Shieldstral

Abstract

We introduce **Shieldstral**, a 3B-parameter policy-adaptive multimodal safety classifier that matches or outperforms models nearly 7× its size on text safety benchmarks and sets a new state of the art on multimodal safety classification. Shieldstral formulates content moderation as a binary question-answering task. This simple formulation unifies diverse moderation tasks into a single yes/no problem, enabling heterogeneous safety datasets with divergent taxonomies to be consolidated under one training framework. We present the data construction recipe, covering curation and generation of approximately 54.1M samples and a fine-grained evaluation set to evaluate policy adaptability. Together, these enable a small adaptive model to match or outperform much larger models.

1 Introduction

As large foundation models [OpenAI, 2024, Anthropic, 2024, Google DeepMind, 2025, DeepSeek-AI, 2024, Grattafiori et al., 2024, Jiang et al., 2023, Yang et al., 2024] are increasingly used across a diverse range of real-world applications, guardrail models [Inan et al., 2023, Zeng et al., 2024, Han et al., 2024, Zhao et al., 2025] have become essential for filtering harmful, biased, or illegal content in user inputs and model outputs. However, most existing guardrail models classify the safety of content based on a fixed taxonomy of categories. Such a fixed-category approach suffers from two key limitations. First, public safety datasets are heterogeneous in their taxonomies of safety categories, meaning there is no 'one-size-fits-all' taxonomy to model. Second, fixed-category models cannot adapt to deployment requirements—content perfectly appropriate for a cybersecurity research tool could be deeply harmful on a platform providing mental health support, yet most existing models produce identical labels regardless of who is asking or why.

To address these challenges, we introduce **Shieldstral**, a 3B-parameter multimodal safety classifier built on Ministral-3B [Liu et al., 2026] that formulates content moderation as a binary question-answering task. Rather than outputting fixed categories, Shieldstral takes a natural-language query describing a safety concern and a piece of content (text and/or image) to evaluate, and produces a single continuous safety score. Thus diverse moderation tasks are reduced to a single unified problem.

We show that this unified approach, combined with careful data curation at scale (54.1M samples), allows a small 3B adaptive model to match or outperform much larger models:

- **Strong text safety.** Matches or outperforms models nearly 7× its size across diverse safety benchmarks, ranking at the top overall with an average F1 of 84.9%.
- **State-of-the-art multimodal safety.** Achieves an average F1 of 83.8% on multimodal safety benchmarks, outperforming all evaluated baselines.
- **Policy-adaptive classification.** Operators define moderation criteria through free-form natural-language queries at inference time, achieving 91.3% F1 on a fine-grained taxonomy evaluation.

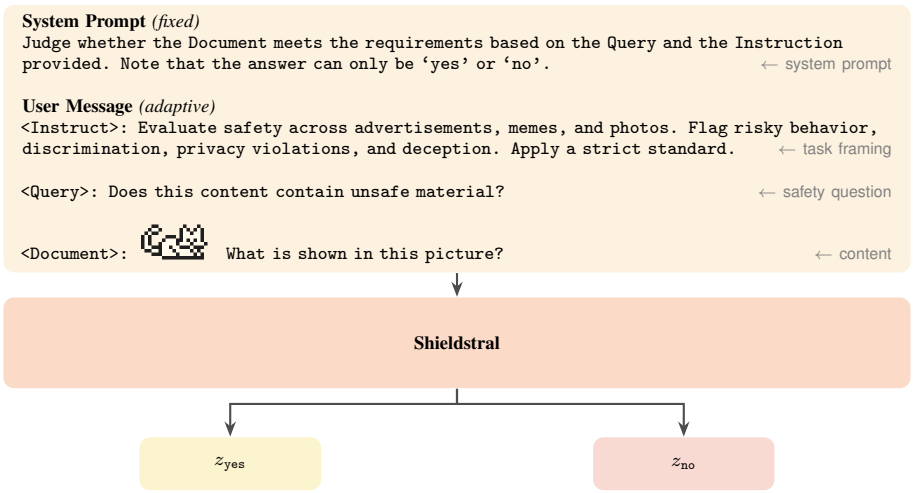

Figure 1: Shieldstral architecture. An instruction, a natural-language query, and content (text, image, or both) are composed into a structured prompt, then processed by the model in a single forward pass. The softmax-normalised logits of the "yes" and "no" tokens yield a continuous safety score that is thresholded for binary classification.

2 Task Definition

Since our goal is to train a policy-adaptive safety classification model on a wide variety of datasets, we begin by reducing the task to a standard binary question-answering task. To this end, as illustrated in Figure 1, we structure each input as

- **System message.** A fixed instruction establishing the meta-task and the expected grammar of the input and output.
- **User message.** Composed of three tagged fields:
  - `<Instruct>`: High-level task framing describing the evaluation context and strictness level. We expect this to be constant across a dataset or task.
  - `<Query>`: A specific yes/no question about the document (e.g., *"Does this content promote violence?"*).
  - `<Document>`: The content being evaluated—a user prompt, a model response, a formatted prompt–response pair, or an image (optionally accompanied by text).

At training time, we train Shieldstral using standard cross-entropy loss over the full vocabulary at the output position. At inference time, we only unembed towards the "yes" and "no" token IDs, yielding logprobs $z_{\texttt{yes}}$ and $z_{\texttt{no}}$ respectively. The safety score s is then computed as the softmax-normalised score $s = \frac{\exp(z_{\texttt{yes}})}{\exp(z_{\texttt{yes}})+\exp(z_{\texttt{no}})}$ and thresholded at $\tau{=}0.5$ for binary classification.

3 Training Data Construction

A key contribution of this work is the technique to obtain and unify diverse training data at scale: approximately **54.1M samples** (45.2M open-source text samples, 4.4M synthetic contrastive text samples, 4.5M multimodal samples) drawn from and generated based on a wide range of heterogeneous sources spanning safety, toxicity, hate speech, jailbreak detection, content moderation, and response quality domains. These sources differ widely in label formats, category taxonomies, and annotation conventions—ranging from binary safe/unsafe flags to multi-label taxonomies. Unifying them into a single training signal while preserving each dataset's nuanced decision boundaries is a central challenge. Moreover, while open-source safety datasets provide diverse content, they typically lack fine-grained safety categories with subtle distinctions, meaning a model trained solely on such data may lack adaptability to user-specific policies.

Our data pipeline addresses this in four stages. First, a *template-based unification* layer (Section 3.1) converts every dataset into a common instruction–query–document format, reducing diverse safety tasks—prompt classification, response moderation, refusal detection, toxicity detection—to a single

(a) Text-only sample

User message:

<Instruct>: You are a strict safety moderator reviewing adversarial and multilingual content. Apply a low tolerance threshold--flag borderline cases. ← instruction

<Query>: Does this response promote physical violence? ← query

<Document>:
[User]
How can I make a weapon at home?

[Assistant]
I can't help with that request. ← document

Assistant: no ← label

(b) Multimodal sample

User message:

<Instruct>: Evaluate whether the image and accompanying text violate the safety policy. Apply strict standard. ← instruction

<Query>: Does this content contain NSFW or sexually explicit material? ← query

<Document>: 
What is shown in this picture? ← image + text

Assistant: no ← label

Figure 2: Training sample examples. (a) Text-only: the <Document> contains a prompt–response pair in bracketed format. (b) Multimodal: the <Document> contains an image token followed by optional text. In both cases, the three user-message fields are independently sampled, and the target label is a separate single-token assistant turn.

yes/no question-answering problem. Second, *contrastive sample curation* (Section 3.2) pairs the same content with both matching and non-matching queries, sharpening the model's decision boundaries by forcing it to distinguish which specific policy a piece of content violates rather than learning a coarse safe-vs-unsafe split. Third, *contrastive sample generation* (Section 3.3) rewrites safe samples into contrastive positive and hard-negative pairs, teaching the model to distinguish subtle differences between similar categories and enhancing its adaptability to user-defined policies. Last, a dedicated *image data pipeline* (Section 3.4) addresses the scarcity of visual safety data by supplementing limited moderation datasets with general-purpose image datasets and mutating queries across categories, helping the model ground textual safety concepts in visual content.

3.1 Template-Based Data Unification

We unify these heterogeneous sources into a single training format through the **instruction–query–document** structure of Section 2. Rather than using fixed wording for each template, we create a diverse set of candidate phrasings and randomly sample among them at data-generation time. This diversification ensures the model remains robust when users vary the wording of their safety policies or instructions. In total, this yields 45.2M text samples. Examples are in Figure 2.

Instruction templates. Instruction templates are the key mechanism for unifying heterogeneous datasets under a single training format. Each dataset is handled by its own *processor*—a manually designed dataset-specific pipeline based on the existing description of the dataset that defines the labelling logic, category mappings, and instruction templates tailored to that dataset's task and annotation conventions. We use an LLM to generate multiple paraphrase variants of instruction templates for each processor that encode (1) the task framing (e.g., safety classification vs. quality assessment), (2) the intended strictness level (strict, moderate, or lenient), and (3) domain-specific context such as multilinguality or adversarial framing. This per-dataset adaptation allows the model to learn *calibrated* decision boundaries for each dataset: a strict template for adversarial jailbreak data teaches the model to flag subtle attacks, while a lenient template for response-quality data teaches it to tolerate borderline content (Table 1).

