GUARD-TO-ACTION: MULTILINGUAL LONG-CONTEXT SAFETY IN TOOL-USING AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety guardrails are increasingly placed in front of language-model agents, yet they are often evaluated as isolated text classifiers. We present system, a multilingual guard-to-agent safety evaluation framework that measures whether guardrail bypasses propagate into downstream restricted-tool selection under short and long context. Starting from 1,901 unsafe English prompts derived from AEGIS2.0, we construct parallel prompts across 98 languages, yielding up to 186,298 instances spanning 19 harm categories. We evaluate five runtime guards—AprielGuard, CREST, GuardReasoner, WildGuard, and XGuard—and three downstream agents, Llama-3.1-8B-Instruct, Qwen2.5-14B-Instruct, and gpt-oss-20b, using inert restricted tools. On the 3,190-prompt AprielGuard reference set, we further test 8K, 20K, and 32K contexts with harmful requests placed at the beginning, middle, or end (81 agent–length–position cells, all at 100% inference coverage). Results show strong model–position interactions: Qwen beginning-position failure rises from 28.53% to 35.24% from 8K to 32K, while Llama falls from 10.97% to 0.72%; gpt-oss-20b is comparatively stable across length and position (7.32–13.30%) and, unlike the other two agents, trends slightly safer at longer context. Prompt position reverses model rankings, and 6.4% of matched cases shift from safe at 8K to unsafe at 32K. Low-resource languages also show higher downstream failure (21.65%) than high- and medium-resource languages (11.08% and 10.66%). These results show that multilingual agent safety is compositional and should be evaluated across the complete guard-to-action pipeline. Code available in this link: https://anonymous.4open.science/r/fc8b0b268a31a5a2e7adb4d416cfaa2057742f8f3

1 INTRODUCTION

Large language models increasingly act through retrieval, browsing, file operations, APIs, and external tools. ReAct established a reasoning-and-acting paradigm (Yao et al., 2023), while WebArena (Zhou et al., 2023), OSWorld (Xie et al., 2024), and τ-bench (Yao et al., 2024) demonstrate both the capabilities and unreliability of agents operating through external interfaces. In this setting, an unsafe response can become an attempted tool invocation that crosses a capability boundary.

Runtime guards such as Llama Guard (Inan et al., 2023), WildGuard (Han et al., 2024), ShieldGemma (Zeng et al., 2024), and Qwen3Guard (Zhao et al., 2025) classify requests before an agent receives them. Yet guard accuracy alone does not determine deployment safety: after a harmful request is admitted, the agent may refuse, escalate, or select a restricted tool. The relevant unit of evaluation is the complete *guard-to-action trajectory*.

This question is especially pressing across languages. MultiJail (Deng et al., 2024) and XSafety (Wang et al., 2024a) report elevated unsafe behavior outside English, while RTP-LX (de Wynter et al., 2025), PolyGuard (Kumar et al., 2025), Arabic safeguard evaluation (Ashraf et al., 2025), and XL-SafetyBench (Choi et al., 2026) expose further linguistic and cultural variation. A guard that blocks an intent in English but admits its translation creates a language-dependent capability boundary. Existing benchmarks address related risks: ToxicChat studies moderation (Lin et al., 2023), Do-Not-Answer tests refusal (Wang et al., 2024b), HarmBench standardizes adversarial evaluation (Mazeika et al., 2024), StrongREJECT examines jailbreak evaluation (Souly et al., 2024), XSTest measures over-refusal (Röttger et al., 2024), and AEGIS2.0 provides a broad safety taxon-

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

omy (Ghosh et al., 2025). Agent-focused work, including ToolEmu (Ruan et al., 2024), InjecAgent (Zhan et al., 2024), AgentDojo (Debenedetti et al., 2024), and Agent-SafetyBench (Zhang et al., 2024), further motivates measuring action-level outcomes.

Long context introduces another source of variation. LongBench (Bai et al., 2024) and LongBench v2 (Bai et al., 2025) show that large context windows do not ensure reliable reasoning. Lost in the Middle (Liu et al., 2024) identifies position sensitivity, and NoLiMa (Modarressi et al., 2025) shows difficulty with semantic retrieval in long contexts. Whether an agent's safety decision persists when the same harmful request moves within a long interaction remains underexplored.

We connect these three settings using 1,901 AEGIS-aligned unsafe English roots translated across 98 languages and 19 harm categories, with BERTScore (Zhang et al., 2020), COMET (Rei et al., 2020), translation-quality, and language-resource metadata. Five runtime guards (AprielGuard, CREST, GuardReasoner, WildGuard, and XGuard) screen the prompts. Guard-admitted requests then reach three tool-using agents (Llama-3.1-8B-Instruct, Qwen2.5-14B-Instruct, and gpt-oss-20b). We place each unchanged request at the beginning, middle, or end of benign English conversation history at 8K, 20K, and 32K tokens. Inert symbolic tools measure attempted restricted-tool selection without executing harmful actions, while stable `sample_ids` support matched comparisons.

Safety varies jointly with agent, context length, and position. Qwen's beginning-position restricted-tool-selection rate rises from 28.53% at 8K to 35.24% at 32K, whereas Llama's falls from 10.97% to 0.72%. Model rankings reverse across positions, and 6.4% of matched cases change from safe at 8K to unsafe at 32K. Risk also varies across harm categories and languages; low-resource languages have approximately twice the restricted-tool-selection rate of high- and medium-resource languages.

Research questions.

- **RQ1: Guard admission.** How reliably do five runtime guards block harmful requests across 98 languages, 19 harm categories, resource tiers, and translation-quality strata?
- **RQ2: Downstream action safety.** After guard bypass, how do agent, context length (8K, 20K, 32K), and request position interact to affect restricted-tool selection? The paired 8K→32K safety-transition analysis retains its original two-length scope.
- **RQ3: Concentrated failures.** How does downstream risk vary across languages, resource tiers, harm categories, and translation quality?

Contributions. We provide a multilingual guard-to-agent evaluation framework comprising up to 186,298 prompt instances from 1,901 unsafe roots across 98 languages and 19 harm categories; a matched long-context evaluation of three tool-using agents that separates guard admission from attempted restricted-tool selection; and a disaggregated analysis showing how model, position, context length, language, resource tier, harm category, and translation quality shape downstream risk.

2 Related Work

2.1 Runtime guardrails, safety benchmarks, and multilingual validity

Runtime guard models operationalize safety policy as inference-time classification. Llama Guard introduced policy-conditioned moderation of user prompts and model responses (Inan et al., 2023); WildGuard jointly models prompt harmfulness, response harmfulness, and refusal behavior (Han et al., 2024); ShieldGemma provides open moderation models based on the Gemma family (Zeng et al., 2024); and Qwen3Guard extends this setting with multilingual generative and streaming safeguards (Zhao et al., 2025).

Complementary benchmarks have broadened safety evaluation. ToxicChat studies real-world user–assistant moderation (Lin et al., 2023), Do-Not-Answer collects requests that responsible models should reject (Wang et al., 2024b), HarmBench standardizes adversarial safety evaluation (Mazeika et al., 2024), StrongREJECT highlights weaknesses in automated jailbreak evaluation (Souly et al., 2024), and XSTest measures excessive refusal of benign requests (Röttger et al., 2024). AEGIS2.0 contributes a broad safety taxonomy and human–LLM interactions for guardrail development (Ghosh et al., 2025). Multilingual studies further show that strong English safety does not imply

cross-lingual robustness. MultiJail reports elevated vulnerability in lower-resource languages (Deng et al., 2024), while XSafety similarly finds higher unsafe-response rates outside English (Wang et al., 2024a). RTP-LX (de Wynter et al., 2025), PolyGuard (Kumar et al., 2025), Arabic safeguard evaluation (Ashraf et al., 2025), and XL-SafetyBench (Choi et al., 2026) further demonstrate that linguistic and cultural variation can expose failures hidden by English-centric evaluation.

