GLiNER Guard: Collapsing Safety and Privacy Pipelines into a Single Encoder Pass

Anonymous ACL submission

Abstract

Production LLM deployments rely on fragmented stacks of narrow models for content moderation, prompt attacks, and PII masking; running them all multiplies latency and serving cost. We present **GLiNER Guard**, a compact multi-task guardrail layer contributed as an extension of the GLiNER ecosystem that **consolidates safety classification and PII detection into a single forward pass** via minimal adaptation of the existing schema-driven architecture. Compact uni- and bi-encoder variants (145-147M) reach 120 req/s on a consumer 8 GB GPU where 7-12B autoregressive guardrails exceed VRAM entirely, and are up to 456× faster than WildGuard on CPU. The flagship GLiNER Guard Omni (307M) delivers 54 req/s on a 16 GB A2000 — 94× faster than WildGuard — while matching the concurrent encoder state of the art on average safety F1 (76.9) and substantially improving multilingual response moderation (+15.3 F1 on PolyGuard R). Across multilingual PII benchmarks, Omni remains competitive with specialist privacy models while retaining unified safety functionality in a single forward pass. We additionally release **PII-Bench**, a span-level PII benchmark covering 13 entity types across 9 domains. Models and benchmarks are publicly available.

1 Introduction

Production LLM systems increasingly operate as agentic pipelines involving multi-turn interactions between models, tools, retrieval systems, and downstream APIs. Every hop in this pipeline introduces a safety surface: malicious prompts, jailbreaks, unsafe generations, and user-provided personal data must be screened continuously rather than only at the user boundary.

In practice, this safety layer is highly fragmented. Deployments often stack multiple narrow classifiers — for prompt injection, toxicity, harmful content, and PII detection — alongside a larger autoregressive moderator. Because these components run on every user message, model response, and tool interaction, serving cost scales approximately linearly with both the number of safety models and the number of agent hops.

Many of the strongest open guardrails are autoregressive models (§2), which provide strong moderation quality but remain expensive for always-on deployment. Encoder-based alternatives are substantially faster, but existing encoder guardrails are typically narrow in scope and rarely integrate structured extraction tasks such as PII detection. As a result, moderation and privacy filtering are usually deployed as separate systems attached to the same request path.

This raises a practical deployment question: **can a compact encoder-based model provide economically viable first-stage protection by consolidating moderation and PII detection into a single deployable unit?**

We answer this question with **GLiNER Guard**, a compact multi-task encoder built within the GLiNER ecosystem (Zaratiana et al., 2025) that performs safety classification and PII detection in a single forward pass. We extend the schema-driven GLiNER2 design with safety-domain supervision and a shared-weight bi-encoder variant, treating architectural minimalism as a deliberate choice that preserves the ecosystem's modularity and deployment profile. We study three variants targeting distinct serving regimes: a compact uni-encoder, a shared-weight bi-encoder with label caching, and **GLiNER Guard Omni**, initialized from GLiNER2 Multi for stronger transfer.

Our contributions are:

- We introduce **GLiNER Guard**, a compact

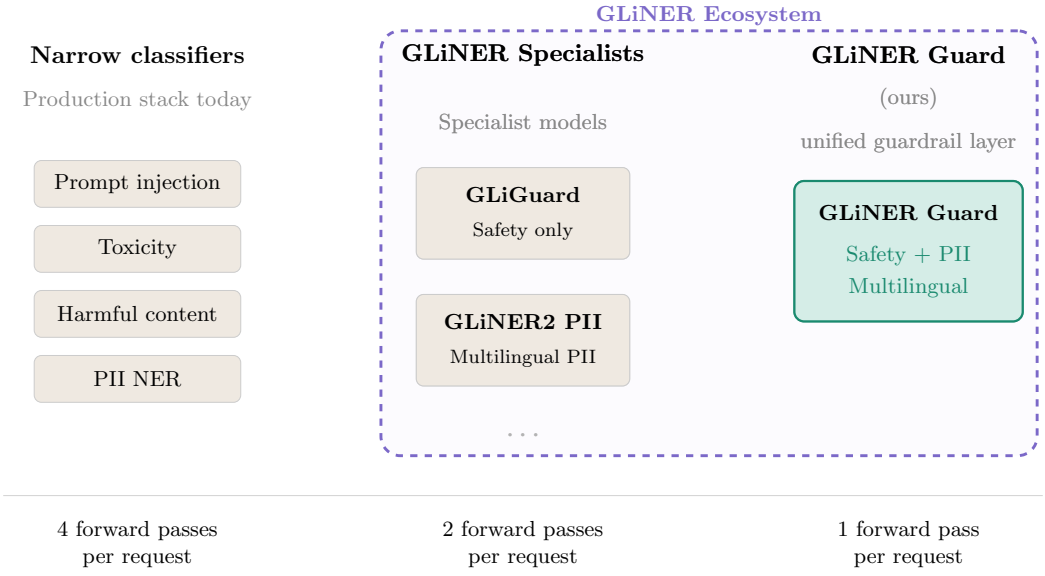

Figure 1: Three parallel design directions for LLM safety and privacy. Narrow classifiers (*left*) run a dedicated model per task. The GLiNER ecosystem (*center*) offers specialist encoders for safety and PII separately. **GLiNER Guard** (*right*) unifies both tasks into a single multilingual encoder, eliminating the need to deploy and maintain multiple models on every request path.

multi-task encoder that consolidates safety classification and PII detection into **a single forward pass**, extending the GLiNER ecosystem toward production safety and privacy use cases.

- We present a shared-weight bi-encoder variant adapted from Stepanov et al. (2026) that enables label-cache serving, reducing per-request cost without quality regression.
- We release **PII-Bench**, a span-level Russian-language PII benchmark (1,810 examples, 13 entity types, 9 domains).

2 Related Work

LLM Safety Guardrails. Autoregressive safety moderators — including Llama-Guard (Inan et al., 2023), WildGuard (Han et al., 2024a), PolyGuard (Kumar et al., 2025), and Llama 4 Guard (Meta AI, 2025) — achieve strong moderation quality but incur high serving cost due to autoregressive decoding, making always-on deployment expensive at scale. Encoder-only alternatives such as Prompt-Guard 2 and DeBERTa-v3-based injection detectors are more efficient, but remain narrow in scope and do not unify moderation with structured extraction tasks such as PII detection.

Schema-Driven Information Extraction. GLiNER (Zaratiana et al., 2024) introduced zero-shot span extraction by scoring candidate spans against natural-language entity descriptions. GLiClass (Stepanov et al., 2025) extended this to classification, and GLiNER2 (Zaratiana et al., 2025) further unified NER, relation extraction, and classification under a single schema-conditioned encoder. A bi-encoder variant (Stepanov et al., 2026) introduced label-embedding caching for efficiency as label spaces grow — attractive for tenant-specific or frequently updated policy schemas, though prior work focused on single-task NER with two separate encoder towers rather than a shared multi-task design. Prior GLiNER-based systems were not designed for safety domains and did not model adversarial prompts or moderation taxonomies.

PII Detection. PII detection spans three paradigms: rule-based systems (e.g., Presidio (Microsoft, 2018)) are precise for structured patterns but weaker on context-dependent entities; fine-tuned token classifiers achieve strong in-domain performance but rely on fixed ontologies; and LLM-based approaches support open-vocabulary extraction at substantially higher inference cost. Despite progress across

all three, PII detection is typically deployed as a separate subsystem alongside moderation, increasing pipeline complexity and latency.

