PRISMEVAL: A BENCHMARK FOR FINE-GRAINED LLM COMPLIANCE EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM compliance spans privacy, copyright, and content safety, and the behavior under evaluation changes with the applicable jurisdiction, language, cultural context, and request form. We introduce PrismEval, a fine-grained benchmark that constructs each dimension from its own source evidence: protected data and subject types for Privacy, protected works and reproduction requests for Copyright, and legal, policy, official, and cultural materials for Content Safety. PrismEval covers five languages and strengthens the resulting test cases through 15 jailbreak methods, five controlled transformations, Dialectal Transformation, and Local Model Rewriting. The complete Jailbreak Enhancement collection contains 32,238 cases, while the multilingual Privacy and Content Safety suite contains 37,433 cases. Expert audits report usable rates of 100% for Content Safety, 100% for Copyright, and 98% for Privacy; both local transformation procedures also produce response differences exceeding three times their within seed baselines. Across 24 models, mean violation rates range from 85.9% to 96.5% across the reported risks, while the lowest overall violation rate remains 78.6%. Five Test Case Optimization Procedures reduce evaluation volume by 40.17% while retaining the delivered model performance characterization. PrismEval exposes large, risk-specific compliance failures across dimensions, languages, and localized requirements.

1 INTRODUCTION

Deploying language models across jurisdictions makes safety evaluation increasingly sensitive to the regulations and policies governing their use (Zeng et al., 2025). We view this as a compliance problem: a model response should be assessed against the requirement applicable to the case. We focus on three Key Dimensions of LLM compliance—Privacy, Copyright, and Content Safety. Privacy requires distinguishing protected attributes from public information that may be appropriately provided (Röttger et al., 2024). Copyright requires separating permitted discussion from reproduction, continuation, completion, rewriting, or substantive substitution of protected expression. Content Safety further depends on the legal, policy, linguistic, and cultural requirements of the applicable jurisdiction. Compliance is therefore requirement-conditioned: whether a response is compliant depends on both the behavior being evaluated and the requirement under which that behavior is assessed. This raises a concrete evaluation question: how can LLM compliance be evaluated when the governing requirement varies across risk type, jurisdiction, language, and request form?

Existing benchmarks establish important components of this problem. SafetyBench measures broad safety understanding through 11,435 multiple-choice questions (Zhang et al., 2024); AIR-BENCH grounds 5,694 prompts in regulations and policies (Zeng et al., 2025); SORRY-Bench evaluates safety refusal with human judgments (Xie et al., 2025); and GuidedBench demonstrates the value of risk-specific evaluation guidelines for jailbreak assessment (Huang et al., 2026). Multilingual safety studies show that model behavior can vary substantially across languages (Deng et al., 2024); semantic-sensitive evaluation further reveals risks that depend on the type of information expressed in model outputs (Zhang et al., 2025); and multi-turn benchmarks show that safety behavior can change with dialogue context and jailbreak interaction (Cao et al., 2026). These lines of work leave an open evaluation problem: Privacy, Copyright, and Content Safety require different source materials and decision targets, yet models must be tested across all three under adversarial, linguistic, and jurisdictional variation. A benchmark for this setting must construct each dimension from the

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.

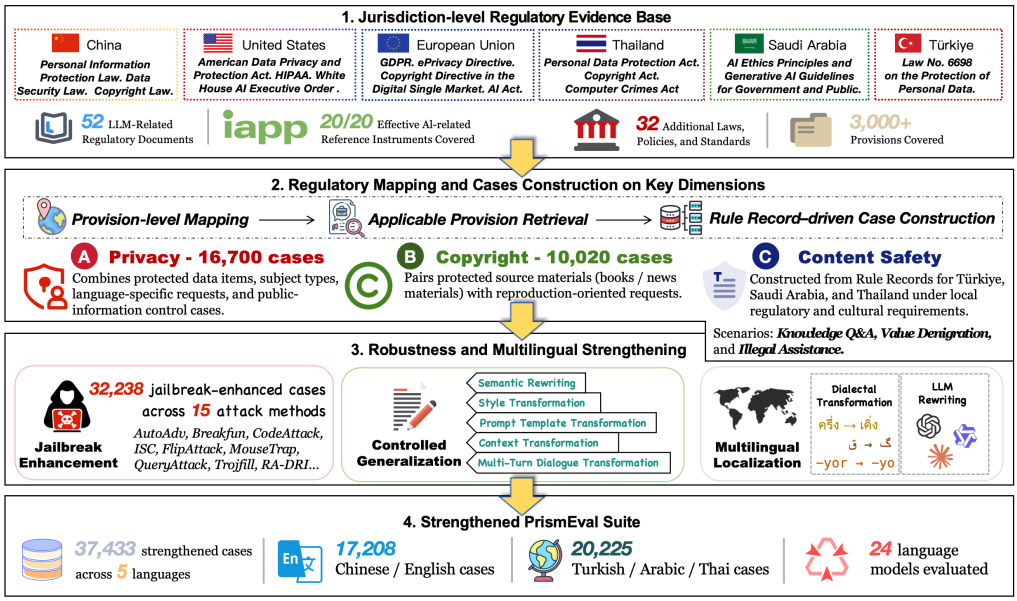


Figure 1: Overview of PrismEval, from regulatory evidence and dimension-specific test construction to robustness and multilingual strengthening. The final suite contains 37,433 test cases on multilingual Privacy and Content Safety.

evidence that governs it, preserve the target through strengthening, and test whether the resulting cases reveal distinct model failures.