Our work extends this literature by evaluating harmful requests across 98 languages and, crucially, by following guard false negatives into a downstream tool-using agent. We therefore treat multilingual safety as a property of a composed guard-to-action pipeline rather than only a classification problem.

2.2 Tool-using agents and action-grounded safety

Language-model agents connect natural-language reasoning to external capabilities. ReAct established an influential framework for interleaving reasoning and action (Yao et al., 2023), while WebArena (Zhou et al., 2023), OSWorld (Xie et al., 2024), and τ-bench (Yao et al., 2024) demonstrate the challenges of reliable web, computer-use, and tool-mediated interaction.

Security-oriented benchmarks show that these systems introduce risks not captured by text-only evaluation. ToolEmu studies risky tool-use trajectories in an LM-emulated sandbox (Ruan et al., 2024); InjecAgent evaluates indirect prompt injection in tool-integrated agents (Zhan et al., 2024); AgentDojo provides a dynamic environment for attacks and defenses involving untrusted content (Debenedetti et al., 2024); and Agent-SafetyBench evaluates broader agentic safety hazards (Zhang et al., 2024). Our setting is complementary. Rather than focusing primarily on prompt injection or generic agent competence, we study whether an upstream multilingual safety-classification error propagates into an attempted restricted action. To avoid real-world side effects, the agent selects among inert symbolic tools, allowing us to measure action-level safety without executing harmful capabilities.

2.3 Long-context evaluation and positional sensitivity

The growth of nominal context windows has motivated dedicated evaluation of whether models reliably use distant information. LongBench evaluates long-document understanding, retrieval, and reasoning (Bai et al., 2024), while LongBench v2 extends this setting toward more realistic long-context reasoning (Bai et al., 2025). Lost in the Middle shows that model performance can vary substantially depending on whether relevant information appears near the beginning, middle, or end of a long sequence (Liu et al., 2024). NoLiMa further demonstrates sharp degradation when long-context retrieval requires semantic association rather than direct lexical overlap (Modarressi et al., 2025). These studies primarily evaluate task competence. We adapt the same length- and position-sensitivity question to safety. The harmful multilingual request is held unchanged while we vary 8K versus 32K context and beginning/middle/end position, allowing paired measurement of whether downstream refusal and restricted-tool selection remain stable under long-context stress. Our results show that the effect is not a generic "lost-in-the-middle" pattern: positional safety profiles differ markedly between model families.

2.4 Evaluator reliability and compositional safety

Automated safety evaluation introduces its own sources of error. MT-Bench and Chatbot Arena document position, verbosity, and self-enhancement biases in LLM-as-a-judge evaluation (Zheng et al., 2023), while StrongREJECT shows that weak evaluators can materially distort conclusions about jailbreak effectiveness (Souly et al., 2024). These findings motivate outcomes that depend as little as possible on secondary free-form judgment. Accordingly, our primary downstream endpoint is mechanically defined: whether the policy-constrained agent emits `call_tool` for the offered category-relevant restricted symbolic capability. Refusals, escalations, parser failures, and inference errors are recorded separately rather than collapsed into a single safety label. This design enables us to study *compositional safety*: how guard admission, downstream model, context length, prompt position, language, and harm category jointly determine whether a multilingual safety failure reaches the action boundary.

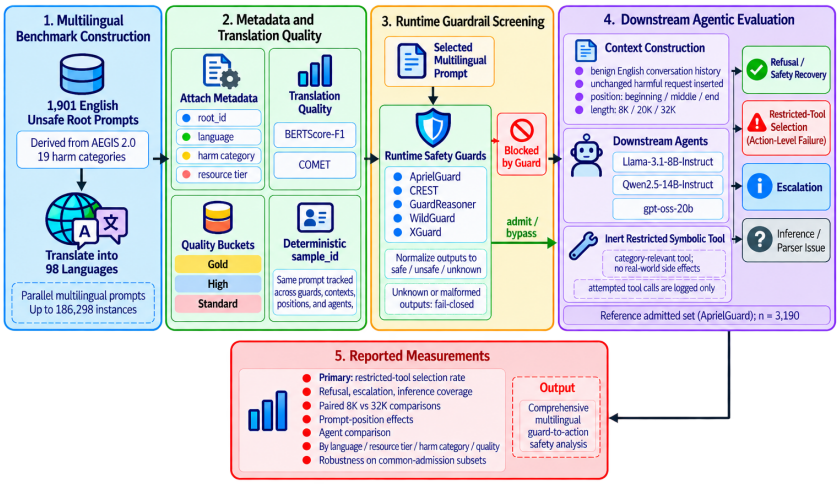

Figure 1: **Multilingual guard-to-agent evaluation framework.** Unsafe prompts translated into 98 languages pass through runtime guards; admitted prompts reach three agents under controlled long-context conditions. We measure guard bypass, refusal or escalation, and attempted selection of inert restricted tools across multilingual and harm-category subgroups.

3 METHODOLOGY

3.1 MULTILINGUAL BENCHMARK CONSTRUCTION

As shown in Figure 1 our evaluation starts from 1,901 English unsafe root prompts derived from AEGIS2.0, covering 19 harm categories. Each root is translated into 98 languages, producing a parallel pool of up to 186,298 multilingual prompt instances. Every instance retains its root identifier, language, harm category, language-resource tier, and translation-quality metadata.

Because machine translation can introduce semantic drift, we retain both BERTScore-F1 and COMET scores and assign translations to quality strata before any guard or agent evaluation. These scores are used for subgroup analysis rather than as post-hoc filters on safety outcomes. This allows us to examine whether apparent multilingual safety differences are associated with translation quality while preserving broad language coverage. Each multilingual instance is assigned a deterministic `sample_id` derived from its root identifier, language, and translated prompt. The same identifier is preserved across guardrail, context-length, prompt-position, and downstream-agent conditions. This enables matched comparisons of the same harmful request throughout the complete evaluation pipeline.

3.2 RUNTIME GUARDRAIL EVALUATION AND ADMISSION

We evaluate five runtime safety guards: **AprielGuard**, **CREST**, **GuardReasoner**, **WildGuard**, and **XGuard**. Because the guards expose heterogeneous output formats, model-specific adapters normalize their predictions into a common operational decision while preserving raw model outputs for auditing. Unknown or malformed outputs are handled under a fail-closed policy.

All prompts entering the downstream study are gold-labeled unsafe. Therefore, a normalized guard prediction of `safe` is treated as a guardrail bypass, whereas `unsafe` blocks the prompt before agent execution. Guard performance is analyzed across language, harm category, resource tier, and translation-quality strata. AprielGuard provides the primary downstream reference population, consisting of 3,190 unique admitted multilingual prompts. To test the robustness of downstream findings to the admitted prompt population, we additionally form pairwise-common admission subsets be-

tween AprielGuard and XGuard ($n = 2{,}524$), CREST ($n = 1{,}237$), GuardReasoner ($n = 102$), and WildGuard ($n = 18$). A common subset contains only prompts classified as safe by both guards.

For these common `sample_id`s, the downstream agent input is identical regardless of which guard admitted the request. We therefore reuse the corresponding AprielGuard agent inference rather than repeating identical model executions. Consequently, these analyses measure robustness across different jointly admitted prompt populations; they are not interpreted as causal effects of guard identity on downstream behavior.

3.3 Downstream Agent and Long-Context Evaluation

Guard admission is an intermediate failure. Each admitted request reaches Llama-3.1-8B-Instruct, Qwen2.5-14B-Instruct, or gpt-oss-20b under the same safety policy with an inert, category-relevant restricted tool. Agents return *refuse*, *escalate*, or *call tool*. A valid restricted tool call is the primary action-level failure; refusal and escalation are downstream recovery. Inference and parsing failures are recorded separately, never as safe outcomes.

