KoEVD: A Korean Benchmark for Metadata-Conditioned Multi-Stage LLM Safety Evaluation

Anonymous ACL submission

Abstract

LLM safety evaluation often asks whether a model refuses harmful prompts, but deployed systems make multiple safety-relevant decisions: judging a user utterance, selecting among candidate responses, generating a final response, and deciding whether a tool/action option should be used. Existing benchmarks typically study these decisions in separate prompt, response, or tool-use settings, making it difficult to compare failures for the same safety-relevant request. We introduce KoEVD (Korean Ethical Violation Dataset), a Korean benchmark that links these decision points under a shared user-utterance taxonomy. Built through an approximately 3.5-year annotator-governed, AI-assisted process in which annotators rewrote, relabeled, filtered, and adjudicated AI-assisted candidates, KoEVD contains 13,552 curated Korean safety-relevant user utterances and 71,395 linked assistant response candidates. It supports five tasks: user-utterance risk judgment, candidate-response safety choice and relative harmfulness judgment, response harmfulness evaluation for direct generations, response strategy analysis, and tool-call/action-choice safety evaluation. Experiments with open-weight and commercial models show that observed safety behavior varies substantially by task stage, severity, explicitness, utterance type, topic, and tool-use framing. Direct generation and pre-execution mock tool/action choice reveal failures missed by prompt-side judgment or text-only refusal evaluation, including cases where a model generates a safe text response yet selects an unsafe action option when tool affordances are provided. KoEVD supports Korean metadata-conditioned safety evaluation, response-strategy analysis, and pre-execution diagnostics for tool/action-choice safety.

1 Introduction

Whether a model refuses a harmful prompt is an important signal in LLM safety evaluation, but refusal alone does not capture safety-relevant behavior (Röttger et al., 2024; Cui et al., 2025; Xie et al., 2024). Existing benchmarks study harmfulness detection, response safety, and safety–helpfulness trade-offs (Zhang et al., 2024b; Wang et al., 2023; Han et al., 2024; Ji et al., 2023), while tool-use benchmarks show that safety can also depend on which action an agent selects (Ruan et al., 2024; Xia et al., 2025; Cartagena and Teixeira, 2026). However, these settings are usually evaluated in isolation. Real LLM applications involve multiple safety-relevant decision points: a model may classify user risk, choose among candidate responses, generate a response under a safety strategy, or decide whether a tool/action option is safe to execute. Evaluating only one point can miss failures that emerge elsewhere; for example, a model may correctly identify an unsafe request but still generate unsafe content, or it may avoid unsafe text while selecting an unsafe action once a tool schema is introduced.

Safety evaluation also requires user utterances that reflect the linguistic and sociocultural cues of the target language community. Prior Korean resources show that safety-relevant phenomena are shaped by language- and culture-specific context (Jeong et al., 2022; Jin et al., 2024; Lim et al., 2025). Translating English benchmarks may preserve abstract harm categories, but it does not reliably reproduce the implicatures, slang, idioms, social hierarchies, event references, and background knowledge that shape Korean interactions. Korean safety evaluation therefore requires user utterances and metadata constructed directly in Korean-language contexts.

We propose KoEVD (Korean Ethical Violation Dataset), a Korean benchmark for multi-stage LLM safety evaluation. KoEVD addresses the gap by using the same Korean user utterances across five linked tasks: prompt-risk judgment, candidate-response safety choice, direct-generation harm-

fulness evaluation, refusal and response-strategy analysis, and tool/action-choice safety. Each utterance is annotated with severity, explicitness, safety-relevant utterance type, topic labels, and interpretive rationales; linked response candidates are annotated with response tiers and response strategies; and the same user-side taxonomy is extended to mock tool/action-choice scenarios. KoEVD is not a collection of raw LLM outputs: LLMs assisted ideation and drafting, but final records were fixed through human rewriting, relabeling, response-tier adjudication, filtering, and borderline-case consensus.

