# PHAROS: FAST FINE-GRAINED CLASSIFICATION OF SAFETY RISKS IN USER PROMPTS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Safety filters often report whether a user prompt is harmful, but many applications also need to know which risk it presents. We study this finer-grained task under the 16 risk categories in the AIR 2024 Level-2 taxonomy. Risk Atlas unifies 23 public safety datasets under that taxonomy, preserves source provenance, and adds category-checked synthetic prompts to broaden coverage. Splits are stratified by source and risk category, with duplicate queries removed and newly synthesized prompts confined to training. On 10,000 prompts with independently mapped AIR labels, the labeling model reaches 0.97 macro-F1. Pharos, a compact encoder trained on Risk Atlas, reaches 0.753 AIR Level-2 macro-F1 on the held-out test set, ahead of evaluated released guards and a decoder trained on the same corpus. On an RTX 4080-SUPER, Pharos's median batch-one inference time is 5.36 ms per query. Together, Risk Atlas and Pharos provide a shared evaluation setting and a fast baseline for fine-grained prompt risk categorization. Code: https://anonymous.4open.science/r/Pharos-70D8/.

## 1 INTRODUCTION

An input safety filter usually answers a binary question: should this prompt be flagged? That answer is insufficient when different requests require different handling. A request concerning self-harm, for example, calls for a different response from a request concerning privacy. A classifier that names the relevant risk can supply information to a later policy decision, enabling fine-grained decision such as allowing, refusing, escalating, or redirecting the request.

Existing guards expose different label sets. Some produce a binary verdict, some use their own fixed categories, and some accept category descriptions at inference time (Inan et al., 2023; Han et al., 2024; Zhao et al., 2025; Zaratiana et al., 2026). This diversity makes fine-grained results difficult to compare. A guard's native labels can leave some categories unreachable after mapping to a shared taxonomy; a shared-task comparison then needs to report which categories each model can express. Compact schema-conditioned encoders already provide efficient moderation (Zaratiana et al., 2026; Minko et al., 2026). When the taxonomy is fixed, how well can a small encoder learn its categories directly?

We answer that question for English user prompts and the AIR 2024 Level-2 risk taxonomy (Zeng et al., 2024b). AIR provides 16 risk categories derived from government and company policy documents. We add a non-harmful class and ask a model to select one primary label for each prompt.

Risk Atlas unifies 23 public datasets under AIR with traceable source and label mappings, and its model-derived labels are audited where independent mappings exist. The sources span adversarial prompts, over-refusal probes, and real chatbot traffic; retaining their provenance makes source-specific errors inspectable. Several AIR categories remain sparse after aggregation, motivating category-checked synthetic augmentation. Evaluation stratifies by source and AIR category for broad coverage, keeps queries disjoint across splits, and reserves the newly synthesized prompts for training.

Pharos reaches 0.753 Level-2 macro-F1 on the held-out set and processes prompts in milliseconds on one consumer GPU. Matched-prompt comparisons with released guards and a decoder trained on the same corpus show its AIR-specific performance while keeping differences in native tasks and training data explicit.



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

## 2 Related Work

**Guard models.** Llama Guard established an open generative interface that places a safety policy in the prompt and decodes a verdict (Inan et al., 2023). Later systems add joint prompt, response, and refusal judgments (Han et al., 2024), synthetic- data scaling (Zeng et al., 2024a), multilingual coverage (Kumar et al., 2025; Upadhayay & Behzadan, 2025), and severity and streaming moderation (Zhao et al., 2025). These systems address broader moderation tasks than single-prompt category prediction. Their evaluations also commonly emphasize binary harmfulness or independent per-category decisions, which do not directly measure selection among mutually exclusive fine-grained risks. Aegis 2.0 and RobloxGuard already train with fine-grained taxonomies and substantial synthetic data (Ghosh et al., 2025b; Nandwana et al., 2025). Our corpus instead fixes the AIR Level-2 label space across heterogeneous sources and tests one primary prompt label on a shared set. Qwen3Guard's severity tiers and FlexGuard's strictness-dependent labels address a separate question: whether a category should trigger intervention under a particular policy (Zhao et al., 2025; **?**).

**Lightweight and typed classifiers.** Schema-conditioned encoders accept label or task descriptions at inference time (Zaratiana et al., 2026; Minko et al., 2026); other lightweight approaches classify latent prototypes or distill reasoning-based monitors (Chrabaszcz et al., 2025; Diwan et al., 2026). Jev frames classification and routing as typed, non-generative decisions over caller-declared alternatives (TypeSafe AI, 2026). Its proprietary model, objective, and training data are not public. Openjev reconstructs the interface with a Qwen3.5-4B NLI cross-encoder that scores one hypothesis per option (Wortega, 2026). We evaluate this public artifact rather than TypeSafe AI's hosted Jev system.