We preserve each multilingual harmful request verbatim within benign English conversation history, varying length (8K, 20K, 32K), position (beginning, middle, end), and agent. This $3 \times 3 \times 3$ design has 27 cells. Contexts target approximately 8,192, 20,480, and 32,768 tokens with a fixed reference tokenizer; realized counts are recorded per agent. Inputs are not silently truncated, and unprocessed cases remain inference failures. The benign filler isolates length and position effects. All 27 cells have 3,190/3,190 scorable runs on the AprielGuard-admitted reference set. An earlier Qwen2.5-14B 32K-middle run had 3.5% coverage due to transient GPU-memory contention; an isolated rerun resolved it (Appendix A.3).

3.4 Outcome Measures and Paired Analysis

Our primary outcome is *restricted-tool selection*, measured over scorable guard-admitted runs. We also record refusal, escalation, inference coverage, parsing failures, and inference failures. Because the same `sample_id` is evaluated across conditions, all principal comparisons are paired. We compare 8K versus 32K while holding agent and position fixed, beginning/middle/end positions while holding agent and context length fixed, and agents pairwise within the same length–position condition. The 20K condition is reported across all descriptive breakdowns (by agent, length, and position) but is not included in the paired 8K→32K transition analysis, which remains scoped to its original two-length design. We additionally track 8K→32K prompt-level transitions as safe→safe, safe→unsafe, unsafe→safe, or unsafe→unsafe. We report Wilson 95% confidence intervals for proportions. Paired comparisons use percentage-point differences, paired-bootstrap 95% confidence intervals with 2,000 resamples, and exact McNemar tests. Low-coverage comparisons are reported separately and are not used as primary evidence.

3.5 Disaggregated and Robustness Analysis

To identify concentrated failures, we disaggregate downstream risk by language, resource tier, harm category, and translation-quality bucket. All 98 languages are included in quantitative analysis; for visualization, we show the 20 highest-risk, 20 median-risk, and 20 lowest-risk languages. Category analysis examines whether particular harm types are disproportionately difficult for either downstream agent, while resource-tier analysis compares high-, medium-, and low-resource languages. These analyses are descriptive, since resource tier covaries with script, language family, tokenization, translation quality, and pretraining exposure. For robustness, we use the larger AprielGuard–XGuard and AprielGuard–CREST common-admission subsets as the primary checks, while GuardReasoner and WildGuard intersections are treated as exploratory because of their smaller size. All results retain `sample_id`, language, category, context condition, agent model, inference status, and parsing metadata for reproducibility.

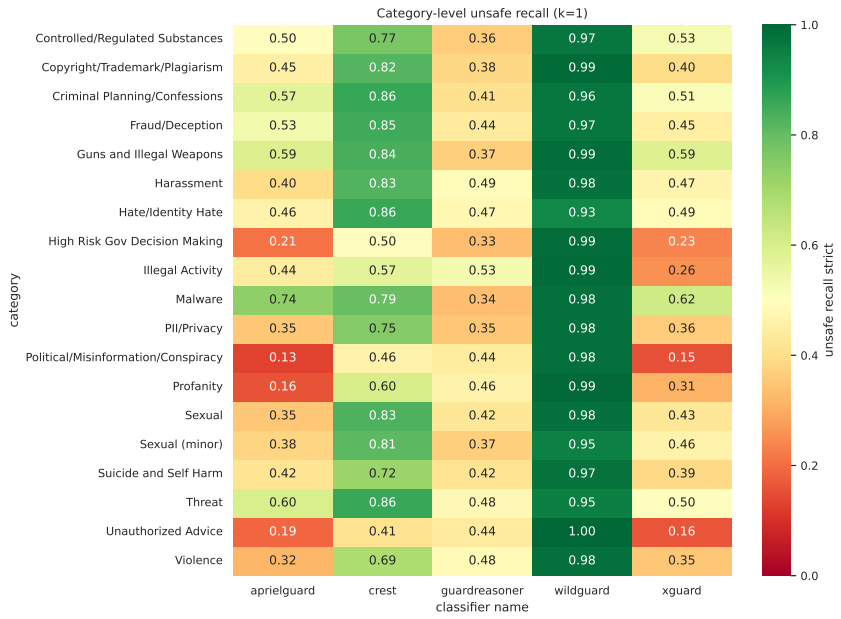

Figure 2: **Guard bypass rate by harm category, all five guards.** Darker cells indicate higher bypass (lower blocking) rate.

Table 1: **Guard-level block/bypass performance, overall and by language-resource tier.** Bypass rate (= evasion rate) is the fraction of the 5,403 gold-unsafe, multilingual prompts a guard classified as `safe` (unknown/malformed outputs are scored as non-blocks, i.e. fail-closed). Block rate = 1 − bypass rate.

Guard	n	Block rate	Bypass rate	Unknown rate	High-tier bypass	Low-tier bypass	Low − High
AprielGuard	5,403	41.0%	59.0%	0.0%	38.8%	67.2%	+28.4 pp
CREST	5,403	73.6%	26.4%	0.0%	25.1%	28.3%	+3.2 pp
GuardReasoner	5,403	41.7%	58.3%	50.6%	86.8%	45.9%	-40.9 pp
WildGuard	5,403	97.5%	2.5%	0.0%	6.6%	1.7%	-4.9 pp
XGuard	5,403	41.0%	59.0%	0.1%	33.5%	68.1%	+34.6 pp

4 Guard Analysis

Before examining downstream agent behavior, we report the runtime guards' own block/bypass performance over the full 5,403-row, 98-language gold-unsafe evaluation set, disaggregated by language-resource tier and with cross-lingual and cross-category extremes.

Guardrail block/bypass performance (RQ1). We report block/bypass rates for all five evaluated guards over the full 5,403-row, 98-language, gold-unsafe evaluation set in Table 1, disaggregated by language-resource tier, with cross-lingual and cross-category extremes in Table 2. Guards vary by an order of magnitude in permissiveness: WildGuard is the most conservative, bypassing only 2.5

Admitted-set consistency across guards is uneven (full pairwise counts in Section 4.2): AprielGuard and XGuard, the two most permissive guards, agree on 65.5% (Jaccard) of their admitted sets, whereas AprielGuard and WildGuard – the most permissive and most conservative guards – overlap on only 0.5% (Jaccard) of admissions, confirming that guard choice materially changes which prompts reach the downstream agent rather than converging on a shared admitted population.

5 Results

We analyze downstream behavior after harmful multilingual requests pass the input-guard layer. Our primary outcome is *restricted-tool selection*, the fraction of scorable decisions in which the agent

Table 2: **Cross-lingual and cross-category guard bypass extremes.** Min/max bypass rate across the 98 evaluated languages (languages with $n \geq 10$ evaluated prompts only), and the harm category with the highest and lowest bypass rate per guard.