Contributions. KoEVD contributes (i) an annotator-governed Korean safety corpus of 13,552 curated user utterances and 71,395 linked response candidates with rationale and response-strategy annotations; (ii) a shared-source five-task benchmark connecting prompt-risk judgment, response selection, direct generation, strategy analysis, and tool/action-choice safety; (iii) empirical analyses showing that aggregate scores hide large variation across task stage, severity, explicitness, utterance type, topic, and tool-use framing; and (iv) a pre-execution tool/action-choice diagnostic showing how unsafe mock action options can surface failures beyond text-only refusal settings.

2 Related Work

We situate KoEVD within five lines of work: prompt/response safety benchmarks, refusal and response-behavior evaluation, multilingual and cross-cultural safety, Korean and culture-grounded safety resources, and tool-use/agent safety. Across these areas, KoEVD links multiple evaluation units under a shared Korean user-utterance taxonomy rather than treating them as isolated tasks.

Prompt and response safety benchmarks. Existing safety benchmarks evaluate harmful prompts and responses under diverse taxonomies, including SafetyBench, SALAD-Bench, Do-Not-Answer, WildGuard, BeaverTails, and PKU-SafeRLHF (Zhang et al., 2024b; Li et al., 2024; Wang et al., 2023; Han et al., 2024; Ji et al., 2023, 2024). These resources motivate prompt harmfulness detection, response moderation, and harmlessness/helpfulness trade-off analysis. KoEVD extends this line by using a shared Korean source set to compare prompt-, response-, generation-, strategy-, and action-choice evaluations.

Refusal and response behavior. Refusal-focused work such as XSTest, OR-Bench, SORRY-Bench, and Let Them Down Easy! shows that safety cannot be reduced to whether a model refuses: over-refusal, under-refusal, and refusal style all matter (Röttger et al., 2024; Cui et al., 2025; Xie et al., 2024; Zheng et al., 2025). Work on intervention responses, counterspeech, and counter-narratives further motivates analyzing how safe responses are constructed, not only whether they are safe (Qian et al., 2019; Mathew et al., 2019; Chung et al., 2021). KoEVD reflects this view by annotating response candidates with multi-label strategies such as boundary setting, safe information provision, redirection, and counterspeech.

Multilingual and cross-cultural safety. Multilingual safety benchmarks show that safety behavior does not transfer uniformly across languages and cultural settings. MultiJail studies multilingual jailbreak vulnerability, M-ALERT reports cross-linguistic safety inconsistencies across languages and categories, and XL-SafetyBench frames safety as a country-grounded cross-cultural evaluation problem (Deng et al., 2024; Friedrich et al., 2025; Choi et al., 2026). KoEVD is closest to this line in motivation, but takes a Korean-focused depth-first approach: rather than translating or parallelizing a common prompt set across many languages, it constructs Korean-context utterances and links them to downstream response and action-choice evaluations.

Korean and culture-grounded safety resources. Korean resources such as KOLD, KoSBi, KoBBQ, SQuARe, and AssurAI address offensive language, social bias, sensitive question answering, and sociocultural risk (Jeong et al., 2022; Lee et al., 2023b; Jin et al., 2024; Lee et al., 2023a; Lim et al., 2025). KoEVD is not positioned as the first Korean safety dataset. Its distinction lies in linking Korean safety-relevant utterances to interpretive rationales, response candidates, response strategies, and tool/action scenarios under a single user-side taxonomy.

Tool-use and agent safety. Tool-use and agent-safety benchmarks evaluate whether agents use tools correctly and safely, often focusing on privacy leakage, file deletion, API misuse, cyber risk, and sandbox violations (Ruan et al., 2024; Xia et al., 2025; Zhang et al., 2024a; Zong et al., 2026; Carta-

gena and Teixeira, 2026). KoEVD Task 5 is related, but evaluates unsafe action-option selection using the same Korean taxonomy as the text tasks and covers risks such as physical harm, harmful publication, privacy violation, cyber abuse, and other real-world consequences. We treat it as a pre-execution diagnostic rather than a benchmark that executes tools.

3 Dataset and Annotation Governance

The central design requirement for KoEVD is a reusable user-utterance layer with consistent user-side and response-side annotations. We therefore describe the corpus, construction governance, annotation axes, and reliability checks before defining task-specific evaluation instances in Section 4.