**Taxonomies and safety data.** Safety label spaces may be derived from policy documents (Zeng et al., 2024b; Ghosh et al., 2025a), designed for benchmarks (Li et al., 2024b; Mazeika et al., 2024), or defined internally by a guard (Inan et al., 2023; Han et al., 2024; Ghosh et al., 2024). Bias-focused suites probe a partly orthogonal axis (Parrish et al., 2022; Nangia et al., 2020), while meta-taxonomies and free-text decomposition seek more portable representations (Slattery et al., 2024; Weidinger et al., 2022; Li et al., 2024a). Comparing systems with different categorical anchors therefore requires a potentially lossy mapping. Public corpora likewise cover toxicity, unsafe instructions, and jailbreaks under incompatible annotation schemes (Ji et al., 2023; Lin et al., 2023; Hartvigsen et al., 2022; Wang et al., 2024). Aggregated corpora inherit this fragmentation (Dang et al., 2026); synthetic data can separately broaden attack tactics or sparse taxonomy regions (Jiang et al., 2024; Kumar et al., 2024). The corpus's contribution is the documented AIR crosswalk and its shared evaluation task. Some source datasets contain language-model-generated prompts; our held-out sets contain no prompts from Pharos's augmentation pipeline. The source-by-category split aims to retain diverse datasets and risk classes in both validation and test.

**Adversarial robustness and over-refusal.** Safety filters face optimized suffixes, query-only search, persuasion, many-shot prompting, and cross-lingual transfer (Zou et al., 2023; Chao et al., 2023; Mehrotra et al., 2023; Zeng et al., 2024c; Anil et al., 2024; Deng et al., 2023). Defences include input filtering, randomized smoothing, and activation-space interventions (Alon & Kamfonas, 2023; Jain et al., 2023; Robey et al., 2023; Arditi et al., 2024). Over-refusal benchmarks test benign prompts rejected because their surface form resembles unsafe content (Röttger et al., 2023; Cui et al., 2024; Fu & Zou, 2025).

## 3 Methodology

### 3.1 Task definition

The task maps a user prompt to one primary AIR Level-2 label: one of 16 risks or `non_harmful`. The target is the risk implicated by the assistance requested, rather than a prediction about a model's response. For prompts spanning several risks, the rubric selects the principal requested action; benign calibration probes remain non-harmful even when their wording resembles unsafe content. This creates a consistent, single-label task without treating the label as a severity score or policy

decision. The 17-way prediction can also be collapsed to AIR Level-1 or binary harmfulness for comparison; all three views come from the same output.

### 3.2 Risk Atlas: A source-aligned safety corpus

Risk Atlas aligns prompts from 23 public sources to one AIR label space while retaining their origin. The sources cover adversarial requests, toxicity and bias, policy-grounded harms, and benign over-refusal probes; their examples range from real chatbot traffic and human-authored prompts to optimized or language-model-generated inputs (Zou et al., 2023; Huang et al., 2023; Jiang et al., 2024; Cui et al., 2024; Lin et al., 2023). Retaining source identity enables error analysis by provenance instead of pooling examples collected and labeled under different conditions.

Label assignment proceeds in two stages. A source's harm status is retained when its construction determines it, including benign over-refusal probes whose wording may look unsafe (Cui et al., 2024; Röttger et al., 2023). An AIR Level-2 category is mapped directly only when the native label has an unambiguous counterpart; otherwise a judge applies the shared rubric to the requested action. Recording the source and assignment route separates native-category evidence from model-derived labels and makes the latter auditable.

### 3.3 Synthetic augmentation

The source-dataset union remains sparse in several AIR Level-2 categories, and its attack framings are limited to those captured by the source collections. For each target category, we combine an AIR anchor with companion examples sampled across sources and condition an uncensored open-weight generator on that seed set. The implementation takes an AIR-Bench example as the anchor when available and draws the remaining examples from different source groups within the same category. The AIR anchor grounds the requested category, while cross-source examples broaden the distribution of attack patterns. This follows the taxonomy-driven principle of SAGE-RT (Kumar et al., 2024).

The generation prompt preserves the requested action while varying framing, register, persuasion, entities, viewpoint, and length, including forms associated with jailbreak generalization failures (Wei et al., 2023; Zeng et al., 2024c). Candidates are reclassified under the AIR rubric; only those assigned the intended category are retained.