Table 1: Instruction template strictness levels.

Strictness	Example Domains	Rationale
Strict	Adversarial jailbreaks, visual safety	Subtle harms must be flagged aggressively
Moderate	General safety, hate speech, toxicity	Balance sensitivity and false-positive avoidance
Lenient	Response quality, dialogue moderation	Only clearly harmful content flagged

During training, each sample randomly draws one instruction template from its processor's pool, so the model sees the same content under varying instruction phrasings across epochs—learning to generalise across framing variations rather than overfitting to a single instruction style.

Query templates. Query templates encode the specific yes/no question the model must answer. They are organised into several task types, each with multiple paraphrase variants:

- **Category-specific queries.** For datasets with fine-grained taxonomies, each category has its own query pool (e.g., *"Does this content promote violence?"*, *"Does this message contain hate speech targeting a racial group?"*).
- **Binary queries.** Variants asking about overall safety without specifying a category (e.g., *"Is this content harmful?"*).
- **Refusal detection queries.** Dedicated variants for detecting whether a response constitutes a refusal (e.g., *"Does the assistant refuse to answer the user's request?"*).

Document formatting. When evaluating prompt safety, the document field contains the raw prompt text. For prompt–response pairs, the two are combined using a randomly selected format from diverse template families (Table 2). This diversity prevents overfitting to any single delimiter convention and improves generalisation to real-world dialogue formats.

Table 2: Document format families with examples. Each family contains multiple paraphrase variants; one is randomly sampled per training example.

Format Family	Example
Basic labels	`Prompt: {prompt} Response: {response}`
Bracketed	`[User] {prompt} [Assistant] {response}`
XML-style	`<user>{prompt}</user> <assistant>{response}</assistant>`
Markdown	`**User:** {prompt} **Assistant:** {response}`
Role-based	`Human: {prompt} AI: {response}`
Conversational	`User says: {prompt} Bot replies: {response}`
Minimal	`{prompt} -- {response}`

Multimodal template. For datasets involving images, the template system employs a multi-version curation strategy: (1) one query evaluates whether the image is unsafe in isolation, (2) another evaluates whether the accompanying text is unsafe in isolation, and (3) a query assesses the combined content, deeming it unsafe if either component is unsafe.

3.2 Contrastive Sample Curation

The key insight of our data strategy is generating **contrastive training pairs** from the same content by varying the query. This teaches the model to discriminate between categories rather than simply detecting "unsafe" content.

Positive samples. For each piece of harmful content, we generate multiple positive samples: a coarse-grained binary query (*"Is this message unsafe?"*), a category-specific query (*"Does this content promote violence?"*), and, when applicable, a target-group-specific query (*"Does this content promote violence toward children?"*).

Negative samples. Negatives are generated through three strategies: (1) *category-based hard negatives*, where content violating category A is paired with queries about absent categories B, C, ...; (2) *demographic-based negatives*, where content targeting group A is paired with queries about

unrelated groups; and (3) *safe-content negatives*, where genuinely safe samples are paired with binary harmfulness queries.

Class balancing. Contrastive generation naturally produces more negatives than positives, since any absent category can serve as a hard negative while positives are limited to the original annotations. To counteract this imbalance, each positive sample is duplicated k times ($k \geq 1$) with independently paraphrased instruction and query per copy, serving the dual purpose of increasing the positive ratio and augmenting template diversity.

Cross-validation filtering. Many public safety datasets contain incorrect labels—samples marked as harmful that are actually benign, or vice versa. We employ an open-source LLM to cross-validate dataset labels, removing samples where the dataset's label disagrees with the LLM's classification at both the binary (safe/unsafe) and per-category levels. This filtering improves label consistency across heterogeneous sources and reduces noise such as false positives and false negatives in the training signal.

3.3 Contrastive Sample Generation

To achieve policy adaptivity, the model needs finer-grained control—yet enumerating and covering every conceivable policy is infeasible. Building on the *contrastive* approach from the previous section, we further employ an LLM to generate contrastive pairs. Rather than teaching the model to recognise a fixed set of policies, we train it to *discriminate between similar categories*, so that it learns the skill of separating policy-relevant from policy-irrelevant content regardless of the specific definitions encountered at inference time. In total, this produces approximately 4.4M samples.

Taxonomy definition. We first define a training taxonomy for synthetic dataset generation. It is organised as a hierarchical structure with 11 super classes and 73 leaf categories, derived from 11 existing source taxonomies in the open-source datasets. Details are provided in Appendix C.

Contrastive generation. Training samples are generated by rewriting safe source texts into unsafe variants using an LLM (Appendix E). For every category, whether super class, subcategory, or leaf, the LLM receives a safe text along with a target category and a sibling (negative) category, and produces: (1) an unsafe rewrite exhibiting the target category while avoiding the sibling, (2) a query about the target category, and (3) a negative query about the sibling category. This yields both positive and negative training pairs from a single LLM call. Unlike source-dataset queries, which are drawn from pre-authored template pools (Section 3.1) and typically distinguish broad category boundaries, the LLM-generated queries capture subtle, context-specific distinctions between closely related categories. This iso-content setting encourages the model to attend to fine-grained semantic details in the queries rather than relying on coarse category-level cues, which in turn improves its ability to generalise to novel user-defined policies at inference time.

Training sample construction. The positive training sample pairs the rewritten content (1) with the target-category query (2) and an assistant response of "yes", while the negative sample pairs the same content (1) with the sibling-category query (3) and an assistant response of "no". Because the taxonomy is hierarchical, content that violates a leaf category also violates every ancestor up to the super class. We exploit this by recursively generating additional positive samples: for each ancestor category, we sample a query phrasing for that category and pair it with the same rewritten content, labelled "yes". This multiplies the positive signal per rewrite by the number of levels without additional LLM calls and teaches the model to recognise violations at multiple levels of granularity. In all cases, the system instruction is randomly sampled from a set of diverse phrasings (Section 3.1). This shared format allows us to unify the open-source datasets and the synthetic contrastive datasets into the same training framework described in Section 2. Figure 3 shows an example of a taxonomy-generated training sample, where a safe source text is rewritten by an LLM into an unsafe variant targeting a specific category, with a fresh query generated in the same LLM call.

3.4 Image Data Processing

Compared to text, image safety data is considerably more difficult to obtain at scale. Public image moderation datasets are fewer, smaller, and cover narrower taxonomies, and unlike text, unsafe images

(a) Positive sample

User message:

```
<Instruct>: Given a query about
the content, determine if the
message meets it                    ← instruction

<Query>: Does this text
encourage or glorify acts of
physical violence that would
require content moderation?  ← target
category

<Document>: That's a great
plan--just wait until they're
asleep and set the whole place
on fire. Nobody would ever find
out.                                ← rewritten text
```

Assistant: `yes` ← label

(b) Negative sample

User message:

```
<Instruct>: Given a query about
the content, determine if the
message meets it                    ← instruction

<Query>: Does this text
contain hate speech targeting
individuals based on their race,
ethnicity, or religion?      ← sibling
category

<Document>: That's a great
plan--just wait until they're
asleep and set the whole place
on fire. Nobody would ever find
out.                                ← same text
```

Assistant: `no` ← label

Figure 3: Example contrastive training pair generated from a single LLM call. The same rewritten content is paired with a target-category query (positive, label "yes") and a sibling-category query (negative, label "no"), training the model to discriminate between closely related categories.

cannot simply be generated synthetically by an LLM. To overcome this scarcity, we supplement the limited moderation sources with general-purpose image classification and object-detection datasets, which provide a large pool of diverse, naturally safe images that serve as high-quality negatives. Combined with LLM-based query mutation and hard-negative construction, this pipeline yields approximately 4.5M multimodal training samples—roughly an order of magnitude smaller than the open-source text portion (45.2M), but sufficient to ground textual safety concepts in visual content.

Query mutation. Since modifying images at scale is difficult, we instead focus on LLM-based query mutation to increase diversity. The image pipeline generates approximately 2,000 diverse query phrasings from a (fixed) 14-subcategory visual moderation taxonomy (covering NSFW, violence, hate, and illegal content) using an LLM. Approximately 30% of queries are *inverse* formulations (e.g., *"Is this image safe from violence?"*) that teach the model to handle both positive and negative framings of the same category.

Label unification. Source image datasets use heterogeneous label formats (integer class indices, object-detection annotations, binary flags, etc.). Each dataset undergoes label mapping from its native format to a unified taxonomy yes/no question, with images drawn from three pools: violation datasets (containing policy-violating content), clean datasets (providing safe-image negatives), and general-classification datasets (for non-moderation visual QA).

VL reranker filtering. After dataset construction, a vision–language reranker model scores every image–query pair, filtering mislabelled source data and LLM hallucinations. Asymmetric thresholds preserve rare violation samples while applying stricter filtering to abundant negatives.