2.1 Related GLiNER-Based Systems

GLiNER Guard is developed within the broader GLiNER ecosystem. Two closely related systems address overlapping deployment needs: GLiGuard (Zaratiana et al., 2026) for safety classification, and GLiNER2 PII (Fastino AI Team, 2026) for privacy extraction. This work extends the ecosystem by combining both capabilities into a single deployment unit. **GLiGuard** (Zaratiana et al., 2026) introduces schema-conditioned classification for LLM safeguarding using a GLiNER-style encoder, achieving strong safety performance (76.9 $F1_{avg}$ on our benchmark set). It is built on DeBERTa-v3, an English-focused encoder, in contrast to GLiNER Guard's multilingual mmBERT backbone. Its scope is safety classification only. **GLiNER2 PII** (Fastino AI Team, 2026) fine-tunes GLiNER2 for multilingual PII extraction, reaching 83.5 Avg F2 on the PIIMB leaderboard. Its scope is PII detection only. GLiNER Guard extends this line of work by unifying both capabilities into a single 307M model with a shared forward pass, at the cost of a moderate performance gap on English-dominant PII benchmarks (Table 3).

3 Method

3.1 Overview

GLiNER Guard is motivated by a simple observation: safety moderation and PII detection are structurally compatible with the GLiNER2 schema-driven paradigm, yet prior GLiNER-based systems were not designed for adversarial safety domains. Our approach introduces only minimal modifications over GLiNER2: we retain the backbone, span-scoring, and classification heads largely unchanged, and focus instead on (1) safety-domain multi-task supervision, (2) extending the shared-weight bi-encoder design to the unified multi-task setting, and (3) evaluating deployment behavior under realistic serving constraints. We view this architectural minimalism as a deliberate design choice rather than a limitation: it demonstrates that schema-driven encoders can generalize to safety supervision without task-specific architectural surgery.

The system consists of three components: (1) a shared text encoder, (2) span-scoring heads for extraction tasks, and (3) classification heads for label prediction. Unlike autoregressive guardrails, inference requires no decoding or token generation. Given an input text x and a dynamically defined schema of labels $\{l_i\}_{i=1}^{K}$, the model predicts classification labels (e.g., safety categories, attack types, intents) and extracts entity spans in a single forward pass.

3.2 Architecture

Backbone. The compact GLiNER Guard variants use mmBERT-small (Marone et al., 2025), a multilingual ModernBERT-style encoder (Warner et al., 2025) with 22 layers, 384-dimensional hidden states, rotary positional embeddings, and support for 100+ languages. We study three deployment-oriented variants: a uni-encoder (147M), a shared-weight bi-encoder (145M), and GLiNER Guard Omni (307M).

Uni-Encoder (147M). The uni-encoder follows the standard GLiNER2 design: input text and schema labels are concatenated and processed jointly with full bidirectional attention. This maximizes text–label interaction but requires re-encoding the schema for every request.

Shared-Weight Bi-Encoder (145M). The bi-encoder encodes text and labels separately with a shared mmBERT-small backbone and projects them into a common matching space. Unlike prior GLiNER bi-encoders (Stepanov et al., 2026), which target single-task NER with separate towers, our variant shares weights across both branches and extends label-cache serving to the multi-task GLiNER2 setting. Formally, given input text x and schema labels $\{l_i\}_{i=1}^{K}$, the model computes pooled representations

$$\mathbf{h}_x = \mathrm{pool}(\mathrm{Enc}(x)),\ \mathbf{e}_i = \mathrm{pool}(\mathrm{Enc}(l_i)), \tag{1}$$

where Enc is shared and pool takes the first-token representation. Text and label vectors are projected to a $\mathbf{p} = W_t\mathbf{h}_x$, $\mathbf{q}_i = W_l\mathbf{e}_i$, and scored with a lightweight interaction head:

$$s_i = \mathrm{MLP}([\,\mathbf{p}_{:d};\ \mathbf{q}_{i,:d};\ \mathbf{p}_{d:} \odot \mathbf{q}_{i,d:}\,])\,. \tag{2}$$

The element-wise product captures multiplicative text–label interaction without full cross-attention, preserving compatibility with label caching.

GLiNER Guard Omni (307M). Omni is initialized from GLiNER2 Multi (Zaratiana et al., 2025) and fine-tuned on the same safety and PII supervision. Unlike the compact variants, which prioritize maximum throughput, Omni targets settings where broader transfer, multilingual robustness, and custom-taxonomy support are more important than edge deployment.

3.3 Training

We train all variants on 467,273 multi-task examples, using a 95/5 split between training and held-out validation data. Each sample may contain up to six supervision signals: span extraction, safety classification, adversarial attack detection, harmful-content categorization, intent recognition, and tone classification. This multi-task setup encourages shared representations across related moderation and extraction objectives.

The training mixture is dominated by classification supervision, while span-level NER annotations are substantially less frequent due to the lower availability of high-quality labeled PII data. In total, 108,702 examples contain span extraction labels covering 32 entity types.

For all public datasets, only designated training splits were used. No validation or test examples from evaluation benchmarks were included in training.

4 Experimental Setup

Our experiments address five questions: (q1–q2) safety moderation quality and quality–efficiency trade-off vs. larger guardrails (§5.1), (q3) unified PII detection (§5.2), (q4) whether a single model can replace a fragmented specialist stack without sacrificing per-task quality (§5.3), and (q5) how label schema design affects classification precision and recall on out-of-distribution safety data (§5.4). Extended serving benchmarks are in Appendix C.

4.1 Benchmarks

Safety. We evaluate on Aegis 2.0 (Ghosh et al., 2025) (prompt and response safety classification), StrongReject (Souly et al., 2024) (strongly harmful request detection), and PolyGuard (Kumar et al., 2025) (multilingual safety with prompt and response splits). We report F1 on the harmful/unsafe class and summarize with the unweighted mean $F1_{avg}$ across all five columns.

PII. We evaluate on three benchmarks. **PII-IMB** (Hoffelen, 2026) is a public zero-shot leaderboard measuring character-level, label-agnostic F2 across 5,000 samples from each of four datasets: OpenPII (Ai4Privacy, 2026) (multilingual, EN-dominant), Gretel PII (Gretel AI, 2024), Nemotron PII (Steier et al., 2025), and Privy (Kilimnik, 2022) (all EN-only). **PII-Bench** is our synthetic Russian-language benchmark (1,810 examples, 13 entity types, 9 domains, IAA 0.87; details in Appendix B). **SPY** (Savkin et al., 2025) is a public benchmark covering legal and medical domains.

4.2 Baselines

Safety baselines span autoregressive moderators (LlamaGuard 3, Llama 4 Guard, WildGuard, ShieldGemma, NemotronGuardV2, GPT-OSS-SafeGuard, YuFeng-XGuard) and prior encoder-only guardrails (PromptGuard 2, Longformer-harmful-ro). Schema-driven architectural baselines are GLiNER2 Multi v1 and GLiClass Instruct Base v1.0. PII baselines are Presidio (rule-based), Llama-3-70B (prompted), DeBERTa-v3 fine-tuned on SPY, and OpenAI Privacy Filter (1.5B sparse MoE, 50M active parameters).