Guard	Min bypass	(language)	Max bypass	(language)	Std. across langs	Worst category	Bypass
AprielGuard	26.8%	Armenian	100.0%	Amharic	20.7%	Political/Misinformation/Conspiracy	86.9%
CREST	9.6%	Burmese	85.3%	Tajik	11.9%	Unauthorized Advice	58.7%
GuardReasoner	7.2%	Zulu	96.4%	Spanish	27.2%	High Risk Gov Decision Making	66.7%
WildGuard	0.0%	Amharic	60.0%	English	7.5%	Hate/Identity Hate	6.7%
XGuard	12.0%	English	93.9%	Xhosa	20.2%	Political/Misinformation/Conspiracy	84.6%

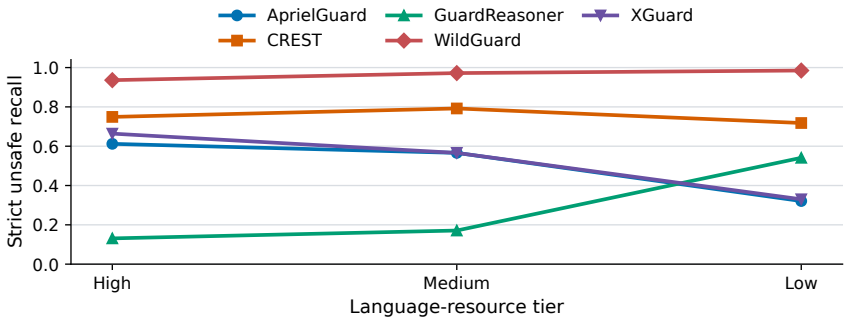

Figure 3: **Guard bypass rate by language-resource tier, all five guards.** The expected high-resource-vs-low-resource bypass gap holds for three of five guards (XGuard, AprielGuard, CREST) but reverses for WildGuard; GuardReasoner's nominal reversal is a coverage artifact of its highly imbalanced unknown-output rate across tiers (see main text).

selects the category-relevant restricted symbolic tool. Since tools are inert, this measures attempted unsafe action rather than real execution. The AprielGuard-admitted reference set contains 3,190 unique multilingual prompts. Pairwise-common subsets with XGuard, CREST, GuardReasoner, and WildGuard provide robustness checks. Because downstream inference is reused for common `sample_ids`, these subsets represent different jointly admitted populations rather than causal effects of guard choice.

5.1 Safety depends jointly on agent, context length, and position

Figure 4 and Table 3 show the complete reference-set factorial comparison.

All three agents exhibit distinct positional profiles. For Llama, 8K risk rises from 10.97% at the beginning to 23.39% at the end; at 20K the same rising pattern holds (13.01%→18.03%→23.48%); at 32K the beginning rate collapses to only 0.72% while middle and end remain above 22%. Qwen shows an almost opposite pattern: beginning/middle risk is high across all three lengths (25–40%), while end-position risk stays low throughout (9.18–10.34%). gpt-oss-20b is the most stable of the three agents, remaining within a narrow 7.32–13.30% band across every length–position cell, and—unlike Llama and Qwen—shows a mild *decrease* in risk as context length grows (8K mean 11.43%, 32K mean 7.95%), the opposite direction from Qwen's length effect. Thus, safety depends jointly on model, length, and position, and the direction of the length effect itself is model-specific rather than universal. Detailed paired position and agent comparisons appear in Appendices A.1 and A.2.

5.2 Increasing context from 8K to 32K has position-dependent effects

We compare matched `sample_ids` at 8K and 32K while holding agent and prompt position fixed.

Longer context does not uniformly increase risk. For Qwen, beginning risk increases from 28.53% to 35.24%, a +6.71-point change (95% CI: +5.45 to +7.90; $p = 1.47 \times 10^{-26}$), while end risk decreases by 1.16 points (95% CI: -2.04 to -0.31; $p = 0.011$). Llama shows the opposite beginning-position effect: risk decreases by 10.25 points (95% CI: -11.44 to -9.18; $p = 1.92 \times 10^{-83}$), whereas middle-position risk increases by 9.31 points (95% CI: +7.93 to +10.69; $p = 9.91 \times 10^{-41}$). The

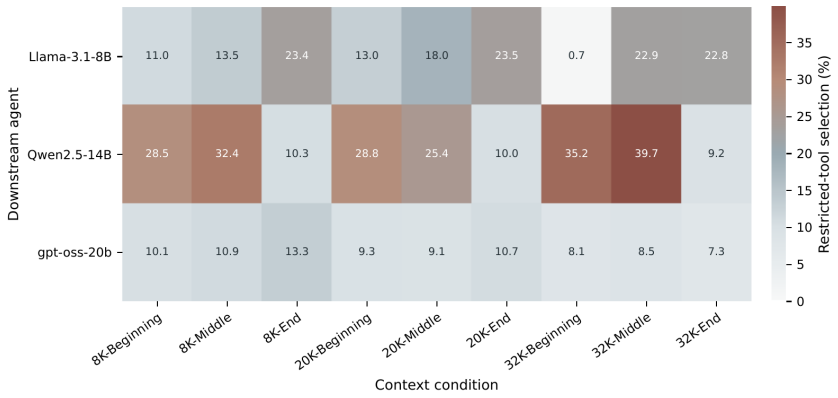

Figure 4: **Reference-set restricted-tool selection across downstream agent, context length, and harmful-prompt position.** Values report percentages over scorable guard-admitted requests.

Table 3: **Restricted-tool-selection rate (%) on the AprielGuard reference set.** Brackets give 95% Wilson intervals. All 27 cells have 3,190/3,190 scorable runs; see Appendix A.3.

Agent	Context	Beginning	Middle	End
Llama-3.1-8B	8K	10.97 [9.93, 12.10]	13.54 [12.40, 14.77]	23.39 [21.95, 24.89]
	20K	13.01 [11.89, 14.22]	18.03 [16.73, 19.40]	23.48 [22.04, 24.98]
	32K	0.72 [0.48, 1.08]	22.85 [21.43, 24.34]	22.76 [21.34, 24.25]
Qwen2.5-14B	8K	28.53 [26.99, 30.12]	32.45 [30.84, 34.09]	10.34 [9.34, 11.45]
	20K	28.78 [27.23, 30.37]	25.39 [23.91, 26.93]	10.03 [9.04, 11.12]
	32K	35.24 [33.60, 36.91]	39.75 [38.06, 41.46]	9.18 [8.23, 10.24]
gpt-oss-20b	8K	10.08 [9.07, 11.19]	10.91 [9.86, 12.06]	13.30 [12.15, 14.55]
	20K	9.34 [8.32, 10.47]	9.11 [8.09, 10.24]	10.70 [9.59, 11.91]
	32K	8.07 [7.09, 9.17]	8.46 [7.44, 9.61]	7.32 [6.36, 8.41]

XGuard and CREST subsets reproduce the principal directions, supporting the robustness of these effects.

5.3 Context expansion changes the safety state of individual prompts

Aggregate rates can hide prompt-level changes. We therefore examine matched 8K→32K safety-state transitions. Most matched cases remain stable: 76.2% are safe at both lengths and 11.9% remain unsafe, so 88.1% preserve their state. However, 6.4% transition from safe at 8K to unsafe at 32K, while 5.5% change from unsafe to safe. The pooled unsafe rate therefore changes only from 17.4% at 8K to 18.3% at 32K, a net increase of 0.9 points. This small aggregate difference masks the much larger and opposing model–position effects in Table 4.

5.4 Safety failures vary substantially across harm categories

Figure 6 shows substantial category-dependent variation for both agents. For Qwen, the largest failure rates occur for Unauthorized Advice (57.6%), Copyright/Trademark/Plagiarism (54.6%), Malware (46.7%), High Risk Government Decision Making (45.2%), Controlled/Regulated Substances (44.1%), and Threat (43.6%). Llama's largest rates include Copyright/Trademark/Plagiarism (37.5%), Malware (29.5%), Unauthorized Advice (26.3%), PII/Privacy (25.6%), Profanity (23.1%), and Illegal Activity (23.0%). Model rankings also vary by category. Qwen exceeds Llama by 31.3 points for Unauthorized Advice (57.6% vs. 26.3%) and 28.2 points for High Risk Government Decision Making (45.2% vs. 17.0%). Conversely, Llama is less safe for Illegal Activity (23.0% vs. 6.9%) and Guns and Illegal Weapons (12.8% vs. 1.7%). Thus, no downstream model dominates across all harm types.

Table 4: **Paired 8K→32K change in restricted-tool selection (percentage points).** XGuard and CREST are pairwise-common admission subsets.