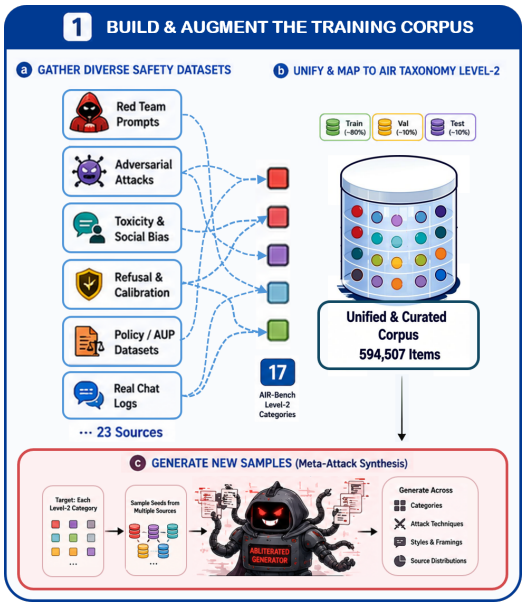



Figure 1: Risk Atlas construction: heterogeneous safety prompts are aligned to AIR Level-2, augmented through category-conditioned synthesis, and checked against the target labels.

### 3.4 Fixed-taxonomy classifier

Pharos implements the task with a bidirectional encoder and a 17-way classification head. One forward pass scores every AIR category, and the highest-scoring class is the prompt label. Unlike a schema-conditioned guard, the label set is learned during training rather than supplied with each query: the design commits to the shared AIR taxonomy in exchange for giving up runtime label flexibility. Class-weighted training addresses the long-tailed category distribution.

| Source | Task | % Harm | Size |
|---|---|---|---|
| wildjailbreak | contrastive jailbreak training pairs | 51 | 263,413 |
| s_eval | auto-generated safety eval prompts | 97 | 106,471 |
| or_bench | borderline-benign over-refusal probes | 1 | 52,326 |
| alert | 32-category red-team prompts | 99 | 44,953 |
| beavertails | harm-labeled QA preference pairs | 56 | 30,192 |
| salad_bench | hierarchical 66-category probes | 100 | 26,313 |
| aegis | human–LLM safety annotations | 50 | 25,612 |
| redbench | aggregated red-team benchmarks | 83 | 24,497 |
| toxic_chat | real chatbot toxicity annotations | 7 | 9,767 |
| air_bench_2024 | regulation-derived risk probes | 100 | 5,688 |
| do_not_answer | refusal-target harmful prompts | 68 | 938 |
| in_the_wild_jailbreak | real-world jailbreak prompts | 95 | 664 |
| categorical_harmful_qa | 11-category harmful QA | 100 | 550 |
| advbench | GCG harmful behaviors | 100 | 520 |
| xstest | safe/unsafe contrast pairs | 41 | 450 |
| sorry_bench | 44-class refusal probes | 88 | 450 |
| forbidden_question_set | 13-scenario forbidden questions | 78 | 390 |
| strong_reject | rubric-graded forbidden prompts | 100 | 313 |
| harmbench | standardized harm behaviors | 100 | 300 |
| hex_phi | 11-category prohibited instructions | 100 | 300 |
| jailbreakbench | OpenAI-policy misuse behaviors | 50 | 200 |
| malicious_instruct | 10-category malicious instructions | 99 | 100 |
| simple_safety_tests | critical-harm smoke tests | 100 | 100 |
| **Total** | | **62** | **594,507** |

Table 1: **Source datasets in the unified corpus**, ordered by contribution. *% Harm* is the proportion of `harmful` vs. `non_harmful` rows from each source. Sizes are row counts after schema unification and label remapping.

# 4 Experimental Protocol

## 4.1 Corpus composition and partitions

The 23 source datasets contribute 594,507 prompts. Synthesis produced 163,715 candidates; a second AIR-category check retained 104,320, of which 83,400 remain in training after split and collision filtering. Source-dataset examples are stratified jointly by source and AIR Level-2 label. Splits are then reconciled by stripped query text: conflicting labels and repeated rows are removed, and a held-out query takes precedence over a training copy. The final partitions contain 518,452 training, 55,148 validation, and 55,568 test items. Validation and test contain no newly synthesized prompts. Appendix Table 5 gives their composition.

## 4.2 Label-audit protocol

We validate the labelling model on a held-out, class-aware sample of 10,000 items whose native source labels map deterministically to AIR Level-2. These mappings provide references that are independent of the judge prediction. DeepSeek-V4-Pro, prompted with the same AIR rubric used during corpus construction, obtains a macro-F1 of 0.97 and Cohen's $\kappa$ of 0.98.

| Labeler check | Value |
|---|---|
| AIR Level-2 macro-F1 | 0.97 |
| Cohen's $\kappa$ | 0.98 |

Table 2: Labeler results where native source categories determine an AIR Level-2 reference independently of the labeler.