5 Results

To contextualise quality and throughput results jointly, we report *moderation utility*—a single cost-efficiency scalar that combines task quality, serving speed, and hardware cost:

$$\text{Moderation Utility} = \frac{\text{Avg F1} \times \text{req/s}}{\text{GPU \$/hr}} \quad (3)$$

where Avg F1 is the unweighted mean F1 (%, 0–100 scale) over the five benchmark columns in Table 1, req/s is single-request throughput (batch=1, no dynamic batching), and GPU $/hr reflects on-demand RunPod pricing. Figure 2 plots this metric across two representative hardware tiers. The encoder-vs-autoregressive gap widens substantially on cheaper hardware—2.4× on an A100 SXM ($1.49/hr) and 6.5× on an RTX A4500 ($0.25/hr)—because autoregressive latency scales worse than encoder latency under reduced compute. Under dynamic batching the advantage grows further: the uni-encoder reaches 193.6 RPS on A100 with ONNX TensorRT (§5.1), roughly 3× the single-request rate used in Figure 2.

Figure 2: Moderation utility (Eq. 3) on two RunPod GPU tiers. Higher is better. Single-request mode; $F1_{avg}$ from Table 1, req/s from Table 2.

5.1 Safety Quality

Table 1 compares GLiNER Guard against autoregressive guardrails, prior encoder baselines, and architecture-matched schema-driven models (including GLiGuard).

GLiNER Guard is strongly competitive on safety moderation despite its compact size. **GLiNER Guard Omni** achieves the best overall encoder result with 76.9 $F1_{avg}$, improving over GLiNER2 Multi (66.6) by 10.3 points and GLiClass (64.9) by 12.0 points—matching GLiGuard (76.9) at 307M parameters. On multilingual response moderation (PolyGuard R), however, Omni leads GLiGuard by 15.3 F1 points (58.9 vs. 43.6), a gap the aggregate average conceals and that reflects the advantage of a multilingual backbone. On prompt-level moderation, the compact variants remain highly competitive: on Aegis 2.0 prompts, both uni-/bi-encoders reach 80.2 F1, within 1.3 points of WildGuard (81.5), while outperforming LlamaGuard 3 (77.2) and Llama 4 Guard (71.5). On StrongReject, the uni-encoder reaches 98.5 F1 and Omni 99.7.

The largest remaining gaps appear in response-level and multilingual moderation, where larger autoregressive models continue to lead. We address this through a cascade design that routes uncertain cases to a stronger second-stage moderator (Appendix F).

Table 2 compares single-request inference across representative hardware tiers, from A100 to a consumer 8 GB GPU and CPU. Compact GLiNER Guard variants consistently achieve the lowest latency and highest throughput. On A100, the bi-encoder reaches 60.5 req/s versus 21.8 for XGuard-Reason-8B and 1.5 for WildGuard. On an RTX A4500, the bi-encoder sustains 64.5 req/s while XGuard drops to 8.6 and WildGuard to 1.3. The advantage holds at the edge: on a Core Ultra 9 185H CPU, the bi-encoder reaches 22.8 req/s while WildGuard drops to 0.05 req/s. Compact encoders also run on a consumer 8 GB GPU (RTX 4060M) at over 100 req/s where larger autoregressive models exceed VRAM entirely. Full results across all eight tested tiers (including H100, RTX 4090, M4 Max) are in Appendix C.

5.2 PII Detection

PIIMB (external leaderboard). PIIMB (Hoffelen, 2026) measures zero-shot, character-level F2 across four public PII datasets (Table 4). GLiNER Guard Omni ranks **9th** on the public leaderboard (Avg F2 80.4%), leading on the multilingual **OpenPII** split at 93.0 F2 and outperforming all PII specialists including OpenMed-PII-SuperClinical-Large (90.7). On EN-only splits, specialized models lead. Omni's Avg F2 is within 0.6 points of its GLiNER2 Multi base (81.0, rank 8), showing that safety adaptation preserves PII transfer. The specialist model gliner2-privacy-filter-PII-multi reaches 83.5 Avg F2 (rank 4) at the cost of safety coverage.

PII-Bench (internal pipeline). Table 5 reports pipeline F1 on our Russian-language benchmark for the two context-dependent entity types—NAME and ADDRESS—that require learned span extraction. Structured entities (phones, card numbers, tax identifiers) are handled by deterministic rules and are excluded here.

GLiNER Guard variants outperform GLiNER2 baselines on both types after pipeline post-processing. The uni-encoder leads on NAME (75.7 F1 vs. 60.6 for GLiNER2 Multi), while the bi-encoder leads on ADDRESS (68.7 vs. 52.1). Raw model-only ADDRESS scores are near zero across all models, as the benchmark expects consolidated multi-token address spans that no model produces directly; pipeline gains reflect successful sub-span detection by the model, aggregated in post-processing. These results confirm that GLiNER Guard provides useful PII coverage for the entity types that

Model	Size	Task	Aegis 2.0		StrongReject	PolyGuard		Avg
			P	R		P	R	
Autoregressive								
Llama3Guard	8B	s-cls	77.2	66.1	98.5	68.5	66.3	75.3
Llama4Guard	12B	s-cls	71.5	64.7	95.3	62.4	54.7	69.7
WildGuard	7B	s-cls	81.5	82.7	99.5	74.7	66.1	80.9
YuFeng-XGuard	8B	s-cls	**86.4**	**80.4**	**100.0**	**85.8**	**79.2**	**86.4**
Encoder baselines								
Longformer-harmful-ro	149M	s-cls	**52.3**	**49.8**	**81.7**	32.5	23.2	**47.9**
Architecture baselines								
GLiClass Base v1.0	150M	cls	74.1	64.5	91.3	57.6	37.1	64.9
GLiNER2 Multi v1	307M	cls+IE	71.4	75.7	95.7	60.3	29.9	66.6
GLiGuard LLM-Safety	208M	cls+IE	**84.4**	**77.7**	**98.1**	**80.6**	**43.6**	**76.9**
Ours								
GLiNER Guard bi-enc	145M	s-cls+IE	**80.2**	74.4	97.7	69.5	49.9	74.3
GLiNER Guard uni-enc	147M	s-cls+IE	**80.2**	**75.7**	98.5	66.2	48.5	73.8
GLiNER Guard Omni	307M	s-cls+IE	79.8	74.6	**99.7**	**71.7**	**58.9**	**76.9**

Table 1: Safety moderation results (F1, %) on Aegis 2.0, StrongReject, and PolyGuard. Avg is the unweighted mean over the five benchmark columns. Abbreviations: P = prompt, R = response.

Model	RTX 4060M 8 GB	RTX A2000 16 GB	RTX A4500 20 GB	A100 80 GB	Core Ultra 9 185H (CPU)
Latency (ms/req) ↓					
GLiNER Guard uni-enc	8.7	16.5	16.3	19.4	62.8
GLiNER Guard bi-enc	**8.3**	**16.4**	**15.5**	**16.5**	**43.8**
GLiNER Guard Omni	10.5	18.6	17.7	18.8	364.7
XGuard-Reason-8B	OOM	OOM	116.7	45.9	1,253
WildGuard	OOM	1,744	751.5	684.9	19,578
Throughput (req/s) ↑					
GLiNER Guard uni-enc	114.3	60.6	61.2	51.6	15.9
GLiNER Guard bi-enc	**120.5**	**60.9**	**64.5**	**60.5**	**22.8**
GLiNER Guard Omni	95.6	53.7	56.4	53.1	2.7
XGuard-Reason-8B	OOM	OOM	8.6	21.8	0.8
WildGuard	OOM	0.57	1.3	1.5	0.05

Table 2: Single-request inference. FP16, batch 1, N=350 prompts. OOM: exceeded VRAM. Full results in Appendix C.

matter most in production, without requiring a separate dedicated NER model.