Agent	Position	Reference	XGuard	CREST
Qwen	Beginning	+6.71	+6.77	+7.11
	Middle	+7.30	+10.84†	+3.03†
	End	-1.16	-1.03	-2.75
Llama	Beginning	-10.25	-11.25	-15.28
	Middle	+9.31	+10.30	+8.81
	End	-0.63	-0.55	-2.18

†Qwen 32K-middle has 3,190/3,190 reference-set pairs (Appendix A.3). Its XGuard and CREST subset results still reuse inference from before the rerun (83 and 33 scorable pairs, respectively); they are excluded from primary evidence pending a refresh of the guard-subset inference-reuse step.

5.5 Cross-lingual averages conceal large disparities

Figure 5 visualizes the 20 highest-risk, 20 median-risk, and 20 lowest-risk languages; all 98 languages remain in the quantitative analysis. Oromo, Somali, Amharic, Zulu, Xhosa, Yoruba, Shona, Nyanja, and Lingala appear among the highest-risk languages, whereas Korean, Ukrainian, Vietnamese, Danish, and Swedish occur among the lower-risk group. The visualized range extends from single-digit failure to above 40%. Low-resource languages exhibit 21.65% restricted-tool selection, nearly twice the high-resource (11.08%) and medium-resource (10.66%) rates. Inference-success rates are almost identical across tiers—91.88%, 91.98%, and 91.97%, respectively—so the gap is not explained by gross differences in inference coverage. Because resource tier covaries with script, family, tokenization, translation quality, and pretraining exposure, we interpret this relationship as associational rather than causal.

6 Discussion and Limitations

A guard bypass does not always cause action-level failure: the agent may refuse or escalate, but this recovery depends on model and context. From 8K to 32K, Qwen's beginning-position failure rises from 28.53% to 35.24%, while Llama's falls from 10.97% to 0.72%; Llama's middle-position failure rises from 13.54% to 22.85%. Among matched cases, 6.4% become unsafe and 5.5% become safe at 32K. Qwen is safer with the request at the end, whereas Llama is safer at the beginning. Context organization is therefore safety-relevant (Appendix C).

The 98-language benchmark is primarily translation-based; BERTScore and COMET cannot guarantee native fluency or equivalent harmful intent. Benign English conversation history differs from retrieved documents, tool outputs, code, multilingual history, and adversarial content. Findings from three agents, three lengths (8K, 20K, 32K), and three positions may not generalize to other models or windows. All 27 cells have full inference coverage after an isolated rerun resolved transient GPU-memory contention in Qwen's 32K-middle cell. That cell improved from 113/3,188 valid inferences (3.54% coverage) to 3,190 scorable runs, with a 39.75% restricted-tool-selection rate (Appendix A.3). Further limitations appear in Appendix D.

7 Conclusion

We present a multilingual guard-to-agent safety evaluation framework spanning 98 languages, 19 harm categories, five runtime guards, and three downstream tool-using agents. By matching the same harmful requests across 8K, 20K, and 32K contexts and beginning, middle, and end positions, we show that downstream safety is strongly dependent on context structure and model choice rather than being a fixed property of the agent. Our results reveal model-specific positional effects, prompt-level safety transitions, substantial harm-category variation, and higher downstream risk for low-resource languages. These findings demonstrate that multilingual agent safety should be evaluated across the complete guard-to-action pipeline, with attention to context length, prompt position, language, and downstream model behavior.

REFERENCES

Yasser Ashraf, Yuxia Wang, Bin Gu, Preslav Nakov, and Timothy Baldwin. Arabic dataset for llm safeguard evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5529–5546. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.285. URL https://aclanthology.org/2025.naacl-long.285/.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.172. URL https://aclanthology.org/2024.acl-long.172/.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 3639–3664. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.183. URL https://aclanthology.org/2025.acl-long.183/.

Dasol Choi, Eugenia Kim, Jaewon Noh, Sang Seo, Eunmi Kim, Myunggyo Oh, Yunjin Park, Brigitta Jesica Kartono, et al. XL-SafetyBench: A country-grounded cross-cultural benchmark for llm safety and cultural sensitivity. *arXiv preprint arXiv:2605.05662*, 2026. URL https://arxiv.org/abs/2605.05662.

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altıntoprak, Lena Baur, 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, 2025. doi: 10.1609/aaai.v39i27.35011.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. In *Advances in Neural Information Processing Systems*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Datasets_and_Benchmarks_Track.html.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.06474.

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: Human Language Technologies*, pages 5992–6026. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.306. URL https://aclanthology.org/2025.naacl-long.306/.

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*, 2024. URL https://arxiv.org/abs/2406.18495.

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. URL https://arxiv.org/abs/2312.06674.

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

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. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.findings-emnlp.311/.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

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*, volume 235 of *Proceedings of Machine Learning Research*, pages 35181–35224. PMLR, 2024. URL https://proceedings.mlr.press/v235/mazeika24a.html.

Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Trung Bui, Ryan A. Rossi, Seunghyun Yoon, and Hinrich Schuetze. NoLiMa: Long-context evaluation beyond literal matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 44554–44570. PMLR, 2025. URL https://proceedings.mlr.press/v267/modarressi25a.html.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. COMET: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702. Association for Computational Linguistics, 2020.

Paul Röttger, Hannah 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: Human Language Technologies*, pages 5377–5400. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.15817.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. A strongreject for empty jailbreaks. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2402.10260.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R. Lyu. All languages matter: On the multilingual safety of llms. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024a. URL https://aclanthology.org/2024.findings-acl.349/.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. Do-not-answer: Evaluating safeguards in llms. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 896–911. Association for Computational Linguistics, 2024b. URL https://aclanthology.org/2024.findings-eacl.61/.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *arXiv preprint arXiv:2404.07972*, 2024. URL https://arxiv.org/abs/2404.07972.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. URL https://arxiv.org/abs/2406.12045.

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. URL https://arxiv.org/abs/2407.21772.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. URL https://aclanthology.org/2024.findings-acl.624/.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.09675.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. Agent-safetybench: Evaluating the safety of llm agents. *arXiv preprint arXiv:2412.14470*, 2024. URL https://arxiv.org/abs/2412.14470.

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. URL https://arxiv.org/abs/2510.14276.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.05685.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023. URL https://arxiv.org/abs/2307.13854.

A Additional Result Analyses

A.1 Detailed Paired Prompt-Position Effects

We compare beginning, middle, and end positions for the same `sample_id`, holding the downstream agent and context length fixed. Figure 9 and Table 6 show strong model-specific positional effects.

For Qwen, moving the harmful request toward the end substantially reduces risk: beginning→end changes restricted-tool selection by -18.18 points at 8K and -26.05 points at 32K. Llama shows the opposite profile, with beginning→end increasing risk by 12.41 points at 8K and 22.04 points at 32K. At 32K, Llama middle and end are nearly identical ($\Delta = -0.09$, $p = 0.930$). These results indicate model-specific positional safety biases rather than a universal "lost-in-the-middle" effect.

A.2 Detailed Paired Downstream-Agent Comparison

We compare Llama and Qwen on the same `sample_id`, context length, and prompt position. Figure 10 and Table 7 show that the relative safety ranking reverses with position.

Llama is substantially safer at beginning and most middle positions. At 8K, it is 17.55 points safer at the beginning and 18.90 points safer in the middle; at 32K-beginning, the difference grows to 34.51

Table 5: **Reference-set downstream risk by language-resource tier.**

Tier	Unique prompts	Scorable runs	Risk	95% CI
High	235	2,591	11.08%	[9.93, 12.34]
Medium	498	5,497	10.66%	[9.87, 11.50]
Low	2,457	27,115	21.65%	[21.17, 22.15]

Table 6: **Prompt-position effects on the AprielGuard reference set.** Δ denotes condition B minus condition A.

