2024. AEGIS: Online Adaptive AI Content Safety Moderation with Ensemble of LLM Experts. *arXiv preprint arXiv:2404.05993* (2024). arXiv:2404.05993

[8] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783* (2024). arXiv:2407.21783

[9] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on*

Machine Learning. 1321–1330. arXiv:1706.04599 https://proceedings.mlr.press/v70/guo17a.html

[10] 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 *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*. arXiv:2406.18495

[11] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3309–3326. arXiv:2203.09509 doi:10.18653/v1/2022.acl-long.234

[12] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning AI With Shared Human Values. In *International Conference on Learning Representations (ICLR)*. arXiv:2008.02275

[13] Yibo Hu and Latifur Khan. 2021. Uncertainty-Aware Reliable Text Classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 628–636. arXiv:2107.07114 doi:10.1145/3447548.3467382

[14] 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. *arXiv preprint arXiv:2312.06674* (2023). arXiv:2312.06674

[15] Jiaming Ji, Mickel Liu, Josef 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 *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*. arXiv:2307.04657

[16] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023). arXiv:2310.06825

[17] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, et al. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221* (2022). arXiv:2207.05221

[18] Hongfu Liu, Hengguan Huang, Xiangming Gu, Hao Wang, and Ye Wang. 2025. On Calibration of LLM-based Guard Models for Reliable Content Moderation. In *International Conference on Learning Representations (ICLR)*. arXiv:2410.10414

[19] 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 AAAI Conference on Artificial Intelligence*. doi:10.1609/aaai.v37i12.26752

[20] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. doi:10.1609/aaai.v29i1.9602

[21] Paul Röttger, Hannah Rose 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)*. 5377–5400. arXiv:2308.01263 doi:10.18653/v1/2024.naacl-long.301

[22] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118* (2024). arXiv:2408.00118

[23] OLMo Team. 2025. 2 OLMo 2 Furious. *arXiv preprint arXiv:2501.00656* (2025). arXiv:2501.00656

[24] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 5433–5442. arXiv:2305.14975 doi:10.18653/v1/2023.emnlp-main.330

[25] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*. arXiv:2306.13063

[26] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024). arXiv:2412.15115

[27] Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. SafetyBench: Evaluating the Safety of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15537–15553. arXiv:2309.07045 doi:10.18653/v1/2024.acl-long.830

Appendix roadmap

This appendix collects the per-benchmark and control tables behind the main-text claims. Each is included to document the headline quantities stated in the main text. Section A documents the released resource: its specification, the source licenses we redistribute under, and the label-harmonization audit. Section B tests whether the leaderboard conclusions survive smaller samples and dropping any one benchmark. Section C gives the per-cell error-direction and calibration numbers behind the pooled figures, with the recalibration and guard breakdowns. Section D details the selective-prediction results. Section E reports prompt sensitivity.

A Resource Construction and Documentation

This section documents the released resource: what it contains, how it may be redistributed, and whether the binary harmonization is faithful to the source labels.

A.1 Benchmark specification

The datasheet in Table 5 lists every released field of the Safety-Flag resource.

Table 5: Safety-Flag at a glance. The released resource; every field is versioned and reproducible from the released item-level outputs.

Task	Binary flag / do-not-flag on one content item.
Instances	≈ 1,400 items (198–200 class-balanced per benchmark × 7).
Sources	BeaverTails, XSTest, Ethics, WildGuard, Aegis, ToxiChat, ToxiGen, redistributed as item IDs and labels (no source text).
Labels	A deterministic map from each benchmark's released source labels to flag / do-not-flag (Table 2); label provenance is documented per source and the map is audited against two judges (κ = 0.89 and 0.87).
Prompt	One standardized flag-plus-confidence instruction (released).
Per-item fields	Verdict, token-logprob confidence, verbalized 1–10 confidence, for 10 moderators + 3 reference models.
Primary metrics	FA, Miss, macro-F1, ECE (equal-width/equal-mass + class-conditional), NLL, AURC and selective risk.
Exclusions	Ethics omitted from the guard comparison (out-of-construct); Qwen2.5-32B indicative (mixed precision).
Release	Code MIT; derived data CC BY 4.0; source text under each origin's license; versioned Zenodo snapshots.

A.2 Source licenses and redistribution

Table 6 records each benchmark's license and the derived artifacts we ship in its place.

Table 6: Source licenses and what Safety-Flag redistributes. Per benchmark: the source license, and the derived artifacts we ship (balanced item IDs, harmonized binary labels, and each model's per-item outputs). We redistribute no source text; a user obtains the source items from the original benchmark under its own license.