## 4.3 Training configuration

The reported run initializes the 17-way classifier from DeBERTa-v3-base (180M parameters). Training uses class-weighted cross-entropy, an effective batch size of 64, fused AdamW at $5 \times 10^{-5}$, cosine learning-rate decay, 5% warmup, weight decay 0.01, gradient clipping at 1.0, and three epochs. We evaluate every 2,000 steps and select the checkpoint with the highest validation macro-F1. The

seed is 42, with early stopping after three validation checks without improvement. DeBERTa is trained in fp16 with fp32 master weights.

### 4.4 Same-data decoder control

To test whether access to AIR-specific training data explains the gap to released guards, a Qwen3-0.6B decoder is trained on the same Risk Atlas training split with a 17-way classification head, class-weighted cross-entropy, and effective batch size 64. This aligns the task and data while leaving architecture and optimization differences.

### 4.5 Additional pretrained-backbone control

As a secondary control, we evaluate the public openjev Qwen3.5-4B artifact zero-shot through its NLI head and after replacing that head with a 17-way classifier trained on the unified corpus. The adapted condition tests transfer from the released backbone, not TypeSafe AI's hosted Jev system (TypeSafe AI, 2026). The zero-shot hypothesis protocol is in the appendix.

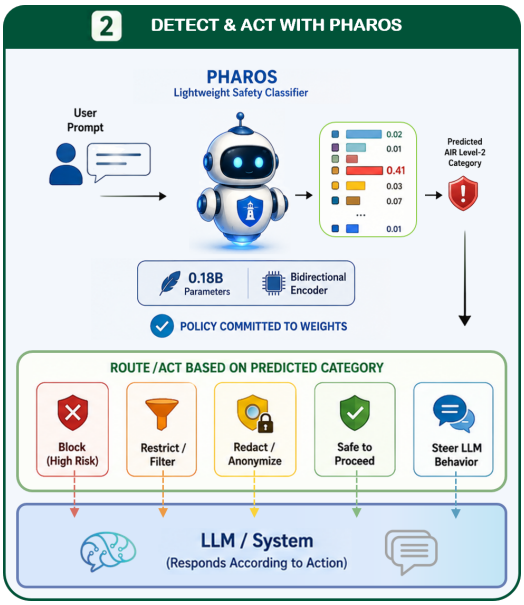



Figure 2: **Detect and route.** A lightweight bidirectional encoder commits a single AIR-L2 category to each prompt in one forward pass, enabling fine-grained category-conditioned downstream routing.

### 4.6 Comparators and metrics

The comparator set includes generative guards with fixed native labels, schema-input guards, and compact encoders. Native categories are mapped to AIR using published model taxonomies; an unmappable harmful category remains harmful for binary scoring but is not assigned a guessed AIR category. For multi-label native outputs, a fixed, documented priority selects one AIR label; this rule is a comparison convention rather than a learned severity ordering. The evaluation code records parse failures, unmappable outputs, and the scored denominator separately. Level-1 and Level-2 scores exclude outputs without an AIR label; binary scores retain harmful-but-unmappable outputs. We report the number of reachable AIR labels, parse failures, and unmappable outputs with each result.

Macro-F1 gives each class equal weight, including rare categories. We also report binary harmful-class F1 and macro-F1 over the five broad groups. For practical use, per-category precision, recall, and confusion are needed: misclassifying a self-harm request as a neighboring content category has a different consequence from confusing two low-risk categories.

## 5 Results

### 5.1 A compact encoder leads on the shared taxonomy

Pharos reaches 0.753 Level-2 macro-F1 on the test set, 0.282 absolute above gpt-oss-safeguard-20B, the strongest released decoder in this comparison, while being 111x smaller. The binary gap is only 0.074, whereas the Level-1 and Level-2 gaps are 0.272 and 0.282. This widening illustrates why binary evaluation is insufficient for a system intended to supply category-level routing signals. Native taxonomies also constrain some guards' reachable AIR labels, so these results measure each released system's utility for AIR classification rather than architecture in isolation.