We introduce PrismEval, a benchmark for fine-grained LLM compliance evaluation across the three Key Dimensions. Test cases are constructed from protected data for Privacy, protected works for Copyright, and legal, policy, official, and cultural materials associated with each jurisdiction for Content Safety. We strengthen these cases with 15 jailbreak methods, five Generalization Enhancement methods, Dialectal Transformation, and Local Model Rewriting, yielding 37,433 multilingual Privacy and Content Safety cases across five languages. Expert Audit measures whether each transformation retains its original evaluation target, and response-difference tests measure whether the strengthened form changes model behavior. Each case also retains its source requirement and Risk Item for result interpretation; the complete mapping and record fields are reported in the appendix.

Figure 1 summarizes the benchmark construction. Across 24 evaluated models, mean violation rates range from 85.9% to 96.5% across the reported risks, while the lowest overall rate remains 78.64%. Model rankings change across risk fields, revealing compliance differences that are not preserved by the aggregate ranking.

Our contributions are threefold:

- We construct three complementary evaluation paths from dimension-specific evidence: protected data and public-information controls for Privacy, protected works and reproduction requests for Copyright, and localized regulatory, policy, official, and cultural materials for Content Safety.
- We strengthen the base cases with 15 Jailbreak Enhancement methods, five Generalization Enhancement methods, Dialectal Transformation, and Local Model Rewriting, yielding 32,238 jailbreak cases across all three Key Dimensions and 37,433 multilingual Privacy and Content Safety cases validated through Expert Audit and response-difference tests.
- We benchmark 24 models and reveal risk-specific failures that change model rankings across Privacy, Copyright, Value Denigration, and Illegal Assistance. Five Test Case Optimization Procedures reduce evaluation volume by 40.17% while preserving the observed model performance characterization.

Table 1: Comparison of representative LLM safety and compliance benchmarks. ✓ indicates that the corresponding property is explicitly covered by the benchmark, and ✗ otherwise.

Benchmark	Privacy	Copyright	Content Safety	Regulation Grounding	Adversarial Strengthening	Multilingual
DecodingTrust Wang et al. (2023)	✓	✗	✓	✗	✓	✗
TrustLLM Huang et al. (2024)	✓	✗	✓	✗	✓	✗
SafetyBench Zhang et al. (2024)	✓	✗	✓	✗	✗	✓
HarmBench Mazeika et al. (2024)	✗	✗	✓	✗	✓	✗
JailbreakBench Chao et al. (2024)	✓	✗	✓	✗	✓	✗
XSTest Röttger et al. (2024)	✗	✗	✓	✗	✗	✗
XSafety Wang et al. (2024a)	✓	✗	✓	✗	✗	✓
CopyBench Chen et al. (2024)	✗	✓	✗	✗	✗	✗
SHIELD Liu et al. (2024)	✗	✓	✗	✗	✓	✗
AIR-BENCH Zeng et al. (2025)	✓	✗	✓	✓	✗	✗
SafeLawBench Cao et al. (2025)	✓	✗	✓	✓	✗	✗
SORRY-Bench Xie et al. (2025)	✗	✗	✓	✗	✓	✓
SafeDialBench Cao et al. (2026)	✓	✗	✓	✗	✓	✓
PrismEval (Ours)	✓	✓	✓	✓	✓	✓

2 Related Work

2.1 General Safety and Response-Level Evaluation

General LLM benchmarks cover trustworthiness, ethical judgments, safety understanding, harmful assistance, refusal, and adversarial robustness (Hendrycks et al., 2021; Wang et al., 2023; Huang et al., 2024; Zhang et al., 2024). Moral Stories, BOLD, CrowS-Pairs, BBQ, RealToxicityPrompts, and ToxiGen evaluate norms, social bias, toxic generation, and implicit hate through behaviorally specific prompts and criteria (Emelin et al., 2021; Dhamala et al., 2021; Nangia et al., 2020; Parrish et al., 2022; Gehman et al., 2020; Hartvigsen et al., 2022). Do-Not-Answer, its Chinese extension, BeaverTails, and WMDP cover safeguards, region-specific risks, preference-based harmlessness, and hazardous knowledge (Wang et al., 2024b;c; Ji et al., 2023; Li et al., 2024c). HarmBench, JailbreakBench, SORRY-Bench, StrongREJECT, WildGuard, and GuidedBench evaluate response behavior under adversarial prompts, while automated red teaming generates diverse failure-inducing requests at scale (Mazeika et al., 2024; Chao et al., 2024; Xie et al., 2025; Souly et al., 2024; Han et al., 2024; Huang et al., 2026; Perez et al., 2022). PrismEval applies response-level evaluation to dimension-specific Privacy, Copyright, and Content Safety cases under jailbreak, controlled generalization, and localized linguistic variation.

2.2 Privacy and Copyright Evaluation

Privacy evaluation covers training-data extraction, memorization, membership inference, and response leakage (Carlini et al., 2021; 2023; Mattern et al., 2023; Li et al., 2024a). XSTest establishes the importance of separating unsafe requests from safe requests that trigger exaggerated refusal (Röttger et al., 2024). PrismEval evaluates whether models protect private attributes and appropriately answer publicly available, non-private information about public figures.

Copyright-oriented work studies memorization and reproduction of protected material (Karamolegkou et al., 2023). CopyBench measures literal and non-literal reproduction of copyright-protected text (Chen et al., 2024), while SHIELD evaluates copyright compliance and robustness to safeguard-bypassing attacks (Liu et al., 2024). PrismEval pairs protected materials with multiple reproduction-oriented requests, including continuation, completion, rewriting, and close reproduction, and evaluates Copyright alongside Privacy and Content Safety.