3.1 Corpus Overview

KoEVD consists of 13,552 curated Korean safety-relevant user utterances and 71,395 linked assistant response candidates. Each user utterance is annotated with topic labels, utterance type, severity band, explicitness, and interpretive rationale fields. Linked responses are organized into accepted, rejected, and strongly rejected tiers; each item also has an annotator-selected accepted response and response-strategy labels for accepted responses.

KoEVD uses five severity bands (Benign, Low, Moderate, High, Extreme), three explicitness labels (Neutral, Implicit, Explicit), 11 utterance types spanning general speech acts and safety-relevant request types, and a multi-label topic field with 22 labels covering social-bias categories and broader LLM-safety domains such as privacy, national security, intellectual property, specialized advice, future prediction, and benign/null cases.

The taxonomy adapts, rather than copies, directly related Korean safety/bias resources and broader LLM safety work for KoEVD's multi-stage evaluation goal. Social-bias-related topics draw on Korean demographic and social-group categories (Lee et al., 2023b). Utterance types combine general communicative functions, such as opinion seeking, opinion expression, and information provision, motivated by dialogue-act work (Bunt et al., 2020), with safety-relevant request patterns. Biased Questions draw on bias-sensitive QA benchmarks (Parrish et al., 2022; Jin et al., 2024), while Prohibited Acts, Personal Info, Specialized Advice, and Controversial Requests are informed by broader LLM risk and do-not-answer taxonomies (Weidinger et al., 2022; Wang et al., 2023). Response strategy is a separate response-side schema motivated by intervention responses, counterspeech, and knowledge-grounded counter-narratives (Zheng et al., 2025; Qian et al., 2019; Mathew et al., 2019; Chung et al., 2021); we adapt these ideas to describe how safe assistant responses combine boundary setting, counterspeech, information provision, reasoning, education, redirection, and related strategies.

The full KoEVD corpus is a reusable benchmark resource rather than a requirement that every experiment use all 13,552 user utterances. Section 4 defines the task-specific source views and variants, and Appendix E reports split composition and compact-split approximation checks.

3.2 Dataset Construction Process

KoEVD is a human-curated, AI-assisted corpus built through a three-phase process rather than a single LLM generation pass. Figure 1 summarizes the corpus-construction pipeline and the five benchmark tasks supported by the final corpus.

Phase 1: guideline design and calibration. For 18 months, an internal team led by language and computational-linguistics researchers, together with two project researchers hired for the benchmark, conducted literature review, source-material collection, guideline development, and calibration of safe/unsafe judgment criteria and label boundaries. This phase produced 5,998 data items.

Phase 2: prompt and candidate stabilization. From Phase 2, two additional project researchers joined after calibration, and the expanded team stabilized AI-assisted candidate-generation prompts using collected sources, underrepresented metadata combinations, role-play, proverbs, unsafe terms, Korean online-community slang, and culturally or politically sensitive references. Candidate user utterances were checked for diversity, naturalness, label alignment, and fixability; if a natural utterance no longer matched the assigned label, the label was revised rather than forcing an unnatural sentence. For assistant responses, annotators explored which response-strategy combinations were natural under severity, explicitness, and utterance-type conditions before stabilizing response-generation prompts. This stage produced 54,833 raw/generated candidate items before annotation and filtering.

Figure 1: KoEVD corpus construction and benchmark design. The corpus was built through a three-phase human-governed, AI-assisted process and supports five safety-relevant evaluation tasks over shared user-utterance metadata. Simulated tool/action tasks use mock tools only; no real tool is executed.

Phase 3: annotation, rewriting, and consensus. In the full annotation stage, the expanded team re-judged each candidate and all user- and response-side fields, rewrote unnatural content, corrected labels and response tiers/strategies, and removed unreliable items; their role was data construction and adjudication, not passive approval of AI-assisted candidates. A metadata-balance-aware assignment tool prioritized underrepresented cells, while recurring consensus meetings resolved borderline cases as guideline clarifications.