Full per-domain and per-entity breakdowns are provided in Appendix D.

5.3 Unified vs. Fragmentation

Where related encoder work consolidates safety and PII into two specialist models—GLiGuard (208M, safety only) and gliner2-privacy-filter-PII-multi (307M, PII only)—requiring 515M hot parameters and two forward passes per request, GLiNER Guard Omni unifies both tasks in a single 307M model with one forward pass.

Table 3 shows that Omni matches the safety specialist in $\mathrm{F1}_{avg}$ (76.9) and stays within 3.1 Avg F2 points of the PII specialist on PIIMB. The trade-off is selective: on English-dominant PII splits, Omni trails the specialist (-3.1 Avg F2 on PIIMB overall), but leads on the multilingual OpenPII split ($+2.4$ F2), suggesting that the cost of unification is concentrated in English-only structured PII extraction — a regime where rule-based fallbacks are often more reliable in practice. Under multilingual production conditions, unification does not require a substantial quality trade-off.

5.4 Schema Sensitivity

The schema-driven interface allows operators to pass arbitrary label sets and descriptions at inference time. We evaluate how three schema designs affect the precision–recall trade-off on Aegis 2.0, a benchmark not seen during training. All three variants use the same GLiNER Guard Omni checkpoint; only the labels differ. *Binary*: bare `safe`/`unsafe` labels without descriptions. *Binary + desc.*: the same two classes with short natural-language descriptions. *Taxonomy*: 8 specific harm categories plus `safe`, aggregated via:

$$\hat{y}(x) = \begin{cases} \texttt{unsafe} & \text{if } \hat{\mathcal{L}}(x) \not\subseteq \{\texttt{safe}\} \\ \texttt{safe} & \text{otherwise} \end{cases} \quad (4)$$

where $\hat{\mathcal{L}}(x)$ is the set of predicted labels for input x. A single fired harm category is sufficient to trigger an unsafe verdict. Full schema listings are in Appendix G.

Approach	Params	Safety F1	Poly F1	PIIMB F2	OpenPII F2
Concurrent specialized models					
GLiGuard (safety only)	208M	**76.9**	62.1	—	—
GLiNER2 PII (privacy only)	307M	—	—	**83.5**	90.6
Specialized stack	515M	**76.9**	62.1	**83.5**	90.6
Ours — unified					
GLiNER Guard Omni	**307M**	**76.9**	**65.3**	80.4	**93.0**

Table 3: Unified vs. specialized deployment. GLiNER Guard Omni (307M, one pass) vs. the two-specialist stack (515M, two passes).

Model	OpenPII	Gretel PII	Nemotron PII	Privy	Avg F2
Specialized PII models					
OpenMed-PII-SC-Large	90.7	**94.5**	**96.3**	**73.4**	**88.7**
privacy-filter-multilingual	85.8	86.4	76.6	46.8	73.9
OpenAI privacy-filter	82.1	86.9	60.7	53.6	70.8
llama-ai4privacy	54.0	68.9	44.1	33.8	50.2
General-purpose / architectural baselines					
GLiNER2 PII	90.6	93.6	85.4	64.6	83.5
GLiNER2 Multi v1	88.3	87.0	82.8	66.1	81.0
GLiNER2 Large v1	81.1	87.2	85.7	60.2	78.6
Ours					
GLiNER Guard Omni	**93.0**	89.1	75.9	63.6	80.4

Table 4: PIIMB zero-shot leaderboard (Hoffelen, 2026): character-level F2 (%) across four datasets. Omni leads on multilingual OpenPII (93.0), confirming safety adaptation preserves PII transfer.

Model	NAME	ADDRESS
Concurrent work (PII specialist)		
GLiNER2 PII	50.0	49.5
Architecture baselines		
GLiNER2 Multi	60.6	52.1
GLiNER2 Large	30.3	35.9
Ours		
GLiNER Guard uni-enc	**75.7**	65.9
GLiNER Guard bi-enc	69.5	**68.7**
GLiNER Guard Omni	52.3	54.6

Table 5: PII-Bench pipeline F1 (%) on NAME and ADDRESS. End-to-end pipeline results (span merging applied).

	User prompts			LLM responses		
Method	$F1_{\uparrow}$	$FPR_{\downarrow}$	$FNR_{\downarrow}$	$F1_{\uparrow}$	$FPR_{\downarrow}$	$FNR_{\downarrow}$
binary	0.798	0.464	**0.074**	0.746	0.142	0.307
binary + desc.	**0.817**	0.385	0.083	**0.752**	0.135	**0.302**
taxonomy	0.813	**0.199**	0.199	0.719	**0.090**	0.381

Table 6: Effect of label schema on GLiNER Guard Omni evaluated on Aegis 2.0 test split. Positive class: `unsafe`. Bold: best per column. Binary + desc. improves F1 over bare binary labels; taxonomy trades recall for a 2× FPR reduction. Schema listings in Appendix G.

Table 6 shows the results. Adding descriptions (*binary + desc.*) improves F1 over bare labels on both user prompts (0.817 vs. 0.798) and LLM responses (0.752 vs. 0.746), confirming that richer label semantics aid zero-shot transfer. The taxonomy schema provides a qualitatively different operating point: it reduces FPR by roughly 2× on user prompts (0.199 vs. 0.385) and 1.5× on responses (0.090 vs. 0.135), at the cost of higher FNR. This reflects the conservative aggregation rule—a text must activate at least one specific harm category to be flagged—making the taxonomy schema suitable for precision-critical deployments where false alarms dominate the cost function.

6 Conclusion

We presented **GLiNER Guard**, a schema-driven encoder family that consolidates safety classification and PII detection into a single forward pass. Rather than claiming state-of-the-art on either task in isolation, our contribution is a different operating point on the throughput–quality–cost frontier: always-on, multi-task first-stage protection that becomes economically viable where large autoregressive moderators do not.

The results support this operating point. GLiNER Guard Omni matches the safety specialist GLiGuard on aggregate $F1_{avg}$ (76.9) while leading on multilingual response moderation (+15.3 F1 on PolyGuard R), approaches the dedicated PII specialist on English-dominant benchmarks (within 3.1 Avg F2), and outperforms it on multilingual PII (+2.4 F2 on OpenPII) — without sacrificing either task. The efficiency advantage over autoregressive moderators reaches 6.5× moderation utility on cheaper hardware,

widening precisely where always-on deployment is most cost-sensitive.

We additionally release **PII-Bench**, a Russian-language span-annotated benchmark (1,810 examples, 13 entity types, 9 domains), to support evaluation in a setting underserved by existing leaderboards. We view unified encoder guardrails not as a replacement for large autoregressive moderators, but as a complementary first-stage tier that collapses a fragmented stack into a single deployable unit under realistic latency and cost constraints. Both GLiNER Guard and its shared-weight bi-encoder variant — adapted from Stepanov et al. (2026) to enable label-cache serving — are contributed as extensions of the GLiNER ecosystem, demonstrating that minimal architectural adaptation suffices to cover production safety and privacy requirements.