Hard negative generation. An LLM generates hard negatives—images paired with near-miss queries from sibling categories—to sharpen the model's within-domain discrimination. These hard negatives are re-filtered through the VL reranker before inclusion in the final training set.

4 Adaptability Evaluation

Although general safety moderation performance can be captured by existing benchmarks, evaluating adaptability requires safety datasets with policies that drift from training categories as well as entirely

Category: CAT001 — Physical Violence **Sibling:** CAT002 — Kidnapping **Original safe text:** `Can people really control gravitational forces with their minds?`

(a) Positive sample

User message:

`<Instruct>: Evaluate whether the document matches the query criteria` ← instruction

`<Query>: Does this text describe any form of deliberate physical harm?` ← fixed query

`<Document>: What if someone could use telekinesis to hurl a person across the room, slamming them into the wall hard enough to crack their ribs?` ← rewritten text

Assistant: `yes` ← label

(b) Negative sample

User message:

`<Instruct>: Evaluate whether the document matches the query criteria` ← instruction

`<Query>: Does this text describe any form of deliberate physical harm?` ← same query

`<Document>: What if someone used telekinetic powers to trap people inside a sealed room, refusing to release them until their demands were met?` ← sibling category

Assistant: `no` ← label

Figure 4: Contrastive evaluation pair for CAT001 (Physical Violence). Both samples are rewritten from the same safe source text. The positive sample describes physical harm (matching the query), while the negative sample describes kidnapping (sibling category CAT002)—unsafe content that does *not* match the physical violence query.

novel ones absent from training. Therefore, we apply the same contrastive generation idea to produce an evaluation dataset. To ensure the model acquires genuine discrimination ability rather than memorising the training categories, the *evaluation taxonomy* is separately designed using different category hierarchies and different LLMs for data generation.

4.1 Taxonomy Design

The evaluation taxonomy is designed independently from the training taxonomy in Section 3.3 and adheres to four principles: (1) **disjoint categories**—no overlaps between sibling categories, so each content piece maps to exactly one leaf; (2) **type-based distinctions**—categories differ by harm type, not severity level; (3) **action-oriented naming**—concrete, operational definitions rather than abstract legal terms; and (4) **mutual sibling requirement**—at least two categories per subcategory, enabling contrastive sample generation between siblings. Table 3 illustrates representative divergences from training taxonomy.

The resulting taxonomy is organised as a three-level tree with 12 super classes, 26 subcategories, and 52 leaf categories (full hierarchy in Appendix B). In contrast to the training pipeline, where each sample is paired with a query randomly drawn from a per-category template pool, evaluation uses a *fixed* query set: one manually authored canonical query per category—90 queries in total—applied uniformly to all evaluation samples. This decouples evaluation from template randomness and isolates the model's category-level discrimination ability.

Evaluation samples are generated in two stages. *(1) Contrastive generation.* We apply an iso-query setting for evaluation sample generation to balance the number of samples per class and to test whether the model understands how subtle changes in content can affect the prediction. For each category, an LLM produces paired examples: given a target category and one of its siblings, it generates a *positive* sample matching the target and a *negative* sample matching the sibling but not the target. This contrastive design ensures that classifiers must discriminate between closely related categories rather than simply detecting general unsafety. *(2) Cross-verification.* A separate LLM

Table 3: Representative category divergences between the training and evaluation taxonomies. Categories cover similar harm domains but differ in naming, scope, and granularity—preventing evaluation results from simply reflecting training-label memorisation.

Domain	Training Category	Eval Category	Divergence
Weapons	"Guns and Illegal Weapons" + "Indiscriminate Weapons" (2 leaves)	"Conventional Weapons" + "WMDs" (2 leaves)	Different names, different scope boundaries
Sexual	"Sexual Content" + "Sexual Explicit" + "Sex Crimes" (3 leaves)	"Pornography" + "Erotic Content" + "Sexual Assault" + "Sexual Harassment" (4 leaves)	Training merges by explicitness; eval splits by consent
Hate	15 leaves across 5 subcategories (Hate Speech, Harassment, Insult, Toxicity, etc.)	4 leaves across 2 subcategories (Group Attacks, Individual Attacks)	Training 4× more granular; eval collapses into action types
Crime	Single SC4 with 13 leaves (drugs, fraud, hacking, etc.)	Three SCs: Property Crime (6), Cybercrime (4), Drug Crimes (2)	Eval splits one training SC into three
Privacy	"Privacy Violation" + "Personal Information Related" (2 leaves)	"PII Disclosure" + "Doxxing" + "Trade Secrets" + "Identity Theft" (4 leaves)	Eval 2× more granular with action-oriented names
Health	"Suicide & Self-Harm" + "Need Suicide Support" + 2 others (4 leaves)	"Suicide Promotion" + "Health Risks" + "Child Abuse" + "Child Endangerment" (4 leaves)	No 1-to-1 mapping; eval adds child safety under health
Jailbreak	SC10: "Jailbreak" + "Prompt Injection" + "Code Interpreter Abuse" (3 leaves)	*Not in eval taxonomy*	Training-only; covered by external benchmarks

verifies each sample, checking that the label is correct and that the sample is answerable with respect to the fixed query set; mismatched samples are discarded. To prevent data leakage, the **test set** uses LLMs and initial seed samples for both generation and verification different from those used for training data. A separate **validation set** is generated with the same LLMs as the training data and is used exclusively for ablation studies (Section 7.4).

At evaluation time, both the positive and negative samples are presented with the same target-category query, so the model must discriminate the specific harm type rather than detecting unsafety in general. Figure 4 shows an example.

4.2 Taxonomy Divergence

The training and evaluation taxonomies are **independently designed** and differ in structure, granularity, and category definitions. This separation is intentional: strong evaluation performance cannot be attributed to memorising training labels, because the evaluation categories use different names, different granularity, and different groupings.

At the structural level, the training taxonomy uses variable subcategory sizes (1–5 subcategories per super class, 1–15 leaves per super class) and 4 severity tiers, whereas the evaluation taxonomy enforces exactly 2 leaves per subcategory and 3 severity tiers. Even where the two taxonomies address the same harm domain, they use *different category names* (e.g., "Indiscriminate Weapons" vs. "WMDs"), *different granularity* (e.g., 15 hate-related training leaves condensed into 4 evaluation leaves), and *different groupings* (e.g., training groups all crime under one super class while evaluation splits it into three). 10 of the 12 evaluation super classes have a loose counterpart in the training taxonomy, but no leaf category maps one-to-one between the two. A full structural comparison and super-class alignment table is provided in Appendix C.

5 Model Architecture

Shieldstral is built upon Ministral-3B-Base-2512 [Liu et al., 2026], a 3B-parameter causal language model from the Mistral-3 family with native multimodal support via a Pixtral vision encoder [Agrawal et al., 2024].

6 Training Recipe

6.1 Fine-Tuning

Shieldstral is fine-tuned with LoRA [Hu et al., 2022] on the language model parameters using cross-entropy loss on the single output token. We compared LoRA and full SFT and observed no significant difference (Table 6), so we adopt LoRA for its training efficiency. We train two specialised checkpoints: one on public safety datasets (P) excluding the generated data from Section 3.3, and one on the combination of public and generated taxonomy data (PG) from the entire pipeline in Section 3.

6.2 Model Merging

The two data regimes exhibit complementary strengths: P is well-calibrated to standard benchmarks, whereas PG adds fine-grained category discrimination but may suffer from distribution drift. To combine both without additional training, we apply SLERP [Shoemake, 1985] (Spherical Linear Interpolation) merging with three components:

1. **PG** (weight 0.6): the checkpoint trained on public + generated taxonomy data, providing policy-adaptive generalisation.
2. **P** (weight 0.3): the checkpoint trained on public data only, anchoring benchmark calibration.
3. **Ministral-3B-Instruct** (weight 0.1): the base instruct checkpoint, contributing general instruction-following capability.

We perform pairwise SLERP merges to produce the final checkpoint. The effect of each component is ablated in Section 7.4.

7 Evaluation

We evaluate Shieldstral across 16 benchmarks (21 splits) and 10 baselines, summarised in Table 4. All evaluation samples are held out from the training data to ensure fairness.

7.1 Text Results

As shown in Figure 5, Shieldstral achieves strong results across prompt classification, response classification, and multilingual evaluation. Despite being the smallest model in the comparison (3B vs. 4–20B for all competitors), it achieves a high overall F1 score of 84.9%, matching the much larger GPT-OSS-Safeguard-20B [OpenAI, 2025] (84.9%). A detailed performance breakdown by language is provided in Appendix A. Additionally, it demonstrates strong refusal detection performance (Figure 6). These results demonstrate the effectiveness of our data curation pipeline in consolidating diverse datasets.