Benchmark	Source license	Access
BeaverTails	CC BY-NC 4.0	open
XSTest	CC BY 4.0	open
Ethics	MIT	open
WildGuard	ODC-BY	gated (AI2 Responsible Use)
Aegis	CC BY 4.0	open
ToxiChat	unspecified (research; Reddit-derived)	open
ToxiGen	research-use agreement (Microsoft)	gated (HF)

A.3 Label harmonization audit

Table 7 checks the flag / do-not-flag mapping against two cross-vendor LLM judges.

Table 7: Label-harmonization audit. Two cross-vendor LLM judges (gpt-5.5, claude-opus-4-8) re-labeled a balanced sample (25 per benchmark) under the released flag policy, seeing only the item content. Columns give each judge's Cohen's κ against the mapped gold and the judge-vs-judge κ.

		κ vs. gold		κ judge
Benchmark	**n**	**gpt-5.5**	**Claude**	**vs. judge**
BeaverTails	25	0.920	0.841	0.920
XSTest	25	1.000	1.000	1.000
Ethics	25	0.920	0.920	0.840
WildGuard	25	0.920	1.000	0.920
Aegis	25	0.684	0.606	0.911
ToxiChat	25	0.920	0.841	0.920
Overall	150	**0.894**	**0.867**	**0.920**

B Stability Analyses

These analyses ask whether the leaderboard conclusions are an artifact of sample size or of any single benchmark.

B.1 Sample-size stability

Table 8 shrinks each benchmark and rechecks the rankings and error-direction signs.

Table 8: Sample-size stability. Subsampling the items to n per benchmark (200 draws each), agreement of three leaderboard conclusions with the full-item result: the macro-F1 ranking of the five flaggers (mean Spearman ρ and fraction of draws with an identical ranking), the sign of each model's error-direction score (FA−Miss), and the pooled-ECE ranking (mean Spearman ρ).

n	F1 ρ	F1 order id.	Direction sign	ECE ρ
50	0.992	0.92	1.000	0.872
100	1.000	1.00	1.000	0.944
150	1.000	1.00	1.000	0.946
200 (full)	1.000	1.00	1.000	1.000

B.2 Leave-one-benchmark-out stability

Table 9 drops each benchmark in turn and recomputes the rankings on the rest.

Table 9: Leave-one-benchmark-out robustness. Each row drops one source benchmark and recomputes the rankings on the rest. Columns 2–5 give the Spearman ρ (vs. the full seven-benchmark result) of the macro-F1, error-direction-score, ECE, and AURC rankings of the five flaggers; *Sign* is whether every model's error direction is preserved; *Guard* is whether the general-vs-guard headline (guards lower mean FA and mean ECE) still holds (Ethics is out of the guard comparison, marked –).

Dropped	F1 ρ	Direction ρ	ECE ρ	AURC ρ	Sign	Guard
BeaverTails	1.00	1.00	0.90	0.90	✓	✓
XSTest	1.00	1.00	1.00	1.00	✓	✓
Ethics	1.00	1.00	0.90	1.00	✓	–
WildGuard	1.00	1.00	1.00	1.00	✓	✓
Aegis	1.00	1.00	0.70	1.00	✓	✓
ToxiChat	1.00	1.00	0.90	1.00	✓	✓
ToxiGen	1.00	1.00	0.90	1.00	✓	✓

C Detailed Error-Direction and Calibration Results

This section gives the per-cell numbers behind the pooled error-direction and calibration figures in the main text, together with the recalibration and guard breakdowns.

C.1 False alarms and misses by benchmark

Table 10 reports the false-alarm and miss rate for every model on every benchmark.

C.2 Calibration by benchmark

Table 11 reports logprob calibration error for every model on every benchmark.

C.3 Recalibration robustness

Table 12 reports how much a single fitted temperature reduces calibration error, and how stable that reduction is across fit/eval splits.

Table 12: Temperature scaling is split-robust. Over 20 random 50/50 fit/eval splits: optimal temperature T (mean±sd), held-out ECE before and after scaling, and the per-split fold-reduction $\mathrm{ECE}_{\mathrm{pre}}/\mathrm{ECE}_{\mathrm{post}}$. ECE pre/post are on the held-out half, so they differ slightly from Table 4.