Released models and benchmark are distributed under the Apache-2.0 license.

Limitations

PII-Bench is Russian-only. PII-Bench covers Russian-language inputs and Russian-specific identifiers (INN, SNILS, OGRN, etc.). Cross-lingual PII evaluation therefore relies on external benchmarks (PIIMB, SPY), limiting direct analysis of multilingual transfer.

Response-level moderation. Response-level moderation remains weaker than large-model baselines, particularly in multilingual settings. The main limitation is data availability: public safety corpora are dominated by prompt-side supervision, while multilingual response annotations remain scarce. Cascade routing to a stronger second-stage moderator (App. F) recovers much of this gap, but does not eliminate the underlying supervision bottleneck.

Future work.

- **ModernBERT-based schema supervision.** mmBERT/ModernBERT backbones provide strong inference efficiency but currently underperform DeBERTa-style encoders on schema-driven supervision.
- **Bi-encoder fusion.** Our shared-weight bi-encoder uses a lightweight projection mechanism; richer fusion strategies (Stepanov et al., 2026) may improve text–label interaction while preserving label caching.
- **Long-context moderation.** Although mmBERT-small supports longer contexts, our supervision remains short-context dominated. Extending training to multi-turn and document-level moderation is left for future work.

Ethical Considerations

Intended use. GLiNER Guard is intended as a first-stage filter in LLM pipelines, not as the sole decision-maker for high-stakes actions. Response-level moderation remains weaker in multilingual settings (§5.1); consequential decisions should therefore involve cascade routing or human review.

Dual use and overblocking. Like any moderation system, GLiNER Guard can be repurposed for censorship beyond its intended scope. Deployments should consider transparency and appeals mechanisms for flagged content.

Synthetic-only PII data. PII-Bench contains no real personal data: all examples were synthetically generated and manually reviewed by native Russian-speaking annotators (Appendix B). Russian-specific identifiers follow valid format conventions but do not correspond to real individuals.

Adversarial robustness. Although the training mixture includes adversarial examples, we do not evaluate robustness against paraphrased attacks or novel jailbreak families. Production deployments should therefore not assume strong out-of-distribution robustness.

References

Ai4Privacy. 2026. OpenPII 1M — multilingual pii masking dataset (19 labels, 23 languages).

Fastino AI Team. 2026. GLiNER2-PII: Multilingual PII extraction via synthetic fine-tuning.

Shaona Ghosh, Prasoon Varshney, Makesh Narsimhan Sreedhar, Aishwarya Padmakumar, Traian Rebedea, Jibin Rajan Varghese, and Christopher Parisien. 2025. AEGIS2.0: A diverse AI safety dataset and risks taxonomy for alignment of LLM guardrails. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5992–6026.

Gretel AI. 2024. GLiNER models for PII detection through fine-tuning on Gretel-generated synthetic documents.

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

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

Camille Van Hoffelen. 2026. PII masking benchmark.

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

Benjamin Kilimnik. 2022. Privy synthetic pii protocol trace dataset.

Knowledgator. 2024. FlashDeBERTa: Memory-efficient attention for deberta. https://github.com/Knowledgator/FlashDeBERTa.

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

Junyu Lin, Meizhen Liu, Xiufeng Huang, Jinfeng Li, Haiwen Hong, Xiaohan Yuan, Yuefeng Chen, Longtao Huang, Hui Xue, Ranjie Duan, Zhikai Chen, Yuchuan Fu, Defeng Li, Lingyao Gao, and Yitong Yang. 2026. YuFeng-XGuard: A reasoning-centric, interpretable, and flexible guardrail model for large language models. *arXiv preprint arXiv:2601.15588*.

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *arXiv preprint arXiv:2509.06888*.

Meta AI. 2025. Llama 4 guard. Available at https://huggingface.co/meta-llama/Llama-4-Guard-12B.

Microsoft. 2018. Presidio – data protection and de-identification SDK. Available at https://github.com/microsoft/presidio.

Maksim Savkin, Timur Ionov, and Vasily Konovalov. 2025. SPY: Enhancing privacy with synthetic PII detection dataset. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*.

Amy Steier, Andre Manoel, Alexa Haushalter, and Maarten Van Segbroeck. 2025. Nemotron-PII: Synthesized data for privacy-preserving AI.

Ihor Stepanov, Mykhailo Shtopko, Dmytro Vodianytskyi, and Oleksandr Lukashov. 2026. The million-label NER: Breaking scale barriers with GLiNER bi-encoder. *arXiv preprint arXiv:2602.18487*.

Ihor Stepanov, Mykhailo Shtopko, Dmytro Vodianytskyi, Oleksandr Lukashov, Alexander Yavorskyi, and Mykyta Yaroshenko. 2025. GLiClass: Generalist lightweight model for sequence classification tasks. *arXiv preprint arXiv:2508.07662*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context fine-tuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Urchade Zaratiana, Mary Newhauser, George Hurn-Maloney, and Ash Lewis. 2026. GLiGuard: Schema-conditioned classification for LLM safeguard.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. GLiNER2: An efficient multi-task information extraction system with schema-driven interface. *arXiv preprint arXiv:2507.18546*.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376. Association for Computational Linguistics.

A Training Details

A.1 Training Hyperparameters

Table 7 reports the optimization settings used for all model variants.

Parameter	uni/bi-encoder	Omni
Backbone	mmBERT-small	mDeBERTa (GLiNER2-Multi-v1)
Max sequence length	384	384
Max span width	12	8
Epochs	3	3
Batch size	64	32
Encoder LR	1×10^{-5}	1×10^{-5}
Task heads LR	2×10^{-5}	2×10^{-5}
Scheduler	cosine	cosine
Optimizer	AdamW	AdamW
Warmup ratio	0.1	0.1
Weight decay	0.01	0.01
Max gradient norm	10.0	10.0
Precision	bf16	bf16

Table 7: Training hyperparameters. Omni was trained with FlashDeBERTa (Knowledgator, 2024) for memory efficiency.

All variants were trained on a single NVIDIA A100 80 GB GPU. The uni-encoder and bi-encoder required approximately 8 GPU-hours each; GLiNER Guard Omni required approximately 30 GPU-hours due to its larger backbone (mDeBERTa) and FlashDeBERTa attention implementation (Knowledgator, 2024). Total training compute for the released model family is approximately 46 GPU-hours.

A.2 Training Data

GLiNER Guard is trained on 467,273 multi-task samples. Each sample carries up to 6 simultaneous tasks: 1 span extraction (NER) and 5 classification tasks. Table 8 lists the public datasets used; the corpus also includes internal Russian translations of WildGuardMix (Han et al., 2024b) and Aegis 2.0 (Ghosh et al., 2025), as well as internal safety data.

Dataset	Domain	Lang	License
WildGuardMix	Safety	EN	ODC-BY
Aegis 2.0	Safety	EN	CC-BY-4.0
Ru Toxic Texts	Toxicity	RU	MIT
PII NER Corpus Synthetic	PII / NER	EN	MIT
Nemotron-PII	PII / NER	EN	CC-BY-4.0
Synthetic PII Finance	PII / NER	Multi	Apache-2.0
Russian Dialogues	Dialogues	RU	MIT
GrandMaster-PRO-MAX	General	RU/EN	Apache-2.0

Table 8: Public data sources used in training.