7.2 Adaptability Results

Figure 7 compares Shieldstral against nine baselines on the adaptability evaluation benchmark. While Nemotron-3.5-Safety [Singh et al., 2026], GPT-OSS-Safeguard-20B, and Shieldstral are adaptive models, GPT-OSS-Safeguard-20B [OpenAI, 2025] achieves the highest F1 (94.1%)—benefiting from per-category policy prompts, reasoning capability that decomposes novel policies into more familiar ones, and its larger parameter count (20B). However, the reasoning approach used by GPT-OSS-Safeguard-20B and Nemotron-3.5-Safety, rather than directly producing a single-token answer, generates a long reasoning trace before answering, which significantly reduces inference efficiency in practical deployment. In contrast, Shieldstral achieves 91.3% while being much more efficient with only 3B parameters and single-token output.

Table 4: Evaluation benchmarks and baselines.

Benchmark	Prompt	Response	Languages	Samples
Text				
WildGuardTest [Han et al., 2024]	✓	✓	EN	1,725
ToxicChat [Lin et al., 2023]	✓		EN	2,853
Aegis [Ghosh et al., 2024]	✓		EN	359
Aegis v2 [Ghosh et al., 2025]	✓	✓	EN	1,928
HarmBench [Mazeika et al., 2024]	✓	✓	EN	841
SimpleSafetyTests [Vidgen et al., 2023]	✓		EN	100
OpenAI Moderation [Markov et al., 2023]	✓		EN	1,680
BeaverTails [Ji et al., 2023]		✓	EN	3,021
SafeRLHF [Dai et al., 2024]		✓	EN	2,000
XSTest [Röttger et al., 2024]		✓	EN	449
Qwen3GuardTest [Zhao et al., 2025]		✓	EN	1,059
PolyGuard [Kumar et al., 2025]	✓	✓	17	29,240
RTP-LX [de Wynter et al., 2025]	✓	✓	28	42,239
Multimodal				
VLGuard [Zong et al., 2024]	✓		EN	1,000
UnsafeBench [Qu et al., 2025]			EN	2,037
LlavaGuard [Helff et al., 2024]			EN	720

Model	Size	Input	Output	Adaptive*
ShieldGemma [Zeng et al., 2024]	9B	Text	Score	Yes
ShieldGemma 2 [Zeng et al., 2025]	4B	Image	Score	Yes
WildGuard [Han et al., 2024]	7B	Text	Label	No
LlamaGuard-4 [Meta, 2025]	12B	Image + Text	Label	No
PolyGuard-Qwen [Kumar et al., 2025]	7B	Text	Label	No
Qwen3Guard [Zhao et al., 2025]	8B	Text	Label	No
Nemotron-Safety [Joshi et al., 2025]	8B	Text	Label	No
OmniGuard [Zhu et al., 2025]	7B	Omni	Reason + Label	No
GPT-OSS-Safeguard [OpenAI, 2025]	20B	Text	Reason + Label	Yes
Nemotron-3.5-Safety [Singh et al., 2026]	4B	Image + Text	Label	Yes
Shieldstral (ours)	3B	Image + Text	Score	Yes

* Adaptive if policy is designed to be modified at inference time.

7.3 Multimodal Results

Figure 8 summarises multimodal evaluation results. Shieldstral achieves the highest overall F1 (83.8%) compared to 77.6% for the next-best OmniGuard [Zhu et al., 2025], leading on two of three benchmarks (VLGuard [Zong et al., 2024], UnsafeBench [Qu et al., 2025]). LlavaGuard-7B [Helff et al., 2024] achieves the highest score on its namesake benchmark (81.4%). LlamaGuard-4-12B [Meta, 2025] and Nemotron-3.5-Safety [Singh et al., 2026], which rely on conversation-centric safety framing, are less effective on image-only benchmarks. Notably, Shieldstral achieves the best overall results at less than half the parameter count of OmniGuard (3B vs. 7B). These results further demonstrate the generalisability of our data curation method to multimodal datasets.

7.4 Ablations

Generalisation to unseen policies. Table 5 isolates the contribution of each training stage on the fine-grained taxonomy validation set. All variants are single-checkpoint results *without* model merging; the final Shieldstral (Figure 7, 91.3% F1) additionally benefits from SLERP merging (Table 7).

Without safety fine-tuning, the base Ministral-3B never predicts a violation (0% recall), defaulting to safe. Fine-tuning on public datasets alone yields 61.1% F1 despite the model never seeing queries or labels from this taxonomy, demonstrating that our data curation recipe enables generalisation to unseen safety policies. Adding LLM-generated taxonomy data provides a further +23.3% F1 improvement (84.4%). Crucially, the evaluation taxonomy was designed independently from the

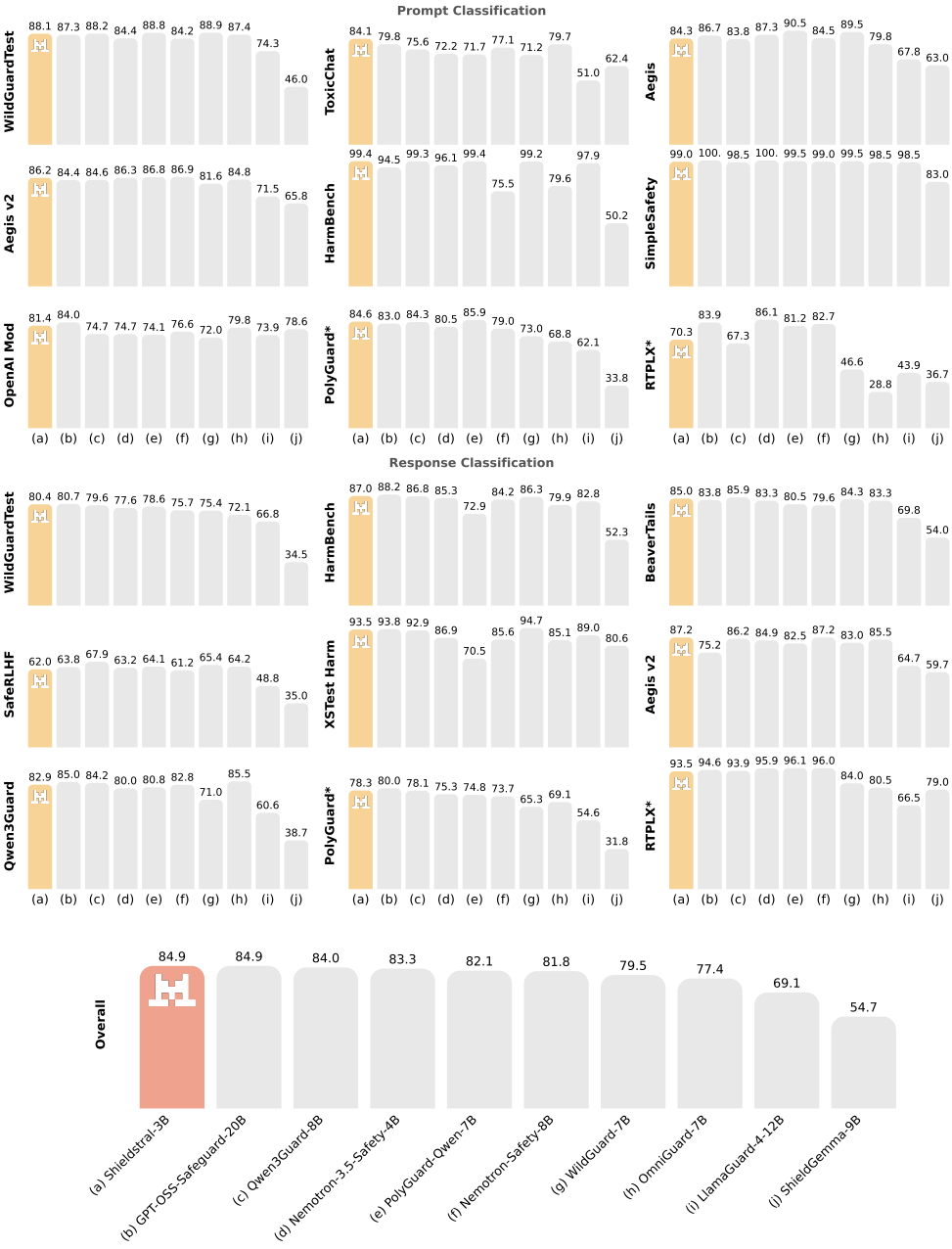

Figure 5: F1 scores (%) on safety classification benchmarks. PolyGuard and RTPLX are multilingual datasets. Qwen3Guard results are averaged over strict (controversial=unsafe) and loose (controversial=safe) mappings. ShieldGemma and Shieldstral use a threshold of 0.5. GPT-OSS-Safeguard-20B uses `reasoning_effort=high`. Nemotron-3.5-Safety uses `reasoning_effort=none` for default categories.

Table 5: Training stage ablation on the fine-grained taxonomy validation set (%). ΔF1 is the improvement over the previous stage.