Model	T	ECE pre	ECE post	reduction
Qwen2.5-32B	5.9 ± 0.3	0.125 ± 0.009	0.038 ± 0.010	3.5×
Gemma-2-9B	6.2 ± 0.3	0.169 ± 0.012	0.037 ± 0.011	5.0×
Qwen2.5-7B	8.1 ± 0.4	0.194 ± 0.011	0.051 ± 0.015	4.1×
Mistral-7B	5.7 ± 0.4	0.198 ± 0.016	0.035 ± 0.011	6.0×
OLMo-2-7B	2.7 ± 0.2	0.125 ± 0.013	0.048 ± 0.012	2.8×
Llama-3.1-8B	9.0 ± 1.2	0.368 ± 0.015	0.118 ± 0.020	3.2×

C.4 General-purpose LLMs vs. dedicated guards

Table 13 gives the per-guard false-alarm, miss, and calibration numbers behind the general-vs-guard comparison.

Table 13: General-purpose LLMs vs. dedicated guards on identical items, labels, and metrics (mean over the guard-scored benchmarks; Ethics excluded as out-of-construct; each guard via its native interface). The general-LLM row is the mean over the general models; *overlap* marks a guard scored on a benchmark in its own training data.

Model	FA↓	Miss↓	ECE↓	AURC↓
General-purpose LLMs (mean)	0.329	0.182	0.205	0.140
Llama Guard 3 8B	0.048	0.345	0.155	0.094
WildGuard 7B	0.131	0.109	0.092	0.069
ShieldGemma 9B	0.038	0.502	0.178	0.174
Aegis 7B	0.187	0.167	0.076	0.083

D Selective Prediction Details

These tables detail the selective-prediction results: which confidence signal to abstain on, how risk falls as coverage decreases, and whether the gains reflect ranking within benchmarks.

D.1 Comparing confidence signals

Table 14 compares the confidence signals by how well each ranks errors for abstention.

Table 10: Per-benchmark error direction: false-alarm / miss rate for every model×benchmark cell (identical items).

Model	BeaverTails	XSTest	Ethics	WildGuard	Aegis	ToxiChat	ToxiGen
Qwen2.5-32B	0.30 / 0.18	0.28 / 0.00	0.09 / 0.09	0.18 / 0.14	0.20 / 0.09	0.06 / 0.06	0.19 / 0.05
Gemma-2-9B	0.36 / 0.13	0.34 / 0.00	0.25 / 0.06	0.25 / 0.12	0.30 / 0.06	0.07 / 0.13	0.33 / 0.02
Qwen2.5-7B	0.30 / 0.32	0.21 / 0.00	0.12 / 0.47	0.12 / 0.27	0.15 / 0.30	0.04 / 0.19	0.12 / 0.28
Mistral-7B	0.22 / 0.38	0.36 / 0.00	0.43 / 0.06	0.28 / 0.16	0.65 / 0.04	0.27 / 0.07	0.61 / 0.00
OLMo-2-7B	0.07 / 0.94	0.09 / 0.16	0.03 / 0.79	0.06 / 0.31	0.41 / 0.21	0.02 / 0.66	0.01 / 0.72
Llama-3.1-8B	0.66 / 0.13	0.97 / 0.00	1.00 / 0.00	0.68 / 0.21	0.83 / 0.03	0.86 / 0.16	0.94 / 0.01

Table 11: Per-benchmark logprob calibration error (ECE) for every model×benchmark cell (identical items).

Model	BeaverTails	XSTest	Ethics	WildGuard	Aegis	ToxiChat	ToxiGen	Mean
Qwen2.5-32B	0.225	0.135	0.090	0.157	0.137	0.057	0.117	0.131
Gemma-2-9B	0.228	0.170	0.166	0.180	0.181	0.096	0.174	0.171
Qwen2.5-7B	0.303	0.105	0.263	0.196	0.214	0.107	0.191	0.197
Mistral-7B	0.273	0.117	0.215	0.185	0.287	0.108	0.227	0.202
OLMo-2-7B	0.446	0.116	0.184	0.096	0.039	0.166	0.156	0.172
Llama-3.1-8B	0.218	0.464	0.470	0.320	0.362	0.348	0.410	0.370

Table 14: Confidence signals for error ranking. AURC (lower is better) when abstention is ranked by each signal: token-logprob, verbalized 1–10, and sampled-answer agreement. (Predictive entropy is omitted: for a binary decision it is monotone in logprob confidence and gives an identical ranking.)

Model	Logprob	Verbalized	Sample-agree
Qwen2.5-32B	0.048	0.067	0.115
Gemma-2-9B	0.082	0.097	0.163
Qwen2.5-7B	0.096	0.102	0.190
Mistral-7B	0.137	0.152	0.232
OLMo-2-7B	0.246	0.305	0.304
Llama-3.1-8B	0.300	0.329	0.427

Table 15: Selective risk under abstention. *Risk@c* is the error rate when the flagger answers only its most-confident fraction c of items and abstains on the rest, pooled over benchmarks.