B PII-Bench Benchmark Specification

PII-Bench is a Russian-language benchmark for span-level PII detection. It uses explicit character-level offsets, enabling evaluation of

Group	Label	Count	%
Person	person	52,754	13.54
	first_name	13,020	3.34
	last_name	7,877	2.02
	alias	3,493	0.90
	title	385	0.10
Location	street	26,227	6.73
	city	24,529	6.30
	country	12,551	3.22
	region	8,904	2.29
	postal_code	7,954	2.04
	unit	3,764	0.97
	address	3,615	0.93
	district	2,263	0.58
	building	2,159	0.55
	landmark	746	0.19
Organization	company	36,365	9.33
	product	3,196	0.82
	government	2,105	0.54
	education	1,193	0.31
	media	708	0.18
Contact	email	31,024	7.96
	phone	25,027	6.42
	social_account	3,070	0.79
	messenger	91	0.02
Identity	national_id	26,782	6.87
	document_id	16,912	4.34
	passport	732	0.19
Temporal	event_date	31,335	8.04
	date_of_birth	11,571	2.97
Financial	bank_account	18,845	4.84
	card_number	9,304	2.39
	crypto_wallet	1,110	0.28

Table 9: NER entity distribution in training data.

complete systems including model predictions, post-processing, and rule-based components. Synthetic examples were generated using Claude Sonnet 4.5 with default parameters, using domain-specific seed prompts for each of the 9 scenarios and prompt rotation across multiple templates per domain. All generated examples were manually reviewed and filtered for realism by two annotators before annotation. Two annotators independently reviewed all examples and resolved disagreements through discussion. Inter-annotator agreement (pairwise F1, exact span match) is **0.87**.

The benchmark covers 13 PII entity types common in Russian-language interactions (Table 10), with 9 domains across two sensitivity levels (Table 11).

Type	Description	N
NAME	Full names (all grammatical cases)	70
PHONE_NUMBER	Phone numbers	70
EMAIL	Email addresses	70
ADDRESS	Physical addresses	70
BANK_CARD_NUMBER	Bank card numbers	70
CVC	CVC/CVV codes	70
INN	Taxpayer ID	70
KPP	Tax registration code	70
OGRN	State registration number	70
OGRNIP	Individual entrepreneur reg. number	70
SNILS	Social insurance number	70
PASSPORT_NUMBER	Passport numbers	70
TOKEN	API tokens, recovery keys	70

Table 10: PII-Bench entity types.

Domain	Description	N	With PII
S-BANK	Banking support	100	65 (65%)
S-TELECOM	Telecom support	100	62 (62%)
S-DELIVERY	Delivery service	100	51 (51%)
S-AUTO	Auto service	100	58 (58%)
S-HR	HR / recruiting	100	65 (65%)
S-RE	Real estate	100	66 (66%)
S-SUPPORT	General support	100	55 (55%)
L-CHAT	Chats / messengers	100	50 (50%)
L-DIALOG	Multi-turn dialogues	100	50 (50%)

Table 11: PII-Bench domains. S = sensitive (recall-priority); L = low-sensitivity (precision-priority).

The benchmark has two complementary splits. The **entity split** (910 examples) groups examples by entity type; the **domain split** (900 examples) groups by scenario with a natural PII/non-PII mix. Total: 1,810 examples; character-level PII density is 19.6%.

C Extended Serving Diagnostics

GLiNER Guard uni-encoder achieves higher throughput than GLiNER2 Multi despite similar parameter counts, primarily due to its mmBERT-small backbone (ModernBERT architecture (Warner et al., 2025)) which uses rotary positional embeddings and more efficient attention patterns compared to mDeBERTa's disentangled attention. Under dynamic batching on a single A100 80 GB (LitServe, max batch 64, timeout 50 ms), GLiNER Guard reaches **193.6 RPS** under ONNX TensorRT —a 58% throughput improvement over GLiNER2 Multi (122.6 RPS) and a 36% P99 latency reduction (900 ms vs. 1,400 ms). GLiNER2 Multi exhibits a 12.95% error rate under concurrent load; GLiNER Guard maintains zero errors across all three tested backends.

Table 12 reports single-request latency and throughput across all eight tested hardware tiers, extending Table 2 with H100 SXM, RTX 4090, and Apple M4 Max results.

Table 13 reports throughput and latency percentiles under dynamic batching (LitServe, max batch 64, timeout 50 ms, A100 80 GB) across three runtimes.

Figure 3 shows the full latency breakdown per backend.

D Extended PII Evaluation

D.1 PII-Bench: Raw Model Extraction

D.2 PII-Bench: End-to-End Pipeline Results

D.3 External Benchmark: SPY

E Generalization: CrossNER

F Cascade Design

The remaining quality gap is concentrated in long outputs, multilingual inputs, and harder response-level judgments. We address this as a routing problem: pairing GLiNER Guard with YuFeng-XGuard (Lin et al., 2026) as a second stage improves moderation quality while routing only a fraction of traffic to the more expensive LLM tier.

G Schema Sensitivity on Aegis 2.0

We compare three label schema designs for GLiNER Guard Omni evaluated on the Aegis 2.0 test split. All variants use the same model and threshold; only the label set passed to the schema-driven interface differs.

Model	RTX 4060M 8 GB	RTX A2000 16 GB	RTX A4500 20 GB	A100 SXM 80 GB	H100 SXM 80 GB	RTX 4090 24 GB	M4 Max (MPS)	Core Ultra 9 185H (CPU)
Latency (ms/req) ↓								
GLiNER Guard uni-enc	8.7	16.5	16.3	19.4	11.8	14.1	18.7	62.8
GLiNER Guard bi-enc	**8.3**	**16.4**	**15.5**	**16.5**	**10.3**	**13.9**	**16.0**	**43.8**
GLiNER Guard Omni	10.5	18.6	17.7	18.8	14.2	15.7	18.6	364.7
XGuard-Reason-8B	OOM	OOM	116.7	45.9	26.8	47.0	679.6	1,253
WildGuard	OOM	1,744	751.5	684.9	407.2	556.6	1,633.6	19,578
Throughput (req/s) ↑								
GLiNER Guard uni-enc	114.3	60.6	61.2	51.6	84.5	71.0	53.4	15.9
GLiNER Guard bi-enc	**120.5**	**60.9**	**64.5**	**60.5**	**96.9**	**71.8**	**62.3**	**22.8**
GLiNER Guard Omni	95.6	53.7	56.4	53.1	70.2	63.5	53.7	2.7
XGuard-Reason-8B	OOM	OOM	8.6	21.8	37.3	21.3	1.5	0.8
WildGuard	OOM	0.57	1.3	1.5	2.5	1.8	0.61	0.05

Table 12: Appendix C. Full single-request inference results across all tested hardware tiers. FP16, batch size 1, N=350 prompts (StrongREJECT). OOM: model exceeded VRAM. RTX 4060M and RTX A2000: large autoregressive models offloaded to CPU and excluded. M4 Max uses MPS backend. XGuard: `max_new_tokens=1`; WildGuard: `max_new_tokens=32`. **Bold**: best per column.