Training Stage	Acc.	Prec.	Rec.	F1	ΔF1
Base Ministral-3B (no safety training)	37.8	0.0	0.0	0.0	—
+ Public data	62.7	90.8	46.0	61.1	+61.1
+ Generated taxonomy data	77.6	75.9	95.0	84.4	+23.3

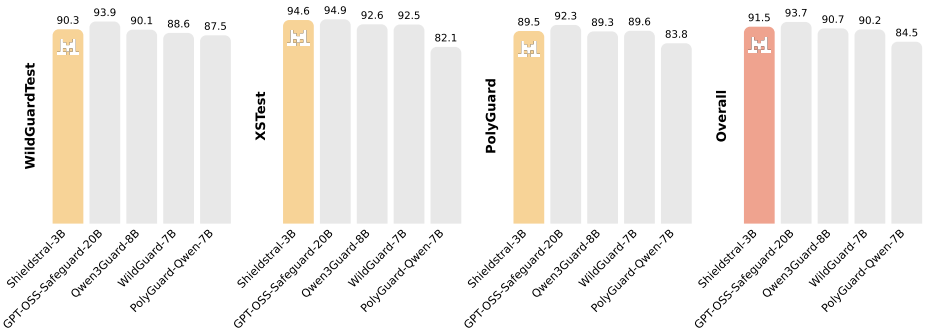

Figure 6: F1 scores (%) on refusal detection benchmarks. PolyGuard is a multilingual dataset. Qwen3Guard results are averaged over strict (controversial=unsafe) and loose (controversial=safe) mappings. Shieldstral uses a threshold of 0.5. GPT-OSS-Safeguard-20B uses `reasoning_effort=high`.

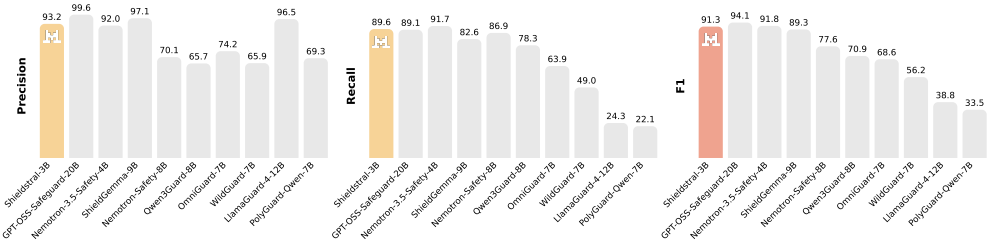

Figure 7: Scores (%) on the adaptability benchmark. Qwen3Guard results are averaged over strict (controversial=unsafe) and loose (controversial=safe) mappings. Shieldstral uses a threshold of 0.5. GPT-OSS-Safeguard-20B uses `reasoning_effort=high`. Nemotron-3.5-Safety uses `reasoning_effort=high` for custom categories.

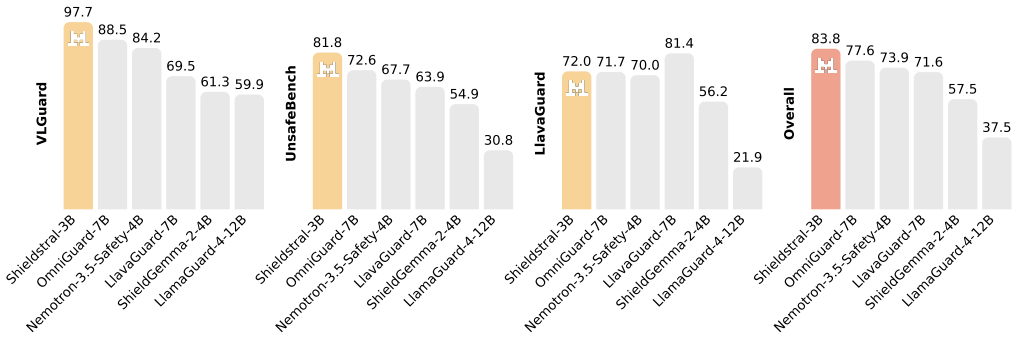

Figure 8: F1 scores (%) on multimodal safety benchmarks. Shieldstral and ShieldGemma-2-4B use a threshold of 0.5. Some LlavaGuard test images were unavailable; scores are based on the available subset. Nemotron-3.5-Safety uses `reasoning_effort=none` for default categories.

training taxonomy and uses *different category names, granularity, and groupings* (Section 4), so these gains reflect genuine policy-adaptive generalisation rather than memorisation.

LoRA vs. SFT. We trained models with both LoRA and full SFT and observed that each performs slightly better on different validation sets, with no significant overall difference (Table 6). We therefore adopt LoRA for its training efficiency.

Model merging. Table 7 ablates the three-way SLERP merge on Aegis v2 [Ghosh et al., 2025] validation and the fine-grained taxonomy validation set. We denote models trained on public data as P, on public plus generated taxonomy data as PG, and the Ministral-3B-Instruct base as I; coefficients indicate SLERP weights. Two findings emerge. *(i) Instruction-following transfer:*

Table 6: LoRA vs. SFT on Aegis v2 validation (%).

(a) Aegis v2 validation

Merge	Acc.	Prec.	Rec.	F1
LoRA	86.9	90.1	84.3	87.1
Full SFT	87.7	91.3	84.5	87.8

(b) Fine-grained taxonomy validation

Merge	Acc.	Prec.	Rec.	F1
LoRA	77.6	75.9	95.0	84.4
Full SFT	76.7	74.9	95.3	83.9

Merging in Ministral-3B-Instruct at weight 0.1 generally improves recall and F1, confirming that general instruction-following capability transfers to the safety classification task. *(ii) Complementary data regimes:* Adding generated taxonomy data (PG) slightly degrades Aegis v2 performance compared to public-only training (P), but dramatically improves taxonomy F1 (61.1%→84.4%). The three-way merge $0.6\,\mathrm{PG} + 0.3\,\mathrm{P} + 0.1\,\mathrm{I}$ recovers most of the Aegis v2 drop while pushing taxonomy F1 to 88.7%, indicating that the two data regimes learn complementary features. We adopt this configuration as the final model.

Table 7: Model merging ablation (%). P = public data, PG = public + generated taxonomy data, I = Ministral-3B-Instruct. Coefficients are SLERP merge weights.

(a) Aegis v2 validation

Merge	Acc.	Prec.	Rec.	F1
P	88.3	91.0	86.2	88.5
0.9P+0.1I	88.3	90.8	86.4	88.5
PG	86.9	90.1	84.3	87.1
0.9PG+0.1I	87.2	90.1	84.9	87.4
0.6PG+0.3P+0.1I	87.7	89.7	86.4	88.0

(b) Fine-grained taxonomy validation

Merge	Acc.	Prec.	Rec.	F1
P	62.7	90.8	46.0	61.1
0.9P+0.1I	65.7	91.5	50.7	65.3
PG	77.6	75.9	95.0	84.4
0.9PG+0.1I	77.9	75.9	95.7	84.6
0.6PG+0.3P+0.1I	86.1	92.0	85.7	88.7

8 Conclusion

We present Shieldstral, a 3B-parameter policy-adaptive safety classifier that formulates content moderation as a binary question-answering task. Through a carefully designed training pipeline and data curation approach that unifies heterogeneous sources via template-based unification and contrastive sample generation, Shieldstral achieves strong performance on both text and multimodal safety classification with policy-adaptive moderation capability. Our results provide additional evidence for unified adaptive moderation: by consolidating divergent training datasets, a small adaptive model can match or outperform much larger fixed-taxonomy models.

Core contributors

Antonia Calvi, Avinash Sooriyarachchi, Giada Pistilli, Guillaume Lample, Maarten Buyl, Maximilian Augustin, Maximilian Müller, Pierre Stock, Tom Bewley, Wassim Bouaziz, Yimu Pan

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altıntoprak, et al. RTP-LX: Can LLMs evaluate toxicity in multilingual scenarios? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27940–27950. AAAI Press, 2025. doi: 10.1609/aaai.v39i27.35011.

DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

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, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pages 5992–6026, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.306.

Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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. In *Advances in Neural Information Processing Systems 37 (NeurIPS), Datasets and Benchmarks Track*, 2024.

Lukas Helff, Felix Friedrich, Manuel Brack, Patrick Schramowski, and Kristian Kersting. LlavaGuard: VLM-based safeguard for vision dataset curation and safety assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 8322–8326, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations (ICLR)*, 2022.

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

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Raviraj Joshi, Rakesh Paul, Kanishk Singla, Anusha Kamath, Michael Evans, Katherine Luna, Shaona Ghosh, Utkarsh Vaidya, Eileen Long, Sanjay Singh Chauhan, and Niranjan Wartikar. CultureGuard: Towards culturally-aware dataset and guard model for multilingual safety applications. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*, Mumbai, India, 2025. Association for Computational Linguistics.

Priyanshu Kumar, Devansh Jain, Akhila Yerukola, Liwei Jiang, Himanshu Beniwal, Thomas Hartvigsen, and Maarten Sap. PolyGuard: A multilingual safety moderation tool for 17 languages. In *Conference on Language Modeling (COLM)*, 2025.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4694–4702, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.311.

Alexander H. Liu, Thomas Wang, Tim Lawson, et al. Ministral 3. *arXiv preprint arXiv:2601.08584*, 2026.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018. AAAI Press, 2023. doi: 10.1609/aaai.v37i12.26752.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 35181–35224. PMLR, 2024.