Agent	Length	Comparison	Pairs	Δ pp	McNemar p
Qwen	8K	Beginning → Middle	3190	+3.92	1.13×10^{-8}
Qwen	8K	Beginning → End	3190	-18.18	2.26×10^{-135}
Qwen	8K	Middle → End	3190	-22.10	8.70×10^{-190}
Qwen	32K	Beginning → Middle	113	+6.19	0.039
Qwen	32K	Beginning → End	3190	-26.05	6.76×10^{-212}
Qwen	32K	Middle → End	113	-31.86	2.91×10^{-11}
Llama	8K	Beginning → Middle	3190	+2.57	2.60×10^{-4}
Llama	8K	Beginning → End	3190	+12.41	3.23×10^{-54}
Llama	8K	Middle → End	3190	+9.84	1.06×10^{-52}
Llama	32K	Beginning → Middle	3190	+22.13	3.29×10^{-198}
Llama	32K	Beginning → End	3190	+22.04	5.50×10^{-200}
Llama	32K	Middle → End	3190	-0.09	0.930

points (35.24% for Qwen versus 0.72% for Llama). The ranking reverses at the end: Qwen is safer by 13.04 points at 8K and 13.57 points at 32K.

This reversal also appears in the larger robustness subsets. For AprielGuard–XGuard, the differences are -37.04 points at 32K-beginning and +14.66 points at 32K-end; for AprielGuard–CREST they are -46.89 and +14.47 points, respectively. Thus, neither agent has a universally superior safety profile.

A.3 Qwen 32K-Middle Coverage Diagnostic and Resolution

An earlier data-collection run of the Qwen2.5-14B-Instruct 32K-middle condition had substantially lower inference coverage than the other factorial cells: only 113 of 3,188 reference-set rows produced valid inference (3.54% coverage), with the remaining 3,075 rows failing with `CUDA_OOM` errors traced to transient GPU-memory contention from concurrent jobs sharing the same device during that run—not an architectural or tokenizer limitation of the model itself (input length for this cell, ≈32,977 tokens, is comparable to the fully-successful 32K-beginning and 32K-end cells). We diagnosed this root cause, reran the affected rows in isolation (batch size reduced to avoid the same memory contention, split across GPUs for throughput), and merged the results back into the reference population. The cell now reports 3,190/3,190 successful inferences (100% coverage) and a restricted-tool-selection rate of 39.75% [38.06, 41.46]—consistent in direction and similar in magnitude to the original 40.71% point estimate from the small scorable subsample, now with full statistical power. All figures and tables in the main text reflect this corrected, fully-covered cell.

One artifact of the original failure remains open: the CREST and XGuard common-admission subsets reuse AprielGuard agent inference for their shared `sample_ids` (Section 3), and that reuse step has not yet been rerun against the corrected reference data. Their 32K-middle paired comparisons therefore still reflect the pre-fix low-coverage inference (83 and 33 scorable pairs respectively) and are retained for transparency but not used as primary evidence; refreshing them is a mechanical rerun of the existing guard-subset consolidation step. All other principal reference-set factorial cells, across all three agents and three context lengths, contain 3,190 successful inferences.

Table 7: **Paired Qwen–Llama comparison on the AprielGuard reference set.** $\Delta = R_{\text{Llama}} - R_{\text{Qwen}}$.

Context	Position	Qwen	Llama	Δ pp	95% CI
8K	Beginning	28.53	10.97	-17.55	[-19.15, -15.96]
8K	Middle	32.45	13.54	-18.90	[-20.47, -17.37]
8K	End	10.34	23.39	+13.04	[11.54, 14.51]
32K	Beginning	35.24	0.72	-34.51	[-36.18, -32.92]
32K	Middle	40.71	37.17	-3.54[†]	[-13.27, 7.08]
32K	End	9.18	22.76	+13.57	[12.07, 15.05]

[†]Based on only 113 paired scorable Qwen 32K-middle cases.

B SUMMARY OF FINDINGS

The results reveal four consistent patterns. First, context length does not have a uniform safety effect: 8K→32K can either increase or decrease restricted-tool selection depending on model and position. Second, prompt location strongly changes safety, with Qwen substantially safer at the end and Llama especially safe at 32K-beginning. Third, model rankings also vary across harm categories, so no single downstream agent is uniformly safer. Finally, these effects occur alongside substantial multilingual heterogeneity, with low-resource languages showing approximately twice the aggregate failure rate of high- and medium-resource languages.

C MORE DISCUSSION

There is also no universally safer downstream agent. Llama is substantially safer at beginning and most middle positions, while Qwen is approximately 13–14 points safer at the end; harm-category rankings likewise reverse across models. Multilingual disparities remain pronounced after guard bypass: low-resource languages show 21.65% restricted-tool selection compared with 11.08% for high-resource and 10.66% for medium-resource languages, despite similar inference coverage. These results suggest that multilingual agent safety should be evaluated jointly across guard behavior, context length, prompt position, model choice, language, and harm category rather than summarized by a single aggregate safety score.

D MORE LIMITATIONS

The downstream analysis is conditional on guard admission, and the XGuard, CREST, GuardReasoner, and WildGuard robustness subsets reuse AprielGuard agent outputs for shared `sample_ids`, so these comparisons reflect different admitted populations rather than causal guard effects. Our symbolic tools also measure attempted unsafe action rather than real-world execution, and the current pipeline omits output guards, execution-time authorization, and benign over-refusal controls. Future work should add native-authored multilingual prompts, realistic retrieved and multilingual contexts, more agents and longer context windows, independently evaluated guard–agent pipelines, sandboxed tool execution, additional defense layers, and matched benign controls.