Model	Runtime	RPS↑	P50 (ms)↓	P95 (ms)↓	P99 (ms)↓	Err (%)↓
GLiNER Guard uni-enc	PyTorch FP16	148.2	570	1500	1700	0.00
	ONNX CUDA FP16	170.6	540	870	1000	0.00
	ONNX TensorRT FP16	**193.6**	**480**	**750**	**900**	0.00
GLiNER2 Multi	PyTorch FP16	83.7	1200	2000	2400	12.95
	ONNX CUDA FP16	90.8	1000	1700	2100	0.00
	ONNX TensorRT FP16	**122.6**	**740**	**1200**	**1400**	0.00

Table 13: Appendix C. Serving performance under dynamic batching (LitServe, max batch 64, timeout 50 ms, NVIDIA A100 80 GB). RPS = requests per second; P50/P95/P99 = end-to-end latency percentiles; Err = HTTP error rate.

Domain	GG uni	GG bi	GG Omni	GLiNER2 Multi	GLiNER2 Large
L-CHAT	75.8	86.5	**92.6**	85.5	64.3
L-DIALOG	29.9	**44.8**	44.0	42.3	24.6
S-AUTO	30.0	40.5	70.6	**74.2**	36.4
S-BANK	16.8	23.2	**71.6**	69.4	45.8
S-DELIVERY	54.2	54.6	54.1	**63.7**	40.6
S-HR	28.9	41.5	**82.6**	78.8	52.1
S-RE	21.2	27.4	64.2	**70.1**	33.8
S-SUPPORT	38.0	61.2	**81.6**	74.8	63.8
S-TELECOM	24.0	33.9	**71.1**	69.9	43.9
Avg	35.4	46.0	**70.3**	69.9	45.0

Table 14: Appendix D1. PII-Bench: model-only F1 (%) by domain.

Entity	GG uni	GG bi	GG Omni	GLiNER2 Multi	GLiNER2 Large
EMAIL	**100.0**	97.2	**100.0**	**100.0**	**100.0**
NAME	74.6	63.7	83.1	**85.2**	27.8
PHONE_NUMBER	21.8	20.6	**87.2**	76.0	57.5
BANK_CARD	19.8	21.3	**69.4**	62.7	67.3
CVC	21.3	0.0	40.5	27.8	**41.1**
PASSPORT	18.8	12.8	**60.6**	51.7	16.1
TOKEN	18.6	9.7	**66.7**	33.1	14.8
ADDRESS	0.5	0.0	**6.6**	5.2	0.0
INN	0.0	0.0	96.6	**97.2**	0.0
KPP	5.7	0.0	81.4	**81.9**	12.5
OGRN	0.0	0.0	88.3	**99.3**	30.4
OGRNIP	19.2	0.0	90.2	**98.6**	10.8
SNILS	14.0	0.0	75.5	**75.7**	75.0
Avg	24.2	17.3	**72.8**	68.8	34.9

Table 15: Appendix D1. PII-Bench: model-only F1 (%) by entity type.

Schema designs. *Binary*: bare `safe`/`unsafe` labels with no descriptions. *Binary + desc.*: the same two labels augmented with short natural-

Domain	Pipe+GG uni	Pipe+GG bi	Pipe+GG Omni	Pipe+GLiNER2 Multi	Pipe+GLiNER2 Large
L-CHAT	**97.1**	94.3	75.2	75.8	60.6
L-DIALOG	**76.6**	75.2	63.6	47.4	25.1
S-AUTO	**74.9**	74.3	73.0	71.7	58.1
S-BANK	**84.9**	83.9	78.5	75.6	68.2
S-DELIVERY	**86.4**	83.3	75.9	77.1	61.9
S-HR	**93.0**	89.4	87.7	83.8	65.9
S-RE	79.8	**80.0**	69.0	71.0	59.9
S-SUPPORT	**94.4**	93.8	90.7	87.2	76.8
S-TELECOM	72.3	**75.2**	70.3	68.4	61.4
Avg	**84.4**	83.3	76.0	73.1	59.8

Table 16: Appendix D2. PII-Bench: full pipeline F1 (%) by domain.

Entity	Pipe+GG uni	Pipe+GG bi	Pipe+GG Omni	Pipe+GLiNER2 Multi	Pipe+GLiNER2 Large
Rule-based (identical across models)					
EMAIL†	100.0	100.0	100.0	100.0	100.0
TOKEN†	100.0	100.0	100.0	100.0	100.0
KPP†	100.0	100.0	100.0	100.0	100.0
OGRNIP†	100.0	100.0	100.0	100.0	100.0
BANK_CARD†	95.5	95.5	95.5	95.5	95.5
OGRN†	95.5	95.5	95.5	95.5	95.5
SNILS†	95.5	95.5	95.5	95.5	95.5
CVC†	94.7	94.7	94.7	94.7	94.7
INN†	88.0	88.0	88.0	88.0	88.0
PHONE_NUMBER†	60.4	60.4	60.4	60.4	60.4
PASSPORT†	49.6	49.6	49.6	49.6	49.6
Model-dependent					
NAME	**75.7**	69.5	52.3	60.6	30.3
ADDRESS	65.9	**68.7**	54.6	52.1	35.9
Avg	**86.2**	85.9	83.5	84.0	80.4

† Rule-based; model has no effect.

Table 17: Appendix D2. PII-Bench: full pipeline F1 (%) by entity type.

Model	Presidio	Llama-3-70B	DeBERTa-v3†	GG (uni)	GG (bi)	GG Omni
Params	–	70 B	184 M	147 M	145 M	209 M
(a) Legal Questions						
Name	79.4	68.9	**93.2**	73.6	72.6	88.8
Email	91.8	88.5	**99.1**	96.0	96.0	95.9
User.	–	59.7	**98.0**	40.9	43.9	43.8
URL	21.3	92.5	**99.0**	–	–	19.5
ID	34.4	62.2	**96.6**	80.7	75.8	80.9
Phone	68.1	92.8	**98.7**	86.8	90.4	89.4
Addr.	–	81.3	94.5	83.4	83.8	**95.5**
(b) Medical Consultations						
Name	80.4	62.9	**88.7**	65.9	62.3	78.2
Email	92.2	90.9	**99.5**	94.2	94.2	93.9
User.	–	70.4	**95.4**	47.7	53.4	54.4
URL	19.4	91.9	**98.9**	–	–	24.1
ID	38.9	75.1	**98.3**	89.1	84.9	86.3
Phone	65.5	90.0	**96.9**	86.6	89.0	87.5
Addr.	–	90.4	**95.1**	82.2	82.7	93.7

† Fine-tuned on SPY cross-domain. GG: GLiNER Guard, zero-shot.

Table 18: Appendix D3. PII entity recall (%) on SPY. "–" = entity type outside training ontology.