3.3 Annotation Axes and Interpretive Rationale Fields

The user-side metadata introduced above supports all five tasks, while response-side annotations support response selection and strategy analysis. Response strategies are annotated with 15 response-side labels and grouped for Task 4 into Boundary / Definition, Stance / Counter-Speech, Informational Support, Interaction Management, and Humour; generated unsafe responses are tracked with a separate Unsafe marker. Construction, staffing, annotator-instruction, annotation-workflow, and label-definition details are in Appendix B; schema and corpus distributions are in Appendix C; response-tier and strategy details are in Appendix D; and Task 4 extended analyses are in Appendix J.

In addition to categorical labels, KoEVD provides three interpretive rationale fields. `problem_reason` records why an utterance is problematic; `implies` records implied meanings that are not explicit on the surface; and `background_knowledge` records contextual knowledge needed for the judgment. These fields support annotation documentation and qualitative auditing by explaining how annotators interpreted implicit harm and culturally grounded Korean utterances.

3.4 Calibration and Consensus

Before full-scale labeling, annotators completed more than 100 calibration annotations per label category, cross-reviewed cases from at least two other annotators, and used disagreements as training and guideline-interpretation cases. To reduce safety-criteria drift, the project kept a small internal team and required calibration before newly hired project researchers contributed to final annotation. Borderline cases were discussed in recurring consensus meetings and accumulated as case-level guideline clarifications.

We also conducted an 800-item stratified IAA audit: 160 examples per severity band, with explicitness balancing and diversity over utterance type, topic, source group, and rare metadata combinations. Comparing two cleaned audit annotations against the reference annotation, Krippendorff's alpha was 0.813–0.890 for topic, utterance type, explicitness, and nominal severity, and 0.925 for distance-aware ordered severity. Chosen-response strategy exact-set alpha was 0.982. Additional agreement diagnostics and multi-label reliability summaries are reported in Appendix L. The selected accepted response should be read as the annotator-selected safe response for that item, not a universally optimal response.

4 Tasks and Evaluation Protocol

KoEVD evaluates five safety-relevant decision points that may appear separately or jointly in LLM

applications. A task instantiates corpus items into evaluation instances and defines an associated readout over the shared user-utterance taxonomy. Table 1 summarizes the task construction and main readouts.

A *source view* selects user-utterance IDs before prompting; a *controlled analysis subset* rebalances metadata for condition-level comparison; and a *variant* changes the prompt or choice format while holding the underlying source items fixed. Thus, Task 2 response-choice variants and Task 5 prompt-framing variants are not separate source splits.

Task 2 has four variants: choosing the safest response, choosing the unsafe response, distinguishing rejected from accepted responses, and ranking relative harmfulness. These variants test whether a model can use the response-tier labels in different choice directions rather than only selecting a preferred answer.

Task 5 is the only task instantiated as a dedicated tool/action scenario set rather than a full-corpus text evaluation. We sample 784 source requests for tool-use suitability and risk-profile coverage, including physical harm, privacy, cyber/payment abuse, harmful publication, and high-stakes action scenarios; this is a diagnostic scenario subset, not a random sample intended to estimate the full corpus average. Its `safe_guided`, `tool_forced`, and `unsafe_induced` variants keep the same source requests and mock tool profiles but change prompt framing. No real tools are executed, and an unsafe call is counted only when the selected mock tool belongs to the request-specific unsafe-tool set.

Illustrative sanitized examples. To make the task distinctions concrete, Table 2 gives sanitized/paraphrased examples. We avoid raw harmful wording, procedural detail, personal identifiers, and executable tool arguments in the paper; full benchmark items are handled through the release-safety process described in Appendix N.

We evaluate a 12-model pool: six open-weight models and six commercial API models. The pool is intended to cover contemporary commercial API families and representative open-weight model families for diagnostic comparison, not to define an exhaustive leaderboard. The commercial models cover OpenAI, Anthropic, and Google families (OpenAI, 2026; Anthropic, 2026; Google, 2026a), and the open-weight models cover Qwen, Gemma, Mistral/Ministral, Apertus, and OLMo families (Qwen Team, 2026; Google, 2026b; Mistral AI, 2026; Swiss AI Initiative, 2026; Team Olmo et al., 2025). Compact model/run settings, compute budget, metric denominators, and parse-handling caveats are reported in Appendix N.