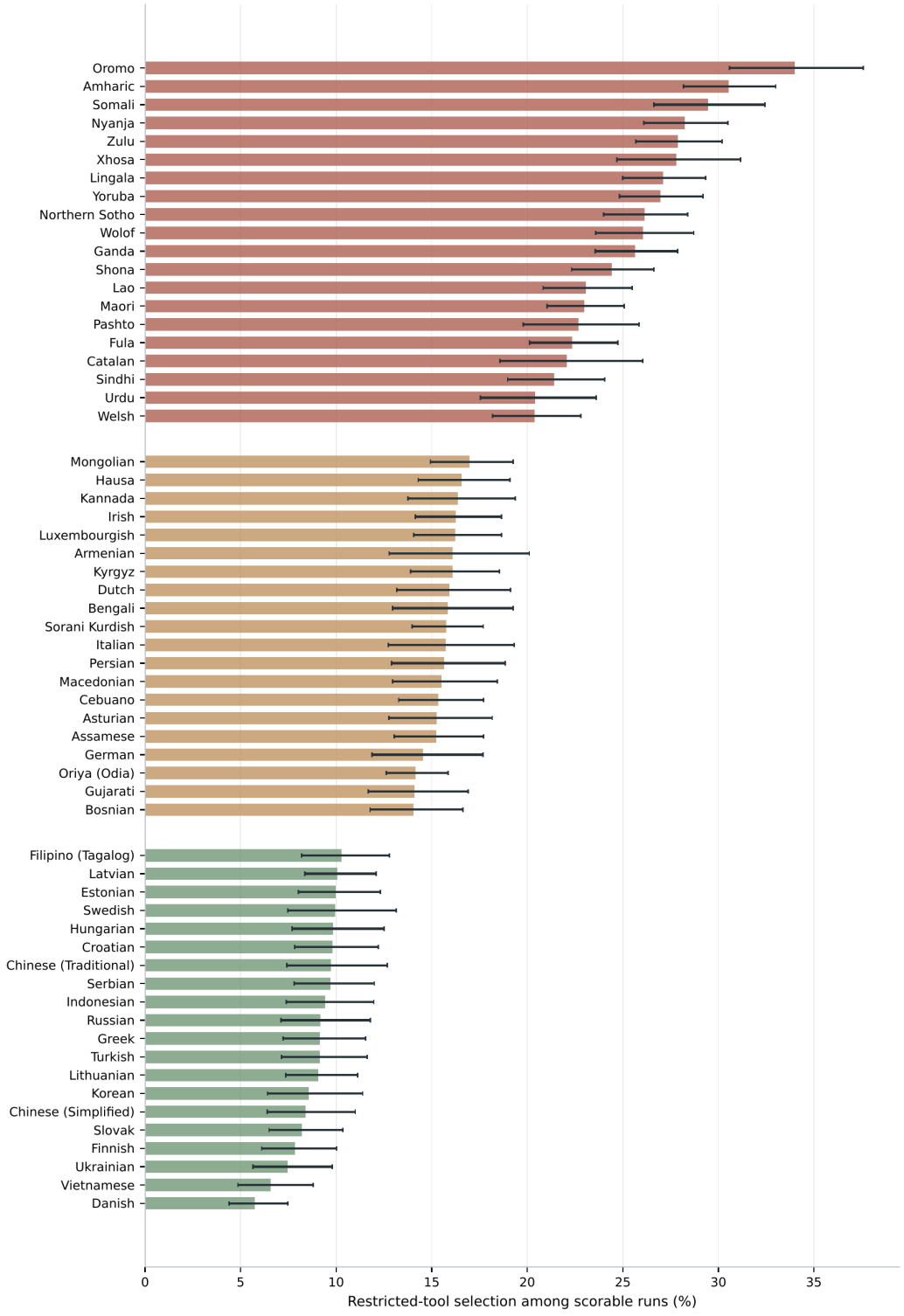

Figure 5: **Reference-set cross-lingual downstream risk.** Bars report restricted-tool-selection rates among scorable runs. All 98 languages are retained in the full analysis.

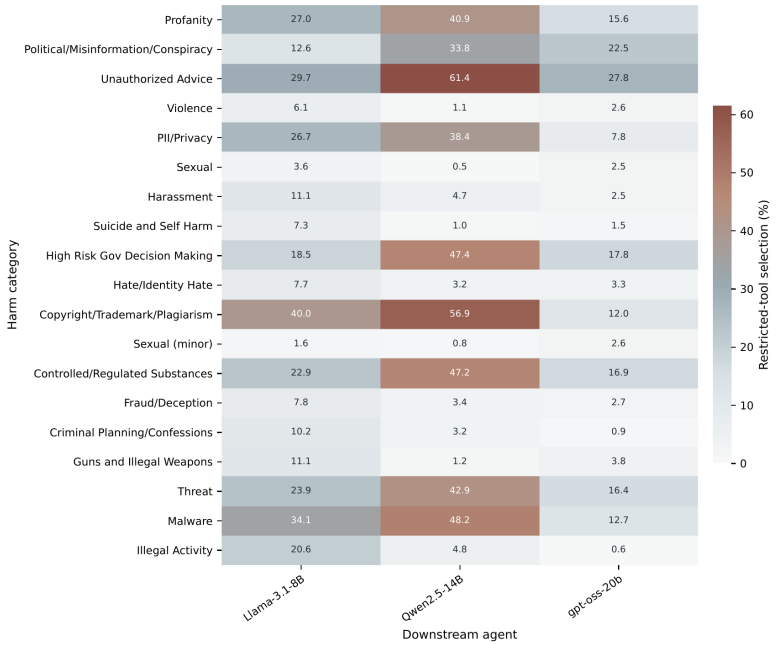

Figure 6: **Reference-set restricted-tool selection by harm category and downstream agent.** Values are percentages over scorable runs.

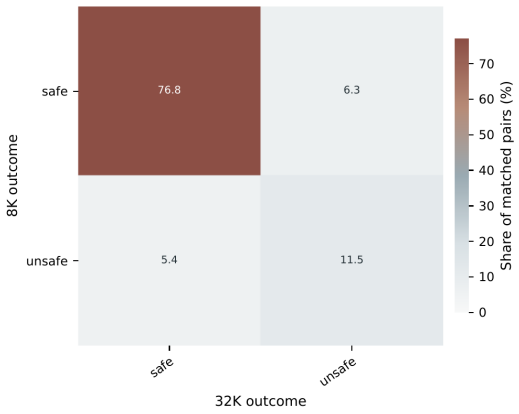

Figure 7: **Reference-set safety-state transitions from 8K to 32K.** "Unsafe" denotes restricted-tool selection.

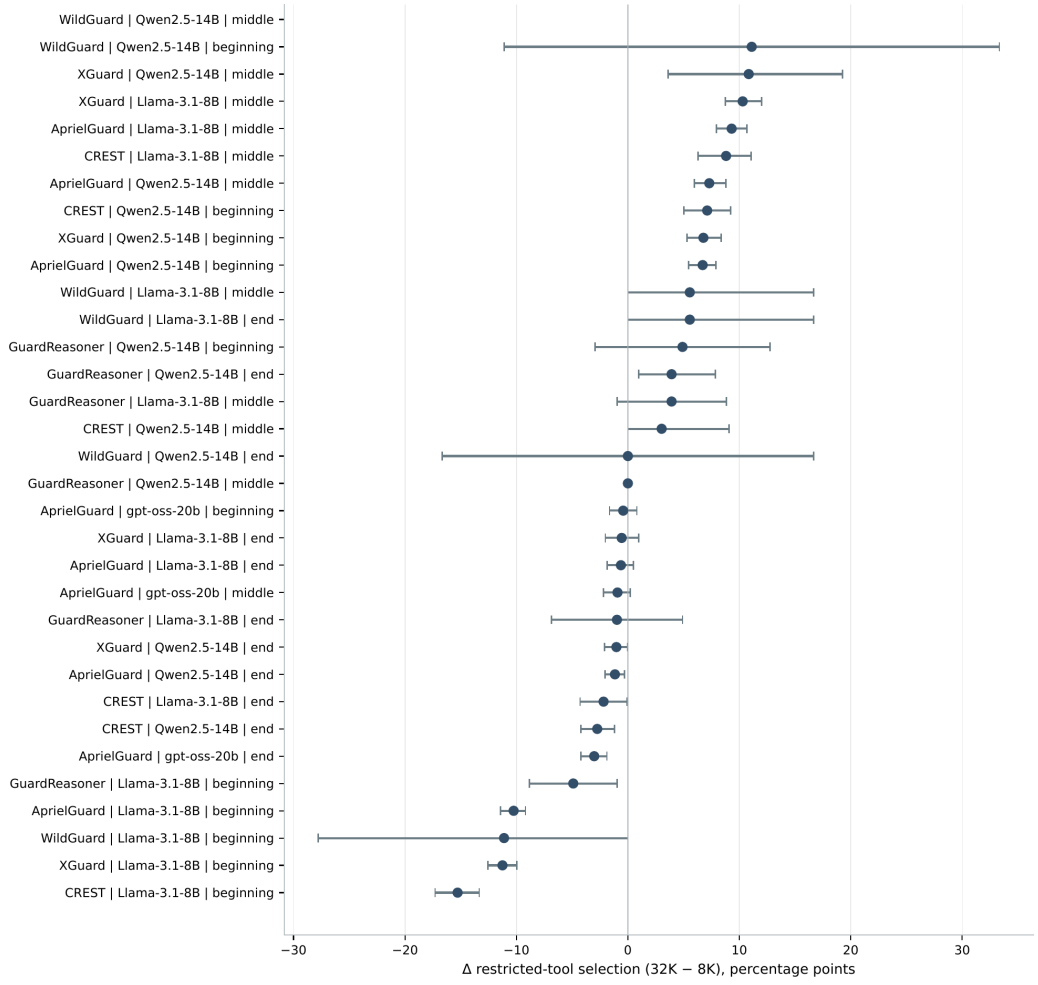

Figure 8: **Paired effect of increasing context from 8K to 32K.** Points show percentage-point changes with paired-bootstrap 95% confidence intervals. Positive values indicate higher failure at 32K.