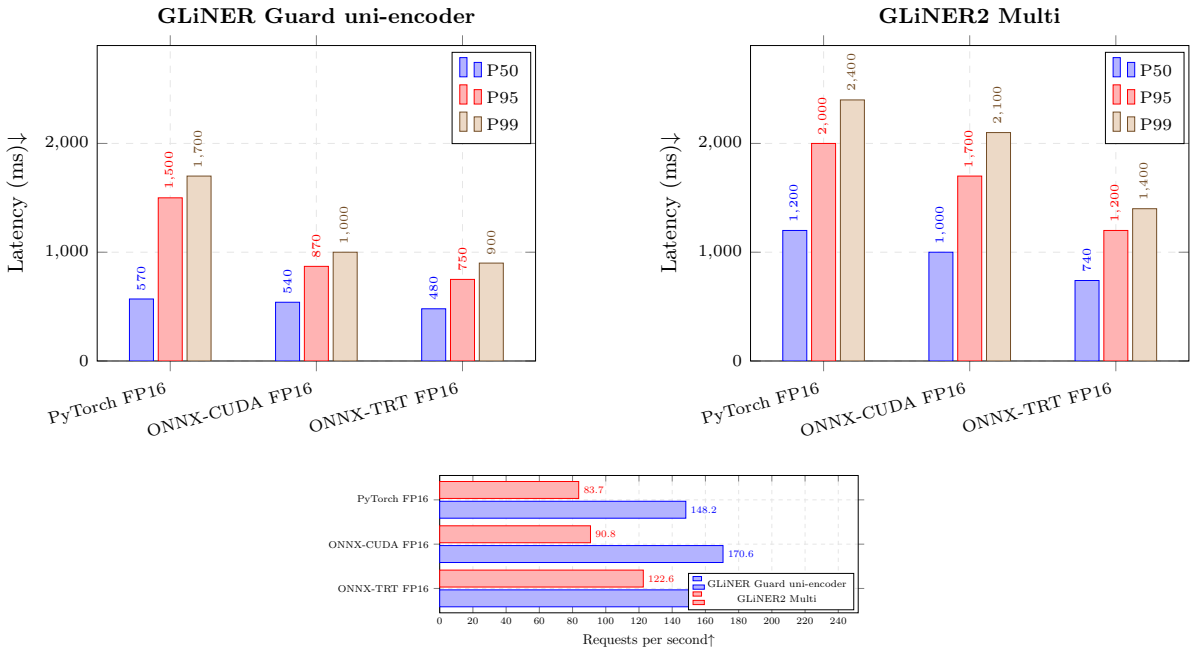

Figure 3: Serving under dynamic batching (LitServe, max batch 64, A100 80 GB). Top: latency percentiles per backend. Bottom: throughput.

Model	Params	AI	Lit.	Music	Politics	Science	Avg
GPT-4o	>100B	54.7	56.1	73.6	63.2	51.8	59.9
GLiNER-M	–	51.8	59.7	69.4	68.6	58.1	61.5
GLiNER2 (Zaratiana et al., 2025)	307M	52.6	56.4	63.2	67.9	54.7	59.0
GLiNER Guard bi-enc	145M	15.3	13.5	10.1	21.2	13.4	14.7
GLiNER Guard uni-enc	147M	14.3	11.7	10.6	19.1	12.3	13.6
GLiNER Guard Omni	307M	45.9	50.3	58.0	56.2	46.8	51.4

Table 19: Appendix E. Zero-shot NER strict F1 (%) on CrossNER. Compact encoders specialize for safety; Omni retains strong open-domain transfer.

language descriptions (Listing 1). *Taxonomy*: a 9-class vocabulary (8 unsafe categories + `safe`), aggregated to safe/unsafe at inference time (Listing 2). Descriptions for the *binary* + *desc.* and *taxonomy* schemas were generated with Claude Opus using only the dataset README from the Aegis 2.0 Hugging Face repository as input, without access to the dataset examples or label definitions beyond those stated in the README.

Results and the aggregation rule (Eq. 4) are in §5.4. The schema listings used in that experiment are provided below.

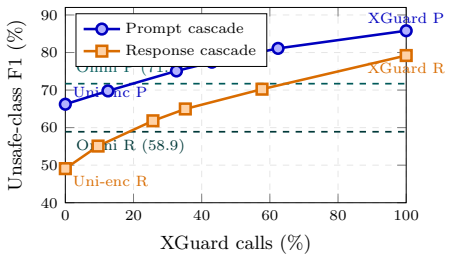

Figure 4: Appendix F. Cascade inference on PolyGuard: unsafe-class F1 vs. XGuard call rate at five GLiNER confidence thresholds ($\tau \in \{0.5, 0.7, 0.9, 0.95, 0.99\}$). **Solid curves**: cascade (leftmost point = Uni-encoder alone; rightmost = XGuard 8B alone). **Dashed lines**: Omni standalone (no cascade), shown for reference. Moving right trades encoder throughput for LLM quality (see Table 20).

Model	Params	Poly P	Poly R	Type
Longformer-harmful-ro	149M	32.5	23.2	encoder
GLiNER2 Multi	307M	60.3	29.9	encoder
Llama4Guard	12B	62.4	54.7	autoregressive
GG uni-enc (alone)	147M	66.2	49.1	encoder
GG bi-enc (alone)	145M	69.5	49.9	encoder
Llama3Guard	8B	68.5	66.3	autoregressive
GG Omni (alone)	307M	71.7	58.9	encoder
WildGuard	7B	74.7	66.1	autoregressive
Cascade τ=0.95	147M+8B	77.4	65.0	cascade
Cascade τ=0.99	147M+8B	81.1	70.2	cascade
YuFeng-XGuard (alone)	8B	85.8	79.2	autoregressive

Table 20: Appendix F. Cascade vs. standalone on PolyGuard (unsafe-class F1, %). P = Prompt, R = Response. Cascade rows combine GG uni-enc with XGuard 8B at two confidence thresholds.

Listing 1: Compact binary schema.

```
AEGIS_LABELS_COMPACT = {
    "unsafe": (
        "harmful content: hate, violence, threats,
      harassment, "
        "sexual content, self-harm, weapons, drugs,
      crime, PII, "
        "fraud, malware, misinformation, unauthorized
      advice, "
        "or manipulation"
    ),
    "safe": (
        "benign content: helpful answers, safe refusals,
      "
        "neutral information, everyday conversation"
    ),
}
```

Listing 2: Taxonomy schema (9 classes); aggregation rule in Eq. 4.

```
AEGIS_LABELS_TAXONOMY = {
    "interpersonal_harm": (
        "violence, physical attacks, killing, gore,
      threats to "
        "harm people, bullying, harassment, stalking,
      intimidation, "
        "hate speech, racism, discrimination against
      identity groups"
    ),
    "criminal_and_dangerous": (
        "planning crimes, theft, breaking and entering,
      evading "
        "law enforcement, confessions of illegal acts,
      guns, "
        "firearms, explosives, illegal weapons, fraud,
      scams, "
        "phishing, hacking, malware, or other illegal
      activity"
    ),
    "substances": (
        "illegal drugs, drug manufacturing or
      trafficking, "
        "alcohol abuse, prescription medication misuse"
    ),
    "sexual_content": (
        "explicit sexual acts, pornographic content,
      sexually "
        "explicit material, or any sexual or suggestive
      content "
        "involving minors"
    ),
    "self_harm": (
        "suicide methods, self-injury, cutting, eating
      disorders, "
        "or encouragement of self-destructive behavior"
    ),
    "privacy_violation": (
        "personal addresses, phone numbers, social
      security "
        "numbers, financial accounts, doxxing, exposing
      private "
        "information about real people"
    ),
    "profanity": (
        "curse words, vulgar expressions, crude or
      offensive "
        "language"
    ),
    "informational_harm": (
        "false claims, misinformation, conspiracy
      theories, "
        "unqualified medical legal or financial advice, "
        "psychological manipulation, copyright
      violations, "
        "or other unethical content"
    ),
    "safe": (
        "benign content: helpful answers, neutral
      information, "
        "educational discussion, or everyday
      conversation"
    ),
}
SAFE_LABELS = {"safe"}
```