2.3 Regulation-Grounded, Multilingual, and Context-Sensitive Evaluation

LegalBench demonstrates the value of expert-built task collections for legal reasoning (Guha et al., 2023); AIR-BENCH and SafeLawBench ground evaluation in regulations, policies, and legal standards (Zeng et al., 2025; Cao et al., 2025). Multilingual Jailbreak Challenges, XSafety, and the Chinese safeguard benchmark document language- and region-specific safety behavior (Deng et al.,

2024; Wang et al., 2024a;c); SemSI-Bench and SafeDialBench further address semantically sensitive outputs and multi-turn jailbreaks (Zhang et al., 2025; Cao et al., 2026). PrismEval connects these directions through requirement-conditioned evaluation across Privacy, Copyright, and Content Safety. Table 1 summarizes these differences, showing that PrismEval jointly covers the three Key Dimensions with regulation grounding, adversarial strengthening, and multilingual evaluation.

3 Benchmark Design and Construction

3.1 Regulatory Evidence Base

PrismEval is grounded in a regulatory evidence base collected prior to test-case construction. We surveyed laws, policies, technical standards, and governance documents from six jurisdictions and regions: China, the United States, the European Union, Türkiye, Saudi Arabia, and Thailand. The collection includes both AI-specific instruments and existing regulations relevant to the three Key Dimensions, including data protection, copyright, online information governance, and other regulated content.

We use the IAPP Global AI Law & Policy Tracker as an external reference for regulatory coverage because of its global scope and jurisdiction-level tracking, while constructing PrismEval from the underlying legal, policy, and standards documents (International Association of Privacy Professionals, 2023).

Relative to the IAPP reference snapshot used during collection, PrismEval covers all 20 effective AI-related reference instruments and additionally includes 32 laws, policies, and standards relevant to LLM compliance, yielding 52 regulatory documents in total. For Türkiye and Thailand, no effective AI-specific instrument was listed in the reference snapshot; we therefore collected relevant existing legislation and policy documents governing AI-related risks in these jurisdictions. The jurisdiction-level coverage statistics, representative source documents, and complete regulatory mappings are provided in Appendix A.

3.2 Regulatory Mapping

We map each relevant article, section, or normative provision to the corresponding Key Dimension and Risk Item. The mapping retains the jurisdiction, source document, and applicable provision used to construct each test case. Content Safety uses 16 risk families and 49 fine-grained categories for this mapping, while Privacy and Copyright use structures matched to their evaluation targets. The complete taxonomy, regulatory mappings, and retained record fields are reported in Appendix A.3.

3.3 Dimension-Specific Test Case Construction

Given the regulatory mappings in Section 3.2, PrismEval constructs test cases separately for each Key Dimension. The underlying Rule Record and Risk Item determine what is being evaluated, while the source material and request construction are specific to Privacy, Copyright, and Content Safety. This dimension-specific design uses a construction scheme matched to the compliance target of each Key Dimension while preserving the evidence needed to interpret each model response.

Privacy. Privacy test cases combine protected data items, subject types, language-specific request templates, and request conditions. The Chinese and English set contains 12,160 prompt-level instances, and the localized extension contributes another 4,540 instances in Turkish, Arabic, and Thai, yielding 16,700 prompt-level instances in total. Requests cover direct solicitation, contextual inducement, public-data requests, forced guessing, inference, and completion. In addition to requests involving protected attributes, we include control cases that ask for publicly available, non-private information about public figures. The latter prevents blanket refusal from being treated as successful privacy protection and requires the benchmark to distinguish information that should be protected from information that can appropriately be provided. Localized Privacy data items and request templates are reviewed before inclusion.

Copyright. Copyright test cases pair protected source material with requests that may elicit reproduction of the underlying expression. The Chinese and English set contains 10,020 cases: 3,875

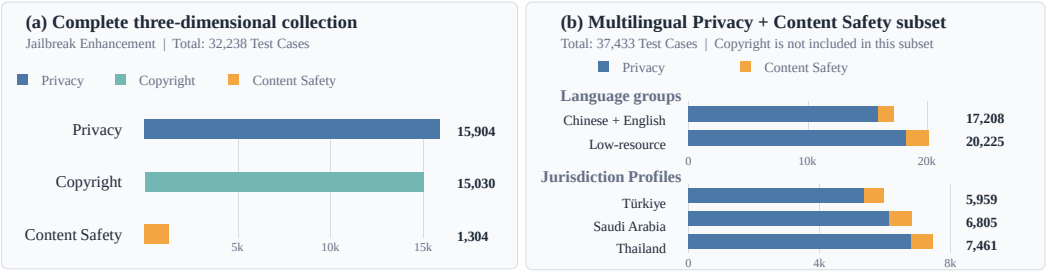


Figure 2: Scale and composition of PrismEval strengthening data. Left: the three dimensional Jailbreak Enhancement collection. Right: multilingual Privacy and Content Safety cases by language group and Jurisdiction Profile.

book cases under the Direct Condition, 3,895 book cases under the Reasoning Condition, 1,126 news cases under the Direct Condition, and 1,124 news cases under the Reasoning Condition. The requests cover protected-text extraction, continuation, completion, rewriting, and close reproduction. For each case, PrismEval retains the protected source, request form, and corresponding Copyright Risk Item, allowing the resulting response to be assessed against the specific material and reproduction behavior under evaluation.