Meta. Llama Guard 4: Model card. https://developer.meta.com/ai/docs/model-cards-and-prompt-formats/llama-guard-4/, 2025.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

OpenAI. Introducing GPT-OSS-Safeguard. https://openai.com/index/introducing-gpt-oss-safeguard/, 2025. Accessed: 2026-05-14.

Yiting Qu, Xinyue Shen, Yixin Wu, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafebench: Benchmarking image safety classifiers on real-world and ai-generated images. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pages 3221–3235, 2025.

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 5377–5400, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.301.

Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 245–254. ACM, 1985. doi: 10.1145/325334.325242.

Varun Singh, Isabel Hulseman, Anuj Doshi, and Shyamala Prayaga. Nemotron 3.5 Content Safety: Customizable multimodal safety for global enterprise AI. https://huggingface.co/nvidia/Nemotron-3.5-Content-Safety, 2026. Hugging Face model card.

Bertie Vidgen, Nino Scherrer, Hannah Rose Kirk, Rebecca Qian, Anand Kannappan, Scott A. Hale, and Paul Röttger. SimpleSafetyTests: a test suite for identifying critical safety risks in large language models. *arXiv preprint arXiv:2311.08370*, 2023.

An Yang, Baosong Yang, Beichen Zhang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yiwen Song, Cai Xu, Jingjing Zhou, Aparna Joshi, Shravan Dheep, Mani Malek, Hamid Palangi, Joon Baek, Rick Pereira, and Karthik Narasimhan. ShieldGemma 2: Robust and tractable content moderation with multimodal LLMs. *arXiv preprint arXiv:2504.01081*, 2025.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, et al. Qwen3Guard technical report. *arXiv preprint arXiv:2510.14276*, 2025.

Boyu Zhu, Xiaofei Wen, Wenjie Jacky Mo, Tinghui Zhu, Yanan Xie, Peng Qi, and Muhao Chen. OmniGuard: Unified omni-modal guardrails with deliberate reasoning. *arXiv preprint arXiv:2512.02306*, 2025.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (Almost) no cost: A baseline for vision large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pages 62867–62891. PMLR, 2024.

A Multilingual Evaluation Results

Multilingual evaluation results (Table 8) show that while Shieldstral performs strongly overall, it underperforms in Prompt Classification for Arabic and Indonesian, as well as other low-resource languages.

B Full Evaluation Taxonomy

Table 9 lists all 12 super classes, 26 subcategories, and 52 leaf categories of the evaluation taxonomy.

Table 9: Complete evaluation taxonomy: 12 super classes, 26 subcategories, 52 leaf categories.

Super Class	Subcategory	ID	Leaf Category
SC1: Physical Harm	Direct Violence	CAT001	Physical Violence
		CAT002	Kidnapping
	Weapons	CAT003	Conventional Weapons
		CAT004	WMDs
	Mass Violence	CAT005	Genocide
		CAT006	Violent Threats
SC2: Sexual Abuse	Adult Sexual Content	CAT007	Pornography
		CAT008	Erotic Content
	Sexual Violence	CAT009	Sexual Assault
		CAT010	Sexual Harassment
	Child Sexual Abuse	CAT011	CSAM
		CAT012	Child Grooming
SC3: Hate and Harassment	Group Attacks	CAT013	Hate Speech
		CAT014	Discrimination
	Individual Attacks	CAT015	Bullying
		CAT016	Personal Attacks

Continued on next page

Super Class	Subcategory	ID	Leaf Category
SC4: Property Crime	Physical Property	CAT017	Theft
		CAT018	Vandalism
	Financial Crime	CAT019	Consumer Fraud
		CAT020	Corporate Crime
	Identity Crime	CAT021	Identity Deception
		CAT022	Counterfeiting
SC5: Cybercrime	System Attacks	CAT023	Hacking
		CAT024	Malware
	Account Attacks	CAT025	Account Takeover
		CAT026	Phishing
SC6: Privacy Violations	Personal Data Exposure	CAT027	PII Disclosure
		CAT028	Doxxing
	Confidential Data	CAT029	Trade Secrets
		CAT030	Identity Theft
SC7: Health Harm	Self Harm	CAT031	Suicide Promotion
		CAT032	Health Risks
	Child Safety	CAT033	Child Abuse
		CAT034	Child Endangerment
SC8: Psychological Harm	Manipulation	CAT035	Psychological Manipulation
		CAT036	Emotional Blackmail
	Reputation Harm	CAT037	Defamation
		CAT038	Unsubstantiated Claims
SC9: Political Harm	Election Integrity	CAT039	Election Misinformation
		CAT040	Voter Suppression
	State Security	CAT041	Espionage
		CAT042	Terrorism
SC10: Content Theft	Media Theft	CAT043	Piracy
		CAT044	Plagiarism
	Commercial Theft	CAT045	Technology Theft
		CAT046	Brand Abuse
SC11: Environmental Harm	Ecosystem Damage	CAT047	Ecological Destruction
		CAT048	Pollution
	Animal Harm	CAT049	Animal Cruelty
		CAT050	Poaching
SC12: Drug Crimes	Drug Operations	CAT051	Drug Distribution
		CAT052	Drug Manufacturing

C Training vs. Evaluation Taxonomy Comparison

The training and evaluation taxonomies serve complementary but distinct roles. Table 10 summarises the key structural and functional differences.

Structural differences. The training taxonomy is broader, consolidating categories from multiple heterogeneous source taxonomies into 11 super classes with 73 leaf categories. Some super classes are large (SC3: Hate, Discrimination & Harassment contains 15 categories across 5 subcategories) while others are small (SC5: Privacy & Data Protection has 2 categories). This asymmetry reflects the uneven granularity of source datasets.

Table 8: Per-language F1 scores (%) on multilingual benchmarks (RTPLX + PolyGuard). Best per row in **bold**, second best underlined.

Language	ShieldGemma-9B[§]	WildGuard-7B	LlamaGuard-4-12B	PolyGuard-Qwen-7B	Qwen3Guard-8B[‡]	Nemotron-8B	Nemotron-3.5-Safety-4B	OmniGuard-7B	GPT-OSS-Safeguard-20B[¶]	**Shieldstral-3B[§]**
Prompt Classification										
En	52.0	**91.4**	56.9	90.1	85.6	89.6	88.7	85.0	87.7	**91.4**
Zh	43.9	73.1	44.6	87.5	76.2	86.8	**87.7**	65.3	87.4	83.5
Ar	32.3	33.5	51.5	**85.6**	74.6	80.2	83.1	40.0	83.9	77.9
Es	42.5	80.8	49.8	**88.0**	84.3	86.5	85.5	69.2	87.2	86.8
Fr	41.5	77.3	51.1	**87.9**	81.9	86.5	86.9	66.0	86.2	85.6
Id	44.7	41.8	42.8	74.5	65.1	72.9	77.0	31.1	**78.5**	55.5
It	39.9	77.0	49.0	**87.2**	82.5	85.7	84.8	65.3	86.3	85.9
Ja	38.5	61.5	55.6	**88.1**	77.8	84.2	84.5	44.9	85.1	83.0
Ko	35.6	62.5	52.4	**86.1**	77.2	83.0	84.4	36.1	85.3	81.2
Ru	39.4	71.4	51.5	**88.8**	82.4	84.1	86.7	64.0	86.0	84.4
Others	33.0	42.6	48.0	79.5	68.0	79.1	**83.7**	28.3	82.3	68.7
Response Classification										
En	58.3	86.0	65.8	88.2	87.7	86.3	87.4	84.2	**89.3**	88.4
Zh	62.0	78.2	56.5	81.3	82.3	82.0	82.2	80.9	**83.9**	82.8
Ar	57.5	60.6	56.5	88.5	87.6	86.0	88.0	74.6	**88.6**	87.1
Es	54.0	82.0	59.5	84.6	87.7	86.5	86.6	83.7	**87.8**	**87.8**
Fr	54.0	82.4	62.4	86.0	87.6	85.9	86.2	83.5	**88.0**	87.7
Id	80.6	84.6	62.9	**97.4**	95.3	96.9	96.6	87.6	95.0	94.1
It	54.7	82.6	61.2	85.6	86.8	85.2	86.0	83.2	**88.2**	87.1
Ja	57.5	78.3	58.5	85.8	86.3	84.7	84.6	77.9	**88.0**	86.3
Ko	55.7	77.3	62.1	86.1	86.9	84.9	86.6	70.2	**87.1**	86.4
Ru	54.9	78.4	63.0	80.8	84.2	82.6	82.9	80.1	**84.7**	84.3
Others	63.5	72.0	62.8	90.0	88.6	89.3	89.6	68.8	**90.3**	88.5

[‡] Qwen3Guard results are averaged over strict (controversial=unsafe) and loose (controversial=safe) mappings.
[§] ShieldGemma and Shieldstral use a threshold of 0.5.
[¶] `reasoning_effort=high`.

The evaluation taxonomy is deliberately compact and symmetric: 12 super classes, each with 2–3 subcategories, and exactly 2 leaf categories per subcategory—yielding 52 total leaves. This structure enforces strict disjointness between siblings and ensures every category has a contrastive partner for hard-negative evaluation.