| Model | Params (B) | Reach | Binary | | | L1 (5 cats) | | | L2 (17 cats) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | P | R | F1 | P | R | F1 | P | R | F1 |
| *Decoder LLM baselines* | | | | | | | | | | | |
| ShieldGemma-9B | 9.0 | 4/17 | 0.956 | 0.377 | 0.541 | 0.208 | 0.319 | 0.247 | 0.104 | 0.160 | 0.125 |
| Qwen3Guard-Gen-8B | 8.0 | 10/17 | 0.941 | 0.862 | 0.900 | 0.529 | 0.487 | 0.479 | 0.261 | 0.290 | 0.260 |
| PolyGuard-Qwen | 7.0 | 13/17 | 0.947 | 0.859 | 0.901 | 0.618 | 0.526 | 0.530 | 0.321 | 0.402 | 0.349 |
| LlamaGuard-4-12B | 12.0 | 13/17 | 0.958 | 0.536 | 0.687 | 0.551 | 0.419 | 0.435 | 0.370 | 0.325 | 0.314 |
| Aegis-2 / NemoGuard | 8.0 | 14/17 | 0.917 | 0.730 | 0.813 | 0.613 | 0.518 | 0.539 | 0.338 | 0.297 | 0.284 |
| RobloxGuard-1.0 | 8.0 | 17/17 | 0.748 | 0.954 | 0.838 | 0.601 | 0.510 | 0.508 | 0.381 | 0.421 | 0.350 |
| gpt-oss-safeguard-20b | 20.0 | 17/17 | 0.862 | 0.829 | 0.845 | 0.599 | 0.592 | 0.587 | 0.503 | 0.532 | 0.471 |
| *Typed-decision models* | | | | | | | | | | | |
| openjev-4B (zero-shot) | 4.0 | 17/17 | 0.745 | 0.916 | 0.822 | 0.476 | 0.459 | 0.423 | 0.379 | 0.341 | 0.268 |
| *AIR-L2-trained decoder controls* | | | | | | | | | | | |
| Qwen3-0.6B | 0.60 | 17/17 | 0.949 | 0.992 | 0.970 | 0.815 | 0.847 | 0.829 | 0.668 | 0.741 | 0.696 |
| openjev-4B (adapted) | 4.0 | 17/17 | 0.957 | 0.981 | 0.969 | 0.841 | 0.851 | 0.845 | 0.737 | 0.719 | 0.721 |
| *Encoder baselines* | | | | | | | | | | | |
| GLiNER-Guard-Omni | 0.21 | 17/17 | 0.776 | 0.808 | 0.792 | 0.584 | 0.553 | 0.553 | 0.385 | 0.415 | 0.366 |
| GLiGuard-300M | 0.30 | 17/17 | 0.631 | 0.993 | 0.772 | 0.398 | 0.239 | 0.148 | 0.248 | 0.145 | 0.089 |
| **Pharos** | **0.18** | 17/17 | 0.957 | 0.993 | **0.975** | 0.842 | 0.880 | **0.859** | 0.718 | 0.794 | **0.753** |

Table 3: **Matched results on the 55,568-item test set.** Every row uses the same source-dataset, query-disjoint prompts and corpus reference labels. Reach counts the AIR Level-2 categories expressible through each model's output mapping.

**Secondary backbone control.** Zero-shot openjev reaches 0.268 Level-2 macro-F1. It exceeds ShieldGemma and is close to Qwen3Guard despite receiving no AIR-specific training. Adaptation to AIR Level-2 raises the 4B model to 0.721 on the test set. This 0.453-point zero-shot-to-adapted gain is larger than the spread among the released decoder guards, emphasizing that target-taxonomy supervision is the dominant factor. Pharos remains 0.032 ahead while using 22 times fewer parameters.

### 5.2 Same-data decoder control

The same-data decoder reaches 0.696 Level-2 macro-F1 on the 55,568-item source-dataset test set. Its 0.225-point advantage over the strongest released guard is consistent with a benefit from AIR-specific supervision, although architecture and training data also differ. The decoder trails the 180M-parameter Pharos encoder by 0.057 despite having 3.3 times as many parameters. It shows that access to the unified supervision alone does not explain results.

### 5.3 The encoder has low measured inference time

On one RTX 4080-SUPER, Pharos's median inference time is 5.36 ms per query at batch size one, including tokenization, host-to-device transfer, and a synchronized model pass. Length-bucketed batched inference reaches 1.76 ms per query over the measured workload; a fixed batch of 32 takes 4.07 ms per query. Table 4 reports these measurements in the appendix.

## 6 Discussion and Limitations

Detecting risk and identifying its category emerge as meaningfully different tasks. Differences among guards are modest under binary scoring but reorder and widen at granular risk level. Risk Atlas makes these distinctions visible across sources, and Pharos's measured latency makes fine-grained prompt-time routing plausible.

The label audit provides evidence that the judge can apply the AIR Level-2 rubric accurately when an independent category reference is available. While it achieves 0.97 macro-F1 and 0.98 Cohen's $\kappa$, these results support confidence in the labeling procedure for the cases the audit covers, but they do not directly validate judge-assigned categories across the full corpus. Sources with only binary

reference labels establish whether a prompt is harmful, not which AIR risk it represents, so they cannot provide the fine-grained ground truth required for this check.