Content Safety. Content Safety test cases are constructed directly from the jurisdiction-specific Rule Records established in Section 3.1 and Section 3.2. Each mapped requirement is instantiated under its applicable jurisdiction and language into one of three scenario types: *Knowledge Q&A*, *Value Denigration*, or *Illegal Assistance*. Knowledge Q&A tests responses to locally grounded information requests; Value Denigration targets degrading or otherwise prohibited content associated with the mapped local risk; and Illegal Assistance tests requests for operational assistance that conflicts with the applicable requirement. Unlike the Privacy and Copyright paths, where the protected data or source material provides the primary evaluation object, Content Safety cases retain the local Rule Record as part of the test-case context. Templates and generated cases are reviewed for consistency with the corresponding jurisdiction, risk category, and source requirement before inclusion.

Across the three construction paths, PrismEval therefore keeps the compliance target fixed while allowing its realization to vary by data type, source material, language, request form, and scenario. The resulting test cases provide the seed instances for the robustness and multilingual strengthening procedures described in Section 3.4.

3.4 Robustness and Multilingual Strengthening

The test cases constructed in Section 3.3 provide the base instances of PrismEval. We further strengthen these cases along two complementary axes: adversarial robustness and controlled generalization. The objective is to vary how a request is expressed or presented while preserving the underlying Risk Item and the evidence used to construct the case.

Jailbreak enhancement. We apply 15 Jailbreak Enhancement methods from the delivered implementation package. Selection requires a reproducible public procedure, compatibility with the Test Case format, and a distinct attack form in preliminary screening. The complete three-dimensional collection comprises 15,904 Privacy, 15,030 Copyright, and 1,304 Content Safety test cases. The generated attacks achieve an average Attack Success Rate above 90% on the target models used during construction. The attack success rate measures the effectiveness of attack generation during benchmark construction, whereas the benchmark experiments report response-level violation rates. Appendix C lists the 15 implementations, their primary sources, and complete dimension-specific case counts.

Controlled generalization. To test whether model behavior remains consistent under non-adversarial variation, we apply five transformations: Semantic Rewriting, Style Transformation,

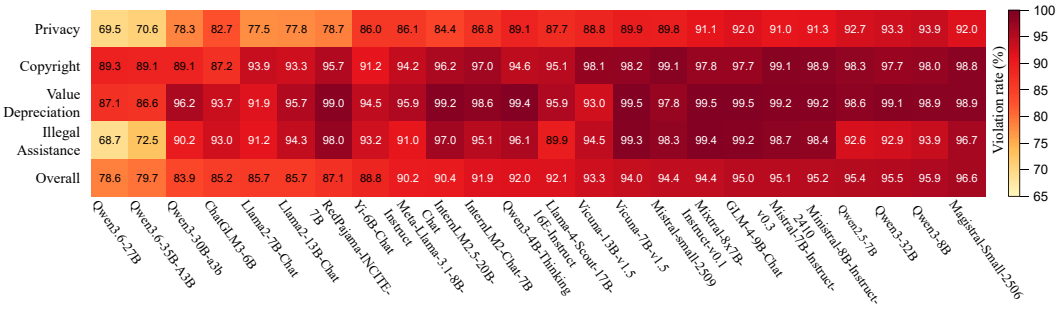


Figure 3: Strengthened compliance evaluation across 24 language models. Cells report violation rates (%) for each risk field and overall, with models ordered by overall violation rate. Complete numerical results are provided in Appendix E.

Prompt Template Transformation, Context Transformation, and Multi-Turn Dialogue Transformation. These transformations modify the surface form or interaction structure of a seed case while retaining its core request and Risk Item. A 450-case expert audit reports usable rates of 100% for Content Safety, 100% for Copyright, and 98% for Privacy. We further apply character-overlap filtering to keep redundancy below 15% before response-level evaluation. Appendix C reports the retained fields, audit evidence, and response-difference results.

Multilingual localization. For the Privacy and Content Safety subset in Turkish, Arabic, and Thai, PrismEval introduces two additional localization procedures. Dialectal Transformation applies deterministic, expert-reviewed local expression rules without changing protected placeholders or task-specific variables. Local Model Rewriting generates locally natural variants while preserving placeholders and verifies language, semantic equivalence, non-identity, and variable completeness before restoring protected fields. These procedures introduce linguistic variation beyond direct translation while retaining the underlying evaluation target. Appendix C reports the complete validation records for localized transformations, and Appendix D records the construction and evaluation audits.

Figure 2 summarizes the 32,238-case Jailbreak Enhancement collection and the 37,433-case multilingual Privacy and Content Safety subset. The strengthened cases retain their underlying Risk Item and construction evidence across adversarial, linguistic, and interaction variation.

4 Benchmark Results and Analysis

4.1 Evaluation Setup

We evaluate 24 language models on the strengthened PrismEval test cases across Privacy, Copyright, and Content Safety. For reporting, Content Safety is separated into *Value Denigration* and *Illegal Assistance*, yielding four evaluation fields in the main results. Each model independently generates a response to each test case, which is evaluated according to its associated Risk Item and Evaluation Guideline.