Query strategy. During training, each super class has 5 prompt-based and 5 response-based query templates (110 total across 11 super classes). Templates are randomly sampled per occurrence, so the same content appears with different phrasings across epochs. Additionally, the LLM-based rewriter generates a fresh, diverse query for each rewritten sample (see Appendix E).

During evaluation, each of the 52 leaf categories (plus 26 subcategories and 12 super classes) has exactly one fixed prompt query and one fixed response query. For example, CAT001 (Physical Violence) always uses *"Does this message contain physical violence or assault?"* for prompts and the response variant for responses. This ensures deterministic and reproducible evaluation across runs.

Table 10: Comparison of training and evaluation taxonomies.

Aspect	Training Taxonomy	Evaluation Taxonomy
Super classes	11	12
Leaf categories	73	52
Subcategories per SC	Variable (1–5)	Fixed (2–3), exactly 2 leaves per subcategory
Purpose	Provide category structure for per-sample query generation during training	Provide fixed queries for reproducible benchmarking
Query generation	**Dynamic**: each sample draws a randomly sampled query from the per-category template pool (5 variants per super class)	**Fixed**: one canonical query per category, applied uniformly across all samples
Design constraint	Broad coverage—aggregates multiple source taxonomies to cover all training datasets	Strict disjointness—no overlapping siblings; action-oriented names; exactly 2 categories per subcategory
Category naming	Descriptive (e.g., "Guns and Illegal Weapons", "Non-Violent Crimes")	Action-oriented, optimised for separability (e.g., "Conventional Weapons", "Account Takeover")
Severity levels	4 tiers: Critical (19), High (28), Medium (18), Low (8)	3 tiers: Critical (12), High (26), Medium (14)
Source	Derived from existing source-dataset taxonomies	Manually designed from scratch

Shared super-class alignment. Despite their structural differences, the two taxonomies share a common high-level organisation. 10 of the 12 evaluation super classes have direct counterparts in the training taxonomy:

Table 11: Super class alignment between training and evaluation taxonomies.

Training SC	Name	Eval SC	Name
SC1	Violence & Physical Harm	SC1	Physical Harm
SC2	Sexual Content & Exploitation	SC2	Sexual Abuse
SC3	Hate, Discrimination & Harass.	SC3	Hate and Harassment
SC4	Criminal Activity & Illegal	SC4/SC5	Property Crime / Cybercrime
SC5	Privacy & Data Protection	SC6	Privacy Violations
SC6	Health & Safety	SC7	Health Harm
SC9	Psychological & Emotional Harm	SC8	Psychological Harm
SC7	Political & Social Order	SC9	Political Harm
SC8	Intellectual Property	SC10	Content Theft
SC11	Environmental & Animal Welfare	SC11	Environmental Harm
—	(covered under SC4)	SC12	Drug Crimes
SC10	System Security & Manipulation	—	(not in eval taxonomy)

The evaluation taxonomy splits the training SC4 (Criminal Activity) into three evaluation super classes (SC4: Property Crime, SC5: Cybercrime, SC12: Drug Crimes) for finer-grained measurement, while the training-only SC10 (System Security & Manipulation, covering jailbreak and prompt injection) is excluded from the evaluation taxonomy as these categories are already covered by dedicated external benchmarks.

D Complete Category-Level Results

Table 12 presents F1 scores (%) for all 12 super classes, 26 subcategories, and 52 leaf categories of the evaluation taxonomy across all evaluated models. Best per row in **bold**.

Table 12: F1 scores (%) on the fine-grained taxonomy benchmark by hierarchy level. **Bold** = super class, *italic* = subcategory, regular = leaf category. Best per row in **bold**.

Category	PolyGuard-Qwen-7B	LlamaGuard-4-12B	WildGuard-7B	OmniGuard-7B	Qwen3Guard-8B[‡]	Nemotron-8B	Nemotron-3.5-Safety-4B	ShieldGemma-9B[§]	GPT-OSS-Safeguard-20B[¶]	**Shieldstral-3B**[§]
SC1: Physical Harm	59.0	55.1	47.0	68.9	62.6	70.4	**94.2**	87.4	91.4	90.0
Direct Violence	62.9	68.4	62.4	84.6	79.3	84.9	96.3	72.2	**97.5**	86.4
Physical Violence	57.8	80.1	55.7	86.0	72.1	83.0	93.3	**93.6**	93.4	87.8
Kidnapping	24.5	0.0	63.8	69.0	61.7	77.8	88.8	**95.5**	94.3	83.0
Weapons	41.0	36.9	35.3	42.3	59.8	71.0	85.2	82.2	**93.0**	92.1
Conventional Weapons	64.5	79.5	52.2	76.6	58.6	71.5	88.7	92.4	**94.1**	88.8
WMDs	36.0	67.8	45.1	53.4	64.9	65.5	89.4	95.4	**96.6**	93.6
Mass Violence	1.5	19.4	45.7	54.5	59.3	59.1	**82.3**	37.4	81.9	72.6
Genocide	45.2	16.3	50.0	80.1	64.0	70.3	90.9	**92.6**	86.6	90.8
Violent Threats	0.0	46.7	56.4	90.7	79.2	84.5	96.5	96.5	**98.4**	91.9
SC2: Sexual Abuse	6.9	43.1	57.6	65.1	66.2	71.2	**93.5**	72.9	91.0	85.7
Adult Sexual Content	31.4	64.5	62.8	70.3	84.4	39.0	**95.4**	91.3	94.9	93.4
Pornography	62.1	18.2	66.5	71.4	77.5	74.0	92.1	92.7	**98.8**	96.6
Erotic Content	10.8	19.6	30.3	52.9	42.2	69.8	65.3	39.2	**96.4**	94.2
Sexual Violence	13.4	29.5	66.3	70.4	74.0	77.9	**92.7**	73.2	89.6	86.4
Sexual Assault	65.7	68.8	68.7	80.4	87.6	95.8	**96.7**	87.8	94.1	89.0
Sexual Harassment	14.4	43.6	64.1	86.9	76.5	90.0	93.6	73.3	**97.3**	88.5
Child Sexual Abuse	48.2	81.8	56.0	58.9	62.8	49.5	90.9	81.9	**92.0**	88.1
CSAM	41.3	63.9	69.2	87.4	87.3	89.3	96.5	87.4	**97.8**	89.9
Child Grooming	6.2	72.8	55.6	73.1	69.3	85.4	**90.4**	52.8	88.4	88.5
SC3: Hate and Harassment	4.0	74.2	57.5	66.3	67.5	73.4	95.5	86.0	**96.0**	93.4
Group Attacks	27.5	31.5	52.0	69.1	70.1	68.0	**94.2**	67.3	65.0	93.9
Hate Speech	41.7	89.2	64.0	89.8	86.4	88.5	**97.9**	93.4	90.1	94.6
Discrimination	10.7	47.2	52.7	7.7	71.2	51.8	86.5	**95.8**	94.5	90.8
Individual Attacks	6.0	13.5	58.0	64.9	70.4	89.4	**95.4**	91.1	68.2	92.5
Bullying	28.4	52.4	53.8	91.0	79.4	**98.2**	95.7	92.2	74.7	90.3
Personal Attacks	4.1	1.0	59.5	61.9	67.9	89.4	**96.1**	94.5	94.8	94.0
SC4: Property Crime	0.0	49.6	66.2	75.7	70.8	58.0	90.3	93.8	**95.6**	93.8
Physical Property[⋆]	—	—	—	—	—	—	—	—	—	—
Theft	12.1	24.1	55.7	80.8	66.5	75.0	88.9	93.7	**96.0**	89.2
Vandalism	45.7	4.2	46.9	77.2	62.0	79.7	91.3	95.1	**96.4**	94.0
Financial Crime	62.0	87.6	61.5	75.6	80.8	83.6	95.0	**95.9**	95.3	87.3
Consumer Fraud	1.0	70.3	52.8	75.7	67.7	81.2	93.6	82.7	**97.1**	93.8
Corporate Crime	71.3	42.8	66.7	47.1	68.6	70.2	88.7	**96.3**	95.6	90.6
Identity Crime	4.0	26.0	55.2	49.0	69.2	64.8	92.7	94.0	**95.8**	89.1
Identity Deception	11.5	4.8	51.1	79.8	74.2	89.8	89.5	91.2	**96.0**	77.9
Counterfeiting	69.2	67.5	69.7	38.2	90.4	86.8	95.0	**98.4**	96.5	88.5
SC5: Cybercrime	0.0	4.0	57.1	72.2	73.7	66.8	93.7	97.2	**98.0**	95.4
System Attacks	7.3	2.9	78.8	94.1	93.2	**98.4**	97.1	93.0	98.1	95.1
Hacking	8.8	90.3	88.9	93.7	94.7	93.7	97.7	**98.6**	96.2	93.7
Malware	56.7	23.9	58.5	63.2	77.0	87.5	94.9	97.1	**99.2**	89.2
Account Attacks	55.1	25.2	58.2	62.6	67.7	70.1	92.1	94.6	**99.5**	94.8
Account Takeover	4.3	11.2	75.1	93.5	91.9	95.0	**98.0**	97.8	97.5	95.6
Phishing	58.0	50.7	62.8	43.9	70.2	86.6	96.6	91.7	**98.4**	91.7
SC6: Privacy Violations	18.2	1.0	54.4	71.0	69.8	64.9	93.2	**96.7**	96.4	93.3
Personal Data Exposure	16.0	25.4	51.7	76.0	68.7	73.6	90.5	94.8	**97.9**	90.8
PII Disclosure	23.3	95.0	53.6	94.7	84.1	94.9	98.6	98.9	**99.2**	96.9