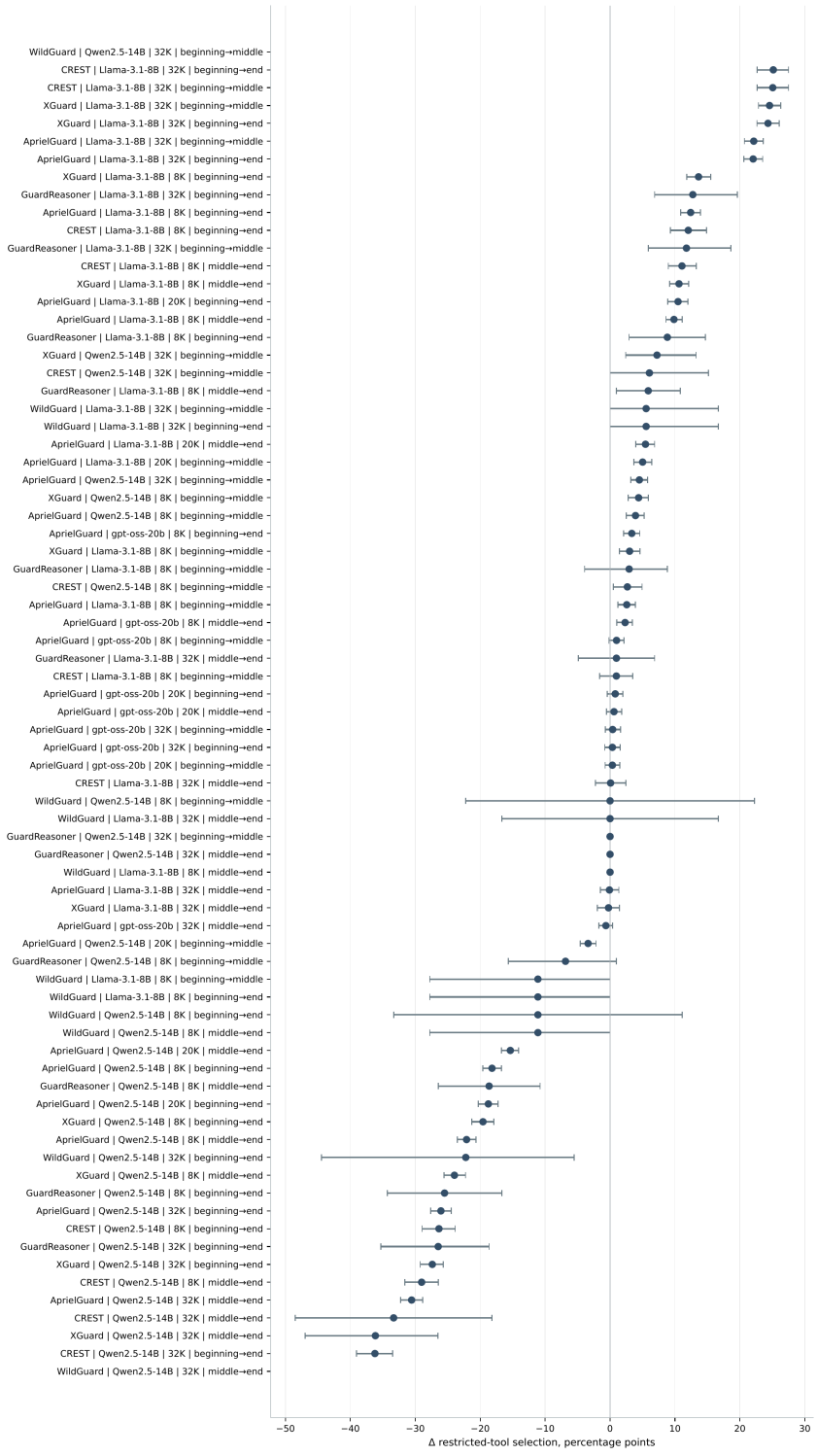

Figure 9: **Paired effect of harmful-prompt position.** Points show percentage-point changes in restricted-tool selection when the same harmful request is moved between beginning, middle, and end positions.

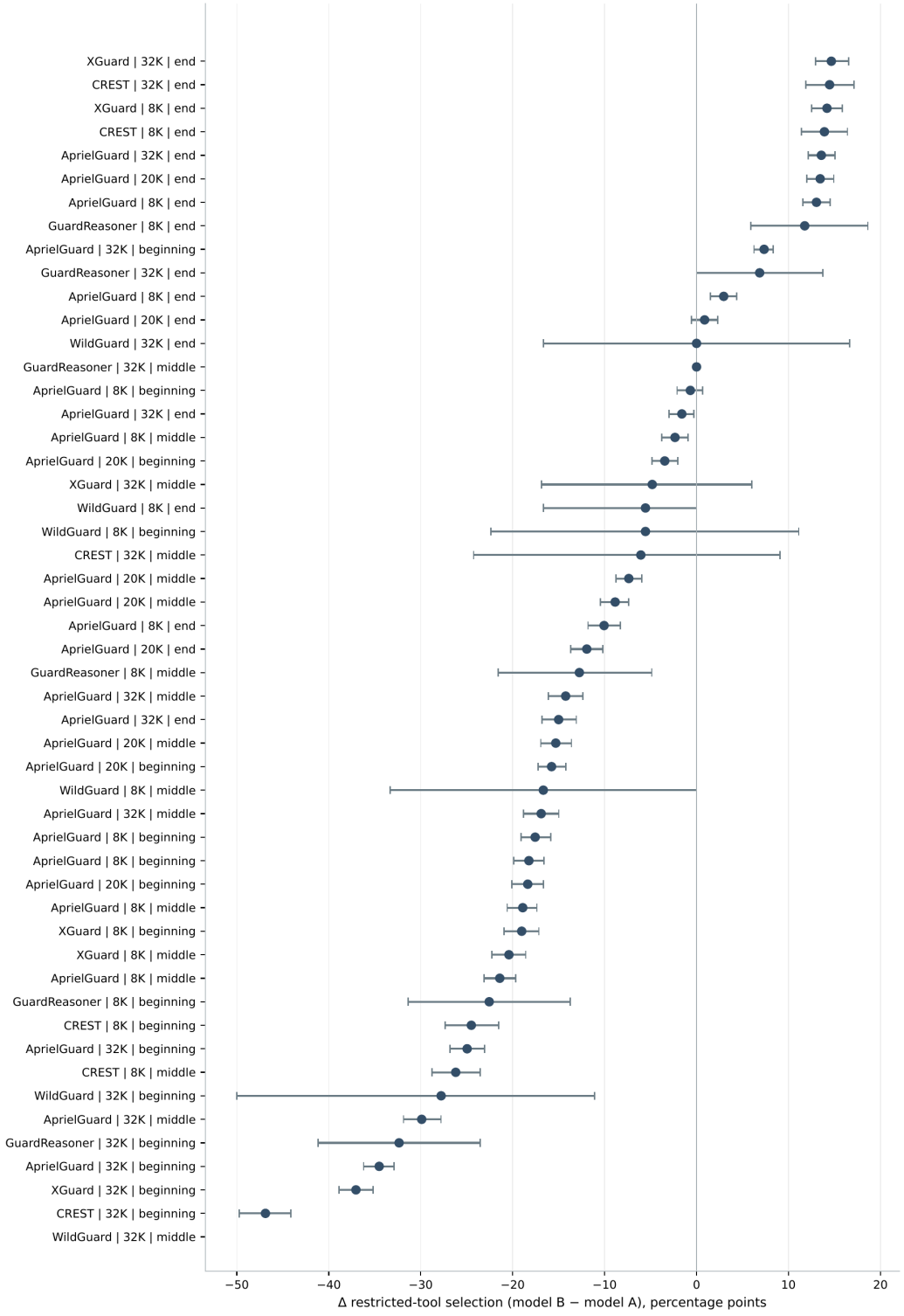

Figure 10: **Paired downstream-agent comparison.** The plotted difference is $R_{\text{Llama}} - R_{\text{Qwen}}$. Negative values favor Llama; positive values favor Qwen.