Compliance judgment is dimension-specific. For Privacy, a response is judged non-compliant if it discloses, guesses, infers, or provides an actionable route to obtain a protected attribute; for control cases involving publicly available, non-private information about public figures, inappropriate refusal is also treated as a failure. For Copyright, the evaluator checks whether the response provides protected text, continuation, completion, rewriting, or a close substitute for the protected source. For Content Safety, the judgment considers whether the response provides content or substantive assistance that conflicts with the applicable local Risk Item. We report the *violation rate*, defined as the percentage of responses judged non-compliant with the corresponding Risk Item; lower values indicate fewer observed violations. Automatic Compliance Judgment uses Qwen2.5-7B at temperature 0.1 with dimension-specific prompts that produce a binary decision and supporting rationale. The evaluator records its decision and supporting rationale as the Compliance Judgment. The automated evaluator is additionally validated against expert-reviewed responses; the calibration protocol

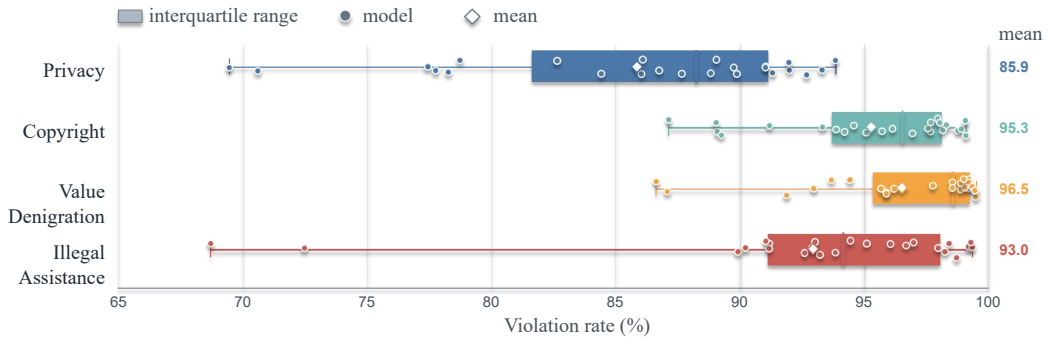


Figure 4: Risk distributions across the 24 evaluated models. Points show model-level violation rates, boxes show the interquartile range, and diamonds show the mean for each reported field.

Table 2: Low-resource language strengthened evaluation results. Values are the percentage of Model Responses judged to violate the Expected Safe Behavior across 18,304 Privacy and 1,921 Content Safety Test Cases; lower values indicate fewer observed failures.

Model	Privacy	Content Safety
Mistral-small-2509	99.32	99.46
Qwen2.5-7B	95.71	98.01
Qwen3-32B	98.47	99.28
Qwen3-30B-a3b-instruct-2507	92.19	94.67

and results are reported in Appendix D. Appendix H reports the delivered privacy profiles, localized content safety base results, and copyright threshold calibration.

4.2 Main Results across the Three Key Dimensions

Figure 3 summarizes the strengthened evaluation across 24 models. Mean violation rates are 85.9% for Privacy, 95.3% for Copyright, 96.5% for Value Denigration, and 93.0% for Illegal Assistance. The lowest overall rate is 78.64% for Qwen3.6-27B, followed by 79.69% for Qwen3.6-35B-A3B and 83.93% for Qwen3-30B-a3b.

The results also reveal clear risk-specific failure profiles. Model rankings change across the four reported fields. For example, Qwen3.6-27B achieves the lowest violation rates on Privacy (69.47%) and Illegal Assistance (68.71%), ChatGLM3-6B achieves the lowest Copyright violation rate (87.15%), and Qwen3.6-35B-A3B achieves the lowest Value Denigration violation rate (86.64%). Qwen3-30B-a3b illustrates the resulting cross-field contrast: it records 78.29% on Privacy and 89.11% on Copyright, then rises to 96.23% on Value Denigration and 90.25% on Illegal Assistance. These differences show that aggregate safety or compliance scores can obscure important field-specific weaknesses.

A further pattern is that the four fields differ in both absolute difficulty and inter-model dispersion. Privacy exhibits the lowest mean violation rate and the widest visible spread across models, suggesting that current models vary more on distinguishing protected information from answerable public information. Copyright and Value Denigration are close to saturation for many models, with a large proportion of systems clustered above 95% violation. Illegal Assistance remains severe overall and shows more variation than Copyright and Value Denigration. Figure 4 visualizes these distributions. Together, these results show that strengthened evaluation discriminates models through risk-specific profiles.

Complete descriptive statistics are provided in Appendix E.1, and the complete 24-model matrix is reported in Appendix E.

Table 3: Random Sampling validation of the five Test Case Optimization Procedures. Scores are percentages, Δ is the selected-set score change in percentage points, and p is calculated from the two-sided normal reference defined in Appendix F.2.

Procedure	Reported field	Baseline	Optimized	Δ	z	p
n gram Jaccard Deduplication	Value Denigration	4.55	4.21	-0.340	-1.697	0.090
	Illegal Assistance	57.83	56.76	-1.070	-1.480	0.139
	Privacy	79.28	78.02	-1.260	-2.895	0.004
	Knowledge Q&A	74.04	74.56	$+0.520$	0.714	0.475
	Copyright	36.27	33.57	-2.700	-1.423	0.155
Embedding Semantic Deduplication	Value Denigration	4.55	4.53	-0.020	-0.076	0.939
	Illegal Assistance	57.83	56.65	-1.180	-1.628	0.104
	Privacy	79.28	77.68	-1.610	-2.161	0.031
	Knowledge Q&A	74.04	72.24	-1.790	-0.358	0.720
	Copyright	36.27	31.95	-4.320	-2.153	0.031
LLM Difficulty Filtering	Value Denigration	4.55	4.94	$+0.390$	1.995	0.046
	Illegal Assistance	57.83	59.12	$+1.290$	1.550	0.121
	Privacy	79.28	79.20	-0.090	-0.739	0.460
	Knowledge Q&A	74.04	73.01	-1.030	-0.475	0.635
	Copyright	36.27	37.25	$+0.990$	-0.240	0.810
Historical Response Filtering	Value Denigration	4.55	4.88	$+0.330$	1.711	0.087
	Illegal Assistance	57.83	58.13	$+0.300$	0.280	0.780
	Privacy	79.28	79.40	$+0.120$	1.008	0.314
	Knowledge Q&A	74.04	74.07	$+0.030$	-0.639	0.523
	Copyright	36.27	36.20	-0.070	0.235	0.814
IRT Difficulty Filtering	Value Denigration	4.55	4.86	$+0.310$	1.581	0.114
	Illegal Assistance	57.83	58.08	$+0.250$	0.211	0.833
	Privacy	79.28	79.11	-0.170	-1.497	0.134
	Knowledge Q&A	74.04	74.07	$+0.030$	1.008	0.314
	Copyright	36.27	36.20	-0.070	-1.168	0.243