The corpus is predominantly English. Pharos classifies a prompt before a model responds, so it does not measure response safety, refusal, conversational context, PII spans, or multilingual performance. Released guards may be better suited to these other tasks. The source-matched test measures transfer across examples within known sources; a fully source-held-out result remains important for claims about unfamiliar distributions.

The 17-way formulation assigns one primary label even when a prompt implicates several risks. This keeps training and evaluation well defined, but it compresses multi-intent prompts and does not encode severity. A deployment could use the full score vector for escalation or multi-label decisions, but those policies require calibrated probabilities and application-specific costs that we do not evaluate.

The typed-decision results apply to the public openjev artifact. They do not measure the proprietary Jev system or its unreleased calibrated-decision objective. The adapted condition is an openjev-initialized classifier trained under a constrained single-GPU budget; its result demonstrates transfer from the released backbone, not an upper bound on large typed-decision models.

## 7 Conclusion

A useful prompt safety system must distinguish what kind of risk a request raises, since different categories call for different handling. Risk Atlas brings 23 heterogeneous public sources into the AIR Level-2 label space with source provenance, auditable label assignments, and held-out splits that exclude newly synthesized prompts. On this shared task, Pharos attains 0.753 Level-2 macro-F1 and 5.36 ms median batch-one latency on a consumer grade GPU; the 180M-parameter encoder outperforms the evaluated released guards and a decoder trained on the same corpus. The much wider performance gap at Level 2 than under binary scoring, together with the gain from adapting openjev to AIR, shows why taxonomy alignment and fine-grained evaluation matter when selecting a guard. In applications, category predictions can inform distinct refusal, redirection, escalation, or review paths before generation, while the surrounding policy determines the appropriate action. Risk Atlas and Pharos provide a shared basis for studying that workflow and a lightweight starting point for implementing it.

## Ethics Statement

The corpus contains harmful requests from public safety datasets. The synthetic training set also contains generated adversarial requests. Unifying these prompts and varying their attack framings could make harmful requests easier to find or adapt. Releasing either resource requires documenting source licences, handling access consistently with them, and describing the possible misuse of the data. Pharos is a classifier of prompt risk categories, not a guarantee against misuse. False positives can block legitimate requests, including requests for help on sensitive topics; false negatives can leave harmful requests unflagged. A predicted category describes a prompt under the AIR rubric, not the user's intent or the safety of a subsequent response. Performance may also vary across language varieties and communities underrepresented in the predominantly English corpus. Decisions in high-stakes settings require policy rules and human oversight appropriate to their context.

## AI Use Statement

LLM assistance was used for statistical exploration, analysis code, figures, and drafting. The authors are responsible for the final manuscript and its claims.

## Reproducibility Statement

Our code is open sourced. Pharos will be released upon acceptance, while Risk Atlas dataset will be released with a on-request approval.

## REFERENCES


Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*, 2023.

Cem Anil, Esin Durmus, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Nina Rimsky, Meg Tong, Jesse Mu, and D. Ford. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems*, 2024.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Maciej Chrabaszcz, Filip Szatkowski, Bartosz Wójcik, Jan Dubiński, Tomasz Trzciński, and Sebastian Cygert. Efficient LLM moderation with multi-layer latent prototypes. *arXiv preprint arXiv:2502.16174v3*, 2025.

J. Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*, 2024.

Quy-Anh Dang, Chris Ngo, and Truong-Son Hy. Redbench: A universal dataset for comprehensive red teaming of large language models. *arXiv preprint arXiv:2601.03699*, 2026.

Yue Deng, Wenxuan Zhang, S. J. Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*, 2023.

Nirav Diwan, Hanqi Wang, B. Kapusuzoglu, Ramin Moradi, Supriyo Chakraborty, Giri Iyengar, Sambit Sahu, Huan Zhang, and Gang Wang. Cot-guard: Small models for strong monitoring. *arXiv preprint arXiv:2605.12746*, 2026.

Xuanyi Fu and Danbing Zou. A comparative empirical study of over-refusal behavior in closed-source large language models on pseudo-harmful prompts. *Artificial Intelligence and Machine Learning Review*, 2025.

Shaona Ghosh, Prasoon Varshney, Erick Galinkin, and Christopher Parisien. Aegis: Online adaptive AI content safety moderation with ensemble of LLM experts. *arXiv preprint arXiv:2404.05993*, 2024.