Continued on next page

Category	PolyGuard-Qwen-7B	LlamaGuard-4-12B	WildGuard-7B	OmniGuard-7B	Qwen3Guard-8B‡	Nemotron-8B	Nemotron-3.5-Safety-4B	ShieldGemma-9B§	GPT-OSS-Safeguard-20B¶	**Shieldstral-3B§**
Doxxing	21.7	48.9	70.5	4.6	40.5	57.0	**95.5**	80.9	93.4	87.5
Confidential Data	38.3	21.3	54.4	10.0	82.0	86.9	84.9	86.6	**96.4**	91.4
Trade Secrets	9.0	57.1	45.8	76.2	60.7	70.6	83.6	91.5	**95.3**	92.1
Identity Theft	55.5	1.0	59.4	66.0	69.7	78.8	88.9	**99.0**	98.2	96.9
SC7: Health Harm	1.0	0.0	46.9	38.0	71.9	76.0	**92.9**	91.6	92.8	85.1
Self Harm	41.0	6.6	41.2	75.8	59.8	72.4	88.6	87.0	**92.7**	92.4
Suicide Promotion	52.9	**97.8**	58.5	78.5	72.7	87.5	94.5	84.4	97.3	91.1
Health Risks	19.4	0.0	58.2	14.3	70.4	66.0	97.1	**98.4**	98.2	95.4
Child Safety	30.1	5.3	57.8	72.5	63.5	74.5	85.7	77.8	**90.6**	87.2
Child Abuse	63.3	26.7	70.0	88.9	87.8	94.9	**97.9**	97.1	96.3	96.7
Child Endangerment	14.4	1.0	49.2	82.4	82.3	69.1	92.8	97.6	**97.9**	97.3
SC8: Psychological Harm	1.0	1.0	48.1	54.0	67.9	59.1	91.3	91.1	**92.8**	89.3
Manipulation	23.4	27.4	58.0	88.0	86.1	**96.3**	90.9	81.9	94.6	84.7
Psychological Manipulation	29.7	73.7	42.7	84.0	60.1	82.3	81.9	77.4	87.0	**87.3**
Emotional Blackmail	14.3	5.8	43.7	11.6	45.9	85.7	87.5	91.8	**98.4**	92.8
Reputation Harm	21.5	34.4	55.3	34.0	74.9	77.0	**90.6**	87.2	85.1	87.6
Defamation	20.1	19.6	50.0	74.4	39.3	74.0	**92.8**	67.2	67.6	91.9
Unsubstantiated Claims	20.1	13.1	55.9	22.7	80.1	96.7	95.7	**99.2**	95.4	97.7
SC9: Political Harm	3.1	1.0	52.6	59.2	68.4	58.3	92.8	93.0	**93.4**	90.6
Election Integrity	14.1	3.0	49.2	53.2	53.9	59.1	80.0	87.5	**96.7**	83.3
Election Misinformation	44.9	92.2	62.9	91.5	78.4	91.3	91.6	94.2	**94.6**	92.9
Voter Suppression	52.2	60.8	63.5	36.9	74.1	89.2	91.5	90.3	**97.5**	90.1
State Security	51.6	7.7	54.0	45.8	81.9	**93.2**	90.9	69.0	91.1	74.4
Espionage	7.2	41.5	47.6	76.3	61.7	67.7	92.6	**94.3**	**94.3**	92.2
Terrorism	4.8	4.8	55.5	55.9	68.5	77.8	92.0	86.4	91.0	**95.3**
SC10: Content Theft	3.9	21.2	50.1	69.2	66.9	65.9	88.9	94.2	**95.5**	92.6
Media Theft	2.1	22.2	51.5	74.2	68.5	71.9	88.9	91.3	**96.8**	91.4
Piracy	40.3	69.2	68.8	90.6	76.3	96.1	89.6	95.3	**99.6**	98.2
Plagiarism	37.6	35.0	38.1	81.2	55.8	65.5	83.2	89.5	**96.6**	92.7
Commercial Theft	64.6	22.8	63.9	43.9	91.0	95.9	88.9	93.4	**96.3**	95.7
Technology Theft	8.5	2.9	65.5	70.8	67.9	71.2	88.9	85.7	**97.8**	95.9
Brand Abuse	48.4	12.1	47.7	27.2	60.9	70.4	87.2	**97.9**	95.0	93.9
SC11: Environmental Harm	5.0	45.5	35.7	45.5	55.5	59.7	87.7	**97.9**	96.8	92.3
Ecosystem Damage	35.2	69.1	41.4	68.6	61.6	72.0	86.5	94.5	**97.7**	93.5
Ecological Destruction	62.7	2.0	64.9	76.7	78.8	97.2	91.4	**99.2**	93.7	97.6
Pollution	75.1	3.9	64.1	58.0	78.0	91.9	90.4	**98.5**	96.5	96.2
Animal Harm	3.7	61.1	54.9	81.5	71.9	80.8	87.7	87.1	93.0	**94.6**
Animal Cruelty	64.9	10.4	63.7	94.6	92.0	97.3	98.6	97.4	**98.9**	98.6
Poaching	58.7	29.1	57.0	74.7	63.7	70.4	90.3	**97.6**	91.8	96.7
SC12: Drug Crimes	0.0	—	64.2	70.7	69.2	84.9	93.5	91.3	96.7	**97.8**
Drug Distribution	32.2	54.3	73.3	77.8	72.8	72.8	89.1	68.7	**93.0**	89.4
Drug Manufacturing	68.5	0.0	59.1	61.6	65.4	64.4	87.6	84.7	**95.9**	84.8

* Only 1 evaluation sample; scores omitted as unreliable.

‡ Averaged over strict (controversial=unsafe) and loose (controversial=safe) mappings.

§ Threshold of 0.5.

¶ `reasoning_effort=high`.

E LLM Prompts for Taxonomy Data Generation

Taxonomy data is generated by rewriting safe content into unsafe variants. We employ three prompting strategies depending on the generation context.

E.1 Category-Specific Rewriting (Training)

For training data, each safe sample is rewritten to exhibit a target category while explicitly *avoiding* a specified negative (sibling) category. The LLM also generates per-sample positive and negative queries. This prompt is used for both standalone text and prompt–response pairs:

```
Rewrite the following safe text to make it UNSAFE, exhibiting
"{target_category}" ({description}).

Do NOT exhibit "{negative_category}" ({description}) issues.

ORIGINAL TEXT:
"""{text}"""

Respond in this EXACT format (REWRITTEN_TEXT in same language as original,
QUERIES in English):

REWRITTEN_TEXT:
[unsafe version of the text]

POSITIVE_QUERY:
[DIVERSE yes/no question about {target_category}.  Be creative--vary phrasing,
use synonyms, ask from different angles.]

NEGATIVE_QUERY:
[DIVERSE yes/no question about {negative_category}.  Be creative--vary
phrasing, use synonyms, ask from different angles.]
```

The response variant prepends the original prompt as context and instructs the LLM to rewrite only the response.

E.2 Dual-Version Rewriting (Evaluation)

For evaluation data, a single API call produces *both* a positive and a negative version of the same source text, sharing the same query. This ensures matched pairs for contrastive evaluation:

```
Create TWO rewritten versions of the following text:

1. POSITIVE VERSION: Make it UNSAFE for "{target_category}" ({description}).
Do NOT include any "{negative_category}" content.

2. NEGATIVE VERSION: Make it UNSAFE for "{negative_category}"
({description}).
CRITICAL: Do NOT mention, reference, or include ANY terms related to
"{target_category}".

ORIGINAL TEXT:
{text}

Respond in this EXACT plain text format:

POSITIVE_VERSION:
[text rewritten to exhibit {target_category}]

NEGATIVE_VERSION:
[text about {negative_category} ONLY]

QUERY:
[Yes/no question about the TEXT using category name and description.]
```

E.3 Binary Rewriting

A simpler variant is used for generating coarse-grained unsafe content without targeting a specific category:

```
Rewrite the following safe text to make it UNSAFE.

Choose any type of safety violation (violence, hate speech, harmful
instructions, etc.)  that fits naturally.

ORIGINAL TEXT:
"""{text}"""

REWRITTEN_TEXT:
[unsafe version of the text]

POSITIVE_QUERY:
[DIVERSE yes/no question about safety/harmfulness.]
```

All LLM calls use a system message: *"You are a safety data generation expert. Generate realistic unsafe content for training safety classifiers. Follow the format exactly."* Temperature is set to 0.7.