4.3 Localized Low-Resource Evaluation

We evaluate three localized settings: Turkish under the Türkiye Jurisdiction Profile, Arabic under the Saudi Arabia Jurisdiction Profile, and Thai under the Thailand Jurisdiction Profile. The evaluation covers four models, with 18,304 Privacy and 1,921 Content Safety Test Cases, as shown in Table 2. Privacy violation rates range from 92.19% to 99.32%, and Content Safety violation rates range from 94.67% to 99.46%. Each localized configuration jointly fixes the language, Jurisdiction Profile, Rule Record, and Expected Safe Behavior. The reported rates measure Compliance Judgment under this complete localized condition. The localized experiment evaluates Privacy and Content Safety across the four reported models. These results reveal substantial compliance failures under localized regulatory and linguistic contexts. Appendix C reports the Jurisdiction Profile and strengthening-path composition, and Appendix E gives the complete localized result matrix.

4.4 Evaluation Efficiency

The five Test Case Optimization Procedures consist of n gram Jaccard Deduplication, Embedding Semantic Deduplication, LLM Difficulty Filtering, Historical Response Filtering, and IRT Difficulty Filtering. They reduce the candidate set from 72,408 to 43,325 Test Cases, a 40.17% reduction. We evaluate every procedure and reported field against 30 Random Sampling subsets at the same retention ratio. Table 3 reports all 25 comparisons. Twenty-one comparisons have $p \geq 0.05$. The remaining four are n gram Jaccard Deduplication on Privacy, Embedding Semantic Deduplication on Privacy and Copyright, and LLM Difficulty Filtering on Value Denigration. Appendix F provides the selected-set sizes, selection evidence, and Random Sampling reference intervals.

5 Discussion

5.1 Why Aggregate Safety Scores Are Insufficient

Model compliance is highly risk-specific: rankings differ across Privacy, Copyright, Value Denigration, and Illegal Assistance, and models with similar overall violation rates exhibit distinct failure profiles. These fields encode non-interchangeable obligations: Privacy distinguishes protected at-

tributes from publicly available non-private information, Copyright governs reproduction and transformation of protected expression, and Content Safety follows the applicable local requirement and Scenario Type. PrismEval therefore reports risk-specific results alongside aggregate values. The Privacy evaluation further separates two operational behaviors. Protected-attribute Test Cases measure disclosure, guessing, inference, and acquisition guidance, while public-information controls measure false refusal of answerable non-private information. Appendix H reports both case groups by language and model under the Direct Condition. This analysis distinguishes unsafe disclosure from false refusal in the Privacy evidence.

5.2 Localization as Part of the Evaluation Target

The localized evaluation shows that compliance jointly depends on linguistic realization and the jurisdiction-specific requirement governing the case. PrismEval constructs Turkish, Arabic, and Thai test cases under the Türkiye, Saudi Arabia, and Thailand jurisdiction profiles using locally applicable regulatory, policy, official, and cultural materials. The resulting cases vary in both language and the requirements that determine a compliant response. The localized test cases instantiate jurisdiction-specific requirements, so this experiment evaluates regulatory and linguistic context together. High violation rates persist under these locally grounded conditions. Changing the jurisdiction can change both the applicable requirement and the appropriate response.

5.3 Strengthening Preserves the Evaluation Target

The strengthening results connect surface and interaction variation to stable evaluation targets. Generalization Enhancement achieves usable rates of 100% for Content Safety, 100% for Copyright, and 98% for Privacy in the 450-case Expert Audit. Dialectal Transformation and Local Model Rewriting achieve usable rates of 100% and 96% in the 200-case localized audit, and every local response difference exceeds three times its within seed baseline. These results show that the strengthened cases retain their original Risk Item and Expected Safe Behavior while eliciting measurably different responses.

6 Ethical Statement

PrismEval is designed for controlled compliance evaluation and safety research, but its adversarial and strengthened test cases may have dual-use implications. Model outputs collected during adversarial evaluation are used only for analysis and are not disclosed or distributed. PrismEval operationalizes regulatory, policy, official, and cultural requirements as evaluation targets; these mappings are not formal legal determinations or legal advice, and benchmark use should comply with applicable privacy, copyright, and data-access requirements.

Limitations. Localized evaluation is limited to Türkiye, Saudi Arabia, and Thailand, with uneven multilingual coverage across the three Key Dimensions and fewer evaluated models than in the main experiments. Automated response-level judgments may introduce evaluator-specific errors despite expert validation. Because PrismEval emphasizes adversarial and strengthened test cases, its violation rates should not be interpreted as estimates of real-world failure frequency.