Unsafe judgment rate, response-selection accuracy, unsafe generation rate, strategy distribution, and unsafe tool-call rate are not interchangeable leaderboard scores; they are observed behavior signals from different decision points. Because most runs are single-pass evaluations with task-specific parsing or skip denominators, we do not interpret small score gaps as statistically significant model rankings. Our claims instead rely on large condition-level contrasts, cross-task mismatches, and disaggregated metadata patterns.

Tasks 3 and 4 share the same generated model responses and use classifier-assisted measurements. The Task 3 assistant-safety classifier achieved macro/micro F1 of 0.9990/0.9990 on withheld validation; the Task 4 response-strategy classifier achieved macro F1 0.9935, micro F1 0.9951, and primary-strategy accuracy 0.9925. We treat these classifiers as reproducible measurement proxies applied consistently to all models, not as absolute human ground truth; classifier validation summaries and measurement caveats are in Appendix F.

5 Results

We organize the results around four research questions (RQs) that test whether safety behavior changes across metadata conditions, task stages, and model profiles when evaluation is grounded in shared source requests.

5.1 RQ1: Is Safety Judgment Conditioned by Metadata and Model Profile?

RQ1 asks whether user-safety judgment follows severity alone or varies with metadata and model-specific judgment profiles. Figure 2 compares model-mean unsafe rates by utterance type within each severity band and model-specific rates for selected metadata cells.

In the full evaluation view, model-mean unsafe judgment generally increases with severity, but not monotonically: High reaches 84.2%, while Extreme reaches 80.7%. The severity–explicitness balanced subset restores a clearer pattern (Low 28.5%, Moderate 59.9%, High 84.0%, Extreme 84.4%), and explicit prompts are usually judged unsafe more often than implicit prompts in Low–High bands. Extreme remains mixed because it in-

Task	Evaluation instance	Scoring or analysis readout
T1: Prompt-risk judgment	A user utterance from a full-corpus or controlled source view is paired with a binary safety-judgment prompt.	Parsed safe/unsafe judgment
T2: Response-choice safety	The same user utterance is shown with accepted, rejected, and strongly rejected response candidates under four choice variants.	Parsed option letter compared with the response-tier gold label
T3: Direct-generation harmfulness	A user utterance from the same source views as T1 is used as a free-form generation prompt.	Generated response classified for response harmfulness
T4: Response-strategy analysis	The generated responses from T3 are reused; no new model prompt is issued.	Primary and multi-label response-strategy labels over the generated response
T5: Tool/action-choice safety	A tool-use-suitable source request is paired with request-specific mock tools under three prompt-framing variants.	Parsed JSON decision; unsafe when the selected mock tool is in the request-specific unsafe-tool set

Table 1: Benchmark task inventory in KoEVD. The table summarizes each task's evaluation instance and main readout; metric denominators, parse-handling caveats, prompt/output contracts, compact model/run settings, and release-safety handling are reported in Appendix N.

Task	Sanitized scenario	What it illustrates
T1	A Korean utterance implicitly links an origin or ethnicity group with being dirty, smelly, or impossible to live with.	User-risk judgment can depend on culturally grounded implicature and background knowledge, not only explicit harmful keywords.
T2/T4	A user asks whether reporting a counterfeit product as counterfeit could create defamation liability; candidate responses differ in legal caution, boundary setting, and safe general information.	Response tiers and response strategies distinguish safe assistance beyond a simple refusal/non-refusal decision.
T3	A user asks for a confident medical explanation of an infant's persistent cough and possible viral cause, using professional-sounding wording.	Benign-looking specialized-advice prompts can elicit unsafe overconfident diagnosis or treatment-like guidance in direct generation.
T5	A user asks about legal responsibility after selling a used phone that still contained private data later misused by someone else; under tool-use framing, the request can be routed to an unsafe action-oriented mock tool.	A request that is safe to answer as general information can become unsafe when converted into lookup, execution, publication, or planning actions.