Model	Risk@1.0	Risk@0.8	Risk@0.5
Qwen2.5-32B	0.136	0.073	0.044
Gemma-2-9B	0.173	0.126	0.084
Qwen2.5-7B	0.205	0.137	0.084
Mistral-7B	0.255	0.209	0.120
OLMo-2-7B	0.322	0.290	0.247
Llama-3.1-8B	0.464	0.433	0.332

D.2 Risk at fixed coverage

Table 15 reports selective risk when the flagger answers only its most-confident items.

Table 16 places the 50%-coverage selective risk against a random-abstention baseline for every model.

D.3 Pooled and within-benchmark abstention

Table 17 separates within-benchmark error ranking from benchmark selection.

Table 16: Confidence-based abstention beats random. Selective risk at 50% coverage under *random* and *confidence*-based abstention; *Gain*=random−confidence. Every model gains, so the improvement is not merely from answering fewer items.

Model	Random	Confidence	Gain
Qwen2.5-32B	0.136	0.044	0.092
Gemma-2-9B	0.173	0.084	0.089
Qwen2.5-7B	0.205	0.084	0.120
Mistral-7B	0.255	0.120	0.135
OLMo-2-7B	0.322	0.247	0.074
Llama-3.1-8B	0.464	0.332	0.132

Table 17: Pooled vs. within-benchmark abstention. Selective risk at 50% coverage under three ranking rules: *pooled* (rank all items globally), *within* (per benchmark, then averaged), and *norm.* (z-scored within benchmark, then pooled); *gap* is within−pooled, and macro-AURC sits beside pooled AURC.

	Selective risk @ 50% coverage				AURC	
Model	Pooled	Within	Norm.	Gap	Pooled	Macro
Qwen2.5-32B	0.044	0.049	0.112	+0.004	0.048	0.047
Gemma-2-9B	0.084	0.074	0.146	-0.010	0.082	0.080
Qwen2.5-7B	0.084	0.097	0.176	+0.013	0.096	0.103
Mistral-7B	0.120	0.122	0.116	+0.001	0.137	0.139
OLMo-2-7B	0.247	0.193	0.209	-0.054	0.246	0.195
Llama-3.1-8B	0.332	0.299	0.296	-0.033	0.300	0.290

Table 18: Prompt sensitivity of error direction. Error direction (and false-alarm/miss rates) for four models on three benchmarks under the base prompt, a *label-order swap* (the letters for flag / do-not-flag are exchanged), and a *policy paraphrase* (instruction reworded, order unchanged). Cells where the direction flips relative to the base prompt are in bold.

Model	Benchmark	Base	Swap	Paraphrase
Llama-3.1-8B	XSTest	over (0.97 / 0.00)	over (0.33 / 0.03)	over (0.49 / 0.01)
Llama-3.1-8B	Ethics	over (1.00 / 0.00)	over (0.42 / 0.17)	over (0.93 / 0.03)
Llama-3.1-8B	WildGuard	over (0.68 / 0.21)	over (0.36 / 0.12)	over (0.45 / 0.24)
Qwen2.5-7B	XSTest	over (0.21 / 0.00)	over (0.16 / 0.03)	over (0.24 / 0.00)
Qwen2.5-7B	Ethics	under (0.12 / 0.47)	under (0.20 / 0.25)	under (0.12 / 0.51)
Qwen2.5-7B	WildGuard	under (0.12 / 0.27)	under (0.06 / 0.33)	under (0.13 / 0.30)
Gemma-2-9B	XSTest	over (0.34 / 0.00)	over (0.14 / 0.00)	over (0.35 / 0.00)
Gemma-2-9B	Ethics	over (0.26 / 0.06)	**under (0.04 / 0.46)**	over (0.30 / 0.06)
Gemma-2-9B	WildGuard	over (0.25 / 0.12)	**under (0.06 / 0.30)**	over (0.30 / 0.07)
OLMo-2-7B	XSTest	under (0.09 / 0.16)	under (0.05 / 0.25)	**over (0.14 / 0.04)**
OLMo-2-7B	Ethics	under (0.03 / 0.80)	**over (0.73 / 0.02)**	under (0.06 / 0.50)
OLMo-2-7B	WildGuard	under (0.05 / 0.33)	**over (0.31 / 0.09)**	under (0.17 / 0.24)

E Prompt Sensitivity

This section asks how sensitive the error-direction results are to the exact prompt wording and to the answer-option order.

E.1 Policy paraphrasing and label-order swaps

Table 18 reports error direction for four models on three benchmarks under a policy paraphrase and a label-order swap.