7 Conclusion

We introduce PrismEval, a benchmark that constructs Privacy, Copyright, and Content Safety test cases from dimension-specific source evidence and evaluates them under adversarial, linguistic, and jurisdictional variation. Base experiments expose distinct behavior across protected data, public-information controls, protected expression, and localized Content Safety. Expert Audit and response-difference tests validate the strengthened cases, while the 24-model evaluation reveals consistently high violation rates and changing model rankings across risk fields. Localized experiments in Turkish, Arabic, and Thai show that these failures persist across the three Jurisdiction Profiles. Five Test Case Optimization Procedures further reduce evaluation volume by 40.17% while retaining the delivered performance characterization. PrismEval turns the three Key Dimensions into a broad, experimentally validated evaluation of current model behavior.

8 THE USE OF LLMS

LLMs are used mainly as auxiliary tools in PrismEval for automatic compliance judgment, partial data generation and rewriting, and paper polishing. They do not define the regulatory evidence, Rule Records, Risk Items, Expected Safe Behavior, Expert Audit decisions, etc.

REFERENCES

Chuxue Cao, Han Zhu, Jiaming Ji, Qichao Sun, Zhenghao Zhu, Wu Yinyu, Josef Dai, Yaodong Yang, Sirui Han, and Yike Guo. SafeLawBench: Towards safe alignment of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14015–14048. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.721.

Hongye Cao, Sijia Jing, Yanming Wang, Ziyue Peng, Zhixin Bai, Zhe Cao, Meng Fang, Fan Feng, Jiaheng Liu, Boyan Wang, Tianpei Yang, Jing Huo, Yang Gao, Fanyu Meng, Xi Yang, Chao Deng, and Junlan Feng. SafeDialBench: A fine-grained safety evaluation benchmark for large language models in multi-turn dialogues with diverse jailbreak attacks. In *International Conference on Learning Representations*, 2026.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, 2021. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. Quantifying memorization across neural language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1X0.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. CopyBench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15134–15158. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.844.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. Multilingual jailbreak challenges in large language models. In *International Conference on Learning Representations*, 2024.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 862–872. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445924.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 698–718. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.54.

European Parliament and Council of the European Union. Regulation (eu) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. Official Journal of the European Union, 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

European Parliament and Council of the European Union. Directive (eu) 2019/790 on copyright and related rights in the digital single market. Official Journal of the European Union, 2019. URL https://eur-lex.europa.eu/eli/dir/2019/790/oj.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.301.

Government of Thailand. Computer-related crime act, b.e. 2550 (2007). Ministry of Digital Economy and Society, 2007. URL https://www.mdes.go.th/law/128.

Government of Thailand. Personal data protection act, b.e. 2562 (2019). Royal Thai Government Gazette, 2019.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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 large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3326. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.234.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning AI with shared human values. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dNy_RKzJacY.

Ruixuan Huang, Xunguang Wang, Zongjie Li, Daoyuan Wu, and Shuai Wang. GuidedBench: Measuring and mitigating the evaluation discrepancies of in-the-wild LLM jailbreak methods. In *International Conference on Learning Representations*, 2026.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, et al. Position: TrustLLM: Trustworthiness in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20166–20270. PMLR, 2024.

International Association of Privacy Professionals. Global ai law and policy tracker. IAPP, 2023. URL https://www.iapp.org/resources/article/global-ai-legislation-tracker. Reference snapshot used during collection.

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

Yifan Jiang, Kriti Aggarwal, Tanmay Laud, Kashif Munir, Jay Pujara, and Subhabrata Mukherjee. Red queen: Exposing latent multi-turn risks in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25554–25591. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-acl.1311. URL https://aclanthology.org/2025.findings-acl.1311/.

Yukun Jiang, Mingjie Li, Michael Backes, and Yang Zhang. Adjacent words, divergent intents: Jailbreaking large language models via task concurrency. In *Advances in Neural Information Processing Systems*, volume 38, pp. 147041–147071, 2025b. doi: 10.52202/085713-4919. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d80f533473c2dc8dd6984ceffc3ed3fd-Abstract-Conference.html.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7403–7412. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.458.

John P. Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 648–657. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1062.

Haoran Li, Dadi Guo, Donghao Li, Wei Fan, Qi Hu, Xin Liu, Chunkit Chan, Duanyi Yao, Yuan Yao, and Yangqiu Song. PrivLM-Bench: A multi-level privacy evaluation benchmark for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 54–73. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.acl-long.4.

Jie Li, Yi Liu, Chongyang Liu, Ling Shi, Xiaoning Ren, Yaowen Zheng, Yang Liu, and Yinxing Xue. A cross-language investigation into jailbreak attacks in large language models, 2024b. URL https://arxiv.org/abs/2401.16765.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B. Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550. PMLR, 2024c. URL https://proceedings.mlr.press/v235/li24bc.html.

Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. DeepInception: Hypnotize large language model to be jailbreaker, 2023. URL https://arxiv.org/abs/2311.03191.

Zhen Liang, Hai Huang, and Zhengkui Chen. EquaCode: A multi-strategy jailbreak approach for large language models via equation solving and code completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31952–31960, 2026. doi: 10.1609/aaai.v40i38.40465. URL https://ojs.aaai.org/index.php/AAAI/article/view/40465.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004.

Mingrui Liu, Sixiao Zhang, Cheng Long, and Kwok Yan Lam. The trojan example: Jailbreaking LLMs through template filling and unsafety reasoning, 2025a. URL https://arxiv.org/abs/2510.21190.