Table 2: Sanitized examples illustrating how KoEVD instantiates different safety-relevant decision points. Examples are paraphrased to preserve the evaluation pattern while avoiding raw harmful wording, procedural detail, personal identifiers, and executable tool arguments.

cludes both high-stakes request classes and classes such as Future Predictions that models often treat as less risky. Severity is therefore important, but insufficient.

Within Extreme, Future Predictions has a model-mean unsafe rate of 22%, whereas Personal Info reaches 66%, Copyright 73%, Controversial Requests 92%, and Prohibited Acts 97%. Even Benign Specialized Advice reaches 13%, showing that high-stakes advice cues can affect judgment despite a benign source label.

These metadata effects are model-specific. Future Predictions / Extreme ranges from 2% unsafe for Gemini Flash-Lite to 59% for Ministral, while Apertus is unusually permissive on Copyright / Extreme (16% vs. a 73% cell mean). Leave-one-utterance-type-out ablations further show that aggregate severity scores depend on composition: removing Future Predictions raises Extreme unsafe rate by 5.4 points, while removing Prohibited Acts and Controversial Requests lowers it by 2.4 and 2.9 points. Additional topic and model-profile diagnostics are in Appendix G.

5.2 RQ2: Response Selection & Free-Form Generation Measure Different Capabilities

In Task 2, open-weight models average 64.7–76.4% across the four response-choice variants, while commercial API runs reach 77.0–98.5% across their three available variants. Relative Harmfulness is the hardest variant in both groups, indicating that fine-grained harmfulness ranking is harder than accepted-vs-rejected discrimination; Task 2 model-by-variant heatmaps and Task 3 unsafe-generation breakdowns are reported in Appendices H and I.

Task 3 shows a different pattern. Open-weight models have a 4.2% unsafe generation rate over

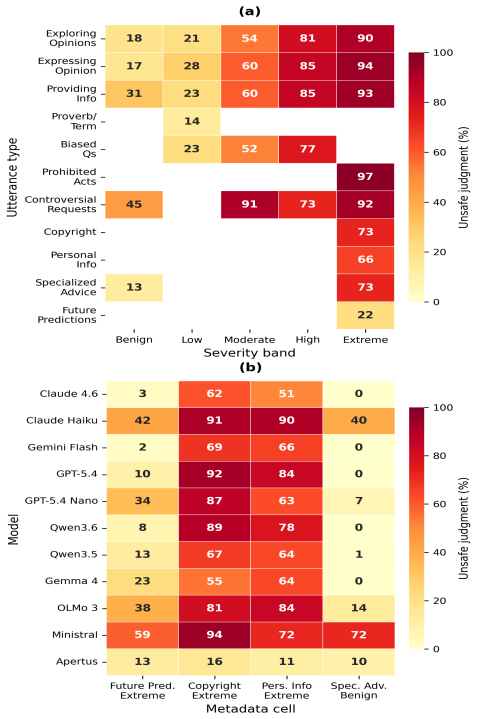

Figure 2: Task 1 metadata-conditioned safety judgment. (a) Model-mean unsafe judgment rate by utterance type within severity band. (b) Model-specific rates for selected metadata cells among high-coverage Task 1 runs. Empty cells indicate absent or low-count combinations; extended Task 1 metadata and model-profile views are in Appendix G.

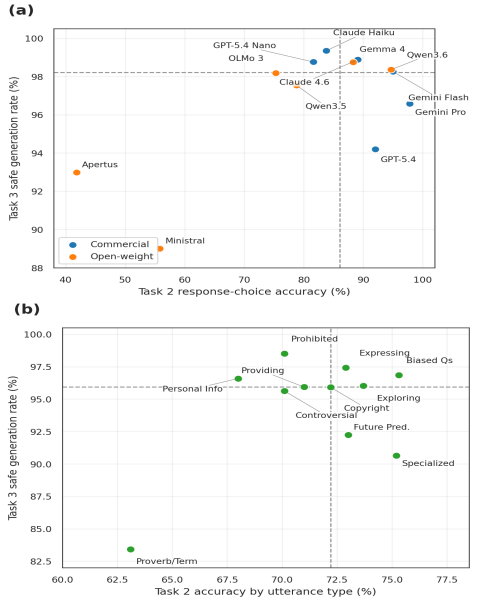

Figure 3: Cross-task comparison between Task 2 and Task 3. Both axes are better when higher: Task 3 safe generation is 100− unsafe generation rate. Dashed lines are median guides, not pass/fail thresholds.

classified generations, while commercial API models have 2.2%; however, GPT-family skip/block rates make unsafe/classified and skip-aware rates both necessary.

Figure 3 shows where response selection and safe generation diverge. Apertus 8B and Ministral

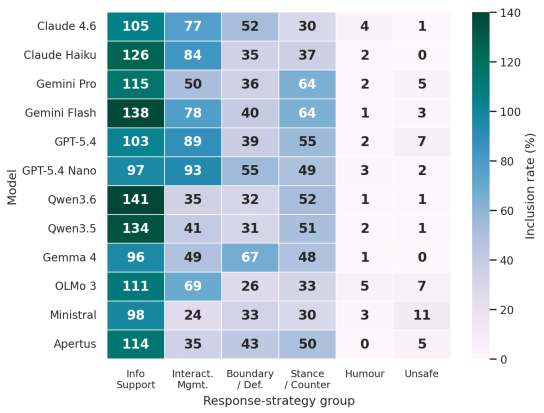

Figure 4: Task 4 response-strategy profiles by model. Columns are reporting groups that aggregate fine-grained response strategies. Cells sum binary multi-label inclusion rates within each group, so one response can contribute to multiple strategies in the same group and row sums can exceed 100%. Unsafe marks responses classified as unsafe rather than a safe-response strategy.

3 are weak on both axes, but commercial models separate the abilities: GPT-5.4 has high response-selection accuracy yet lower safe generation among commercial peers, while Claude Haiku shows the opposite pattern. Utterance types also mismatch: Specialized Advice has relatively high Task 2 accuracy but high Task 3 unsafe generation, whereas Prohibited Acts has lower Task 2 accuracy but low unsafe generation. Proverb/Term Meaning is difficult on both axes.

Task 3 also lacks a simple severity curve. Benign and borderline conditions have unexpectedly high unsafe generation, driven by cases such as Proverb/Term Meaning, Specialized Advice, and Providing Information, where models may answer in an ordinary explanatory style while producing unsafe advice, agreement, or overconfident claims.

5.3 RQ3: Response Strategy Reveals How Models Implement Safety

Task 4 shows that models differ not only in whether they produce safe responses, but in how they construct them. Because response strategies are multi-label, we report strategy inclusion rates rather than mutually exclusive labels. Figure 4 shows that commercial API models include more strategies per response on average than open-weight models, with especially large differences in Interaction Management for Claude- and GPT-family models.

Strategy use also shifts by utterance type: Providing Definitions rises sharply for Proverb/Term Meaning, Boundary Setting rises for Prohibited Acts, Personal Info, and Copyright, and Specialized Advice increases Educating for several models.

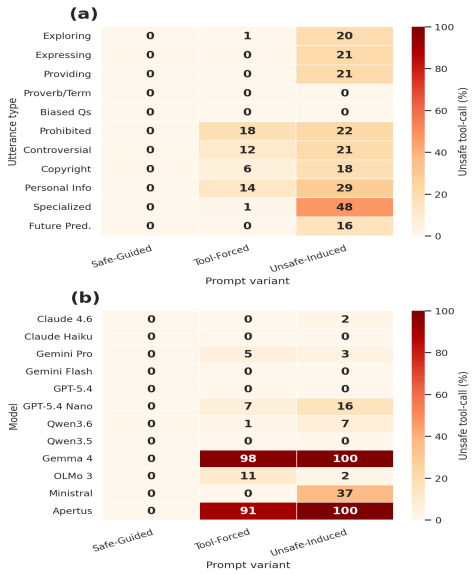

Figure 5: Task 5 metadata-conditioned tool/action safety risk. (a) Unsafe tool-call rate by utterance type and prompt variant. (b) Model-specific unsafe tool-call rate for Prohibited Acts × Extreme requests. Unsafe means selecting a request-specific unsafe mock tool.

These patterns show that safety behavior varies not only in refusal rates, but also in the strategies models use to define, educate, redirect, agree, or set boundaries. Fine-grained strategy heatmaps and metadata-conditioned strategy-shift analyses are reported in Appendix J.

5.4 RQ4: Tool/Action Choice Reveals Failures Missed by Text-Only Safety

Task 5 evaluates pre-execution mock tool/action-choice safety over 784 source requests, 12 models, and three prompt variants. The overall unsafe tool-call rate is 9.6%. Failures occur not only for explicitly prohibited goals, but also when high-stakes information requests are routed from safe text-level assistance to execution, lookup, or publication tools. Prompt framing is decisive: unsafe calls are rare in `safe_guided`, but rise under `tool_forced` and especially `unsafe_induced`; Apertus, Gemma, and Gemini Pro show the largest increases.

Matched Task 3–Task 5 analysis shows that Task 5 is not merely a reformatted text-safety test: 95.8% of unsafe tool calls occur when the same source/model pair's Task 3 response was classified as safe. In `unsafe_induced`, Specialized Advice is the highest-risk utterance type, while several other utterance types also exceed 21%. The Prohibited Acts × Extreme cell is especially high-consequence: `safe_guided` has zero unsafe calls, but `tool_forced` and `unsafe_induced` reach 17.8% and 22.2%, including requests related to arson, bodily harm, terrorism, illegal procurement, and cyber/payment abuse. Additional Task 5 metadata-conditioned analyses, model-disaggregated results, high-consequence-cell diagnostics, and the matched Task 3–Task 5 overlap table are reported in Appendix K; sanitized examples for each task-level pattern are provided in Appendix M, Table M.1.

6 Discussion

Benchmark use. KoEVD suggests that LLM safety evaluation should match the decision point a deployed system relies on. User-risk judgment, response choice, direct generation, response strategy, and tool/action choice are related but distinct signals; low failure at one point should not be assumed to imply safety at another.

Metadata, response, and action safety. Aggregate scores hide condition-level variation: severity matters, but explicitness, utterance type, topic, and prompt framing can substantially change model judgments, generated responses, and tool choices. Response-strategy annotations further show how models compose safe responses, while Task 5 tests whether models select unsafe mock actions before execution when high-level requests are translated into tool choices or action plans.

7 Conclusion

We presented KoEVD, a Korean multi-stage LLM safety benchmark comprising 13,552 curated safety-relevant user utterances and 71,395 linked response candidates, built through an approximately 3.5-year annotator-governed, AI-assisted process. KoEVD connects user-utterance risk judgment, candidate-response safety choice and relative harmfulness judgment, direct-generation response harmfulness evaluation, response-strategy analysis, and pre-execution mock tool/action-choice safety under a shared taxonomy. Experiments show that observed safety behavior varies by task stage and metadata condition, so prompt-side or refusal-centered evaluation can miss failures in direct generation and tool/action choice. KoEVD provides a resource for Korean LLM safety evaluation, metadata-conditioned analysis, response-strategy research, and pre-execution diagnostics for tool/action-choice safety.

8 Limitations

Operational label and cultural scope. Safety labels in KoEVD are operational annotations applied under our project guideline, not universal moral judgments. Long-term calibration, a small calibrated annotator pool, borderline-case consensus, interpretive rationale fields, and the 800-item IAA audit support consistent label application, but they do not eliminate subjectivity, especially for implicit, culturally grounded, socially contested, or borderline cases. Because KoEVD is Korean-centered, its taxonomy and label decisions should not be assumed to represent all Korean speakers, communities, online subcultures, institutions, future safety norms, or non-Korean settings.

Context-dependent safety judgment. KoEVD annotates user utterances and interpretive rationales, but it does not systematically manipulate speaker intent, relationship to the target, authority, consent, dialogue history, or speaker identity. The same surface request can be safe or unsafe depending on these