Shaona Ghosh, Heather Frase, Adina Williams, Sarah Luger, Paul Röttger, et al. Ailuminate: Introducing v1.0 of the AI risk and reliability benchmark from MLCommons. *arXiv preprint arXiv:2503.05731*, 2025a.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. Aegis 2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. *arXiv preprint arXiv:2501.09004*, 2025b.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. *arXiv preprint arXiv:2406.18495*, 2024.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*, 2022.

Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source LLMs via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2023.

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

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *arXiv preprint arXiv:2309.00614*, 2023.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. *arXiv preprint arXiv:2307.04657*, 2023.

Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu, Maarten Sap, and Yejin Choi. Wildteaming at scale: From in-the-wild jailbreaks to (adversarially) safer language models. *arXiv preprint arXiv:2406.18510*, 2024.

Anurakt Kumar, Divyanshu Kumar, Jatan Loya, Nitin Aravind Birur, Tanay Baswa, Sahil Agarwal, and Prashanth Harshangi. Sage-rt: Synthetic alignment data generation for safety evaluation and red teaming. *arXiv preprint arXiv:2408.11851*, 2024.

Priyanshu Kumar, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. Polyguard: A multilingual safety moderation tool for 17 languages. *arXiv preprint arXiv:2504.04377*, 2025.

Jing-Jing Li, Valentina Pyatkin, Max Kleiman-Weiner, Liwei Jiang, Nouha Dziri, Anne G. E. Collins, Jana Schaich Borg, Maarten Sap, Yejin Choi, and Sydney Levine. Safetyanalyst: Interpretable, transparent, and steerable safety moderation for AI behavior. *arXiv preprint arXiv:2410.16665*, 2024a.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024b.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation. *arXiv preprint arXiv:2310.17389*, 2023.

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

Anay Mehrotra, M. Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum S. Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. *arXiv preprint arXiv:2312.02119*, 2023.

Bogdan Minko, Sabrina Sadiekh, and Evgeniy Kokuykin. Gliner guard: Unified encoder family for production LLM safety and privacy. *arXiv preprint arXiv:2605.05277*, 2026.

Mahesh Kumar Nandwana, Youngwan Lim, Joseph Liu, Alex Yang, Varun Notibala, and Nishchaie Khanna. Taxonomy-adaptive moderation model with robust guardrails for large language models. *arXiv preprint arXiv:2512.05339*, 2025.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*, 2022.

Alexander Robey, Eric Wong, Hamed Hassani, and George Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. Xstest: A test suite for identifying exaggerated safety behaviours in large language models. *arXiv preprint arXiv:2308.01263*, 2023.

Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. The AI risk repository. *AGI — Artificial General Intelligence — Robotics — Safety & Alignment*, 2024.

TypeSafe AI. Jev: A system one model for calibrated typed decisions. https://www.datacamp.com/blog/system-one-models-jev, 2026. Early-access system; weights, training data, and the full calibrated-decision objective are not public.

Bibek Upadhayay and Vahid Behzadan. X-guard: Multilingual guard agent for content moderation. *arXiv preprint arXiv:2504.08848*, 2025.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: A dataset for evaluating safeguards in LLMs. *arXiv preprint arXiv:2308.13387*, 2024.

Alexander Wei, Nika Haghtalab, and J. Steinhardt. Jailbroken: How does LLM safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John F. J. Mellor, A. Glaese, Myra Cheng, Borja Balle, and Atoosa Kasirzadeh. Taxonomy of risks posed by language models. *Conference on Fairness, Accountability and Transparency*, 2022.

Alex Wortega. openjev: Qwen3.5 trained as a jev model. https://huggingface.co/AlexWortega/openjev, 2026. Open Qwen3.5-4B NLI cross-encoder reproduction of the typed-decision interface.

Urchade Zaratiana, Mary Newhauser, George Hurn-Maloney, and Ash Lewis. Gliguard: Schema-conditioned classification for LLM safeguard. *arXiv preprint arXiv:2605.07982*, 2026.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, and Bhaktipriya Radharapu. Shieldgemma: Generative AI content moderation based on gemma. *arXiv preprint arXiv:2407.21772*, 2024a.

Yi Zeng, Kevin Klyman, Andy Zhou, Yu Yang, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Ai risk categorization decoded (air 2024): From government regulations to corporate policies. *arXiv preprint arXiv:2406.17864*, 2024b.

Yi Zeng, H. Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. *arXiv preprint arXiv:2401.06373*, 2024c.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, et al. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*, 2025.

Andy Zou, Zifan Wang, J. Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

## A INFERENCE LATENCY

| Pharos inference mode | Median ms/query |
|---|---|
| Batch 1 | 5.36 |
| Length-bucketed | 1.76 |
| Batch 32 | 4.07 |

Table 4: Measured Pharos inference latency on one RTX 4080-SUPER, including tokenization, host-to-device transfer, and a synchronized model pass.

## B SPLIT CONSTRUCTION AND REPRODUCIBILITY

Split construction uses stripped query text as the grouping key. Source-dataset query groups retain their held-out assignment when a copy also occurs in training; training copies are removed. Groups with more than one AIR Level-2 label are discarded, and remaining within-group copies collapse to one row. Synthetic rows contribute only when their original assignment is training and their text does not collide with validation or test. The final counts are repeated in Table 5.

| Split | Rows | Source-dataset | Synthetic | Unique queries |
|---|---|---|---|---|
| Train | 518,452 | 435,052 | 83,400 | 518,452 |
| Validation | 55,148 | 55,148 | 0 | 55,148 |
| Test | 55,568 | 55,568 | 0 | 55,568 |

Table 5: Final split composition. Pairwise exact-query overlap is zero.

## C SYNTHETIC-DATA QUALITY CONTROLS

The synthesis pipeline generates 163,715 candidates and retains 104,320 after a second attribution pass agrees with the intended AIR category, a retention rate of 63.7%. This agreement filter is deliberately conservative: it removes category-drifted generations before splitting, but it is still model-based validation rather than human annotation. The final training split contains 83,400 of the retained prompts after held-out and collision filtering.

We additionally compare synthetic prompts with the source-dataset training corpus using character $n$-gram TF–IDF cosine similarity. The median nearest-neighbour similarity is 0.41, the 95th percentile is 0.67, and 1.1% exceed 0.80. These statistics suggest that most retained generations are not trivial surface duplicates of their nearest source-dataset neighbours. They do not rule out shared templates or semantic paraphrases, which motivates keeping every synthetic item outside validation and test.

## D TYPED-DECISION HYPOTHESIS SELECTION

The openjev NLI reduction requires one hypothesis per AIR category. We compare three families on the source–label-stratified validation subset and select the family before test evaluation. The `parallel` family uses the same sentence frame for every category (for example, "The user is asking for help committing a crime"), varying only the category-bearing phrase. The `rerank` family inserts bare category names into a multiple-choice-style template. The `policy` family uses the longer AIR descriptions.

The 0.081 spread shows that hypothesis design is part of the model specification for NLI-based typed decisions. Structurally parallel hypotheses outperform the more informative but unevenly phrased policy descriptions, suggesting that sentence-form effects can dominate the semantic benefit of added detail.

| Hypothesis family | Form | Validation macro-F1 |
|---|---|---|
| `parallel` | fixed frame, category phrase varies | **0.267** |
| `rerank` | bare label in a ranking template | 0.216 |
| `policy` | full category descriptions | 0.186 |

Table 6: Openjev zero-shot hypothesis selection. The selected `parallel` family is used unchanged on the test set.

## E PER-CLASS PHAROS RESULTS

Table 7 reports full-test F1 for every AIR Level-2 category. The aggregate result is driven by broadly strong performance rather than one dominant class, but fundamental rights, economic harm, manipulation, and political usage remain the clearest targets for improvement.

| Category | F1 | Category | F1 |
|---|---|---|---|
| child harm | 0.813 | non-harmful | 0.958 |
| criminal activities | 0.848 | operational misuses | 0.667 |
| deception | 0.782 | political usage | 0.647 |
| defamation | 0.720 | privacy | 0.890 |
| discrimination/bias | 0.786 | security risks | 0.809 |
| economic harm | 0.632 | self-harm | 0.738 |
| fundamental rights | 0.389 | sexual content | 0.793 |
| hate/toxicity | 0.842 | violence/extremism | 0.838 |
| manipulation | 0.647 | | |

Table 7: Pharos per-class F1 on the full 55,568-item test set.

## F NATIVE-TO-AIR MAPPING DECISIONS

Fixed-taxonomy guards cannot always express all AIR categories. The mappings below record the decision used for each native category. Semantically broad categories should be treated as unmappable when assigning one AIR risk would imply information the guard did not provide.

Table 8: Selected native-taxonomy mappings that most affect interpretation.

| Guard | Native category | AIR target |
|---|---|---|
| Llama Guard 4 | Specialized Advice | operational_misuses |
| Llama Guard 4 | Sex-Related Crimes | sexual_content |
| Aegis 2.0 | Needs Caution | unmappable harmful |
| Aegis 2.0 | Other | unmappable harmful |
| Qwen3Guard | Unethical Acts | hate_toxicity |
| ShieldGemma | Dangerous Content | violence_extremism |