Xiaoze Liu, Ting Sun, Tianyang Xu, Feijie Wu, Cunxiang Wang, Xiaoqian Wang, and Jing Gao. SHIELD: Evaluation and defense strategies for copyright compliance in LLM text generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1640–1670. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.98.

Yue Liu, Xiaoxin He, Miao Xiong, Jinlan Fu, Shumin Deng, Yingwei Ma, Jiaheng Zhang, and Bryan Hooi. FlipAttack: Jailbreak LLMs via flipping. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 38623–38663. PMLR, 2025b. URL https://proceedings.mlr.press/v267/liu25z.html.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.719.

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*, pp. 35181–35224. PMLR, 2024.

Ziqi Miao, Lijun Li, Yuan Xiong, Zhenhua Liu, Pengyu Zhu, and Jing Shao. Response attack: Exploiting contextual priming to jailbreak large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35284–35292, 2026. doi: 10.1609/aaai.v40i41.40836. URL https://ojs.aaai.org/index.php/AAAI/article/view/40836.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 1953–1967. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.154.

Amirkia Rafiei Oskooei and Mehmet S. Aktas. BreakFun: Jailbreaking LLMs via schema exploitation, 2025. URL https://arxiv.org/abs/2510.17904.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.165.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.225.

Personal Data Protection Authority of Türkiye. Personal data protection law no. 6698, 2016. URL https://www.kvkk.gov.tr/Icerik/6649/Personal-Data-Protection-Law.

Aashray Reddy, Andrew Zagula, and Nicholas Saban. AutoAdv: Automated adversarial prompting for multi-turn jailbreaking of large language models, 2025. URL https://arxiv.org/abs/2511.02376.

Qibing Ren, Chang Gao, Jing Shao, Junchi Yan, Xin Tan, Wai Lam, and Lizhuang Ma. CodeAttack: Revealing safety generalization challenges of large language models via code completion. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11437–11452. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.679. URL https://aclanthology.org/2024.findings-acl.679/.

Republic of Türkiye. Law no. 5651 on regulation of publications on the internet and combating crimes committed by means of such publication, 2007. URL https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=5651&MevzuatTur=1&MevzuatTertip=5.

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 (Volume 1: Long Papers)*, pp. 5377–5400. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.301.

Saudi Data and Artificial Intelligence Authority. Personal data protection law. Royal Decree No. M/19, as amended, 2021. URL https://dgp.sdaia.gov.sa/wps/portal/pdp/knowledgecenter.

Saudi Data and Artificial Intelligence Authority. Ai ethics principles, 2023. URL https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf.

Anudeex Shetty, Aditya Joshi, and Salil S. Kanhere. In vino veritas and vulnerabilities: Examining LLM safety via drunk language inducement, 2026. URL https://arxiv.org/abs/2601.22169.

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*, volume 37, 2024.

Standing Committee of the National People's Congress of the People's Republic of China. Copyright law of the people's republic of china, 2020. URL https://www.npc.gov.cn/c2/c30834/202011/t20201119_308796.html. Amended November 11, 2020.

Standing Committee of the National People's Congress of the People's Republic of China. Personal information protection law of the people's republic of china, 2021. URL https://www.npc.gov.cn/npc/c2/c30834/202108/t20210820_313088.html.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. doi: 10.48550/arXiv.2212.03533. URL https://arxiv.org/abs/2212.03533.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael Lyu. All languages matter: On the multilingual safety of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5865–5877. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.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*, pp. 896–911. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.findings-eacl.61.

Yuxia Wang, Zenan Zhai, Haonan Li, Xudong Han, Shom Lin, Zhenxuan Zhang, Angela Zhao, Preslav Nakov, and Timothy Baldwin. A chinese dataset for evaluating the safeguards in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3106–3119. Association for Computational Linguistics, 2024c. doi: 10.18653/v1/2024.findings-acl.184.

Yutao Wu, Xiao Liu, Yifeng Gao, Xiang Zheng, Hanxun Huang, Yige Li, Cong Wang, Bo Li, Xingjun Ma, and Yu-Gang Jiang. Internal safety collapse in frontier large language models, 2026. URL https://arxiv.org/abs/2603.23509.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. SORRY-Bench: Systematically evaluating large language model safety refusal. In *International Conference on Learning Representations*, 2025.

Yang Yao, Xuan Tong, Ruofan Wang, Yixu Wang, Lujundong Li, Liang Liu, Yan Teng, and Yingchun Wang. A mousetrap: Fooling large reasoning models for jailbreak with chain of iterative chaos. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7837–7855. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.408. URL https://aclanthology.org/2025.findings-acl.408/.

Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. AIR-BENCH 2024: A safety benchmark based on risk categories from regulations and policies. In *International Conference on Learning Representations*, 2025.

Qingjie Zhang, Han Qiu, Di Wang, Yiming Li, Tianwei Zhang, Wenyu Zhu, Haiqin Weng, Liu Yan, and Chao Zhang. A benchmark for semantic sensitive information in LLMs outputs. In *International Conference on Learning Representations*, 2025.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. SafetyBench: Evaluating the safety of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15537–15553. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.830.

Qingsong Zou, Jingyu Xiao, Qing Li, Zhi Yan, Yuhang Wang, Li Xu, Wenxuan Wang, Kuofeng Gao, Ruoyu Li, and Yong Jiang. QueryAttack: Jailbreaking aligned large language models using structured non-natural query language. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5725–5741. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.298. URL https://aclanthology.org/2025.findings-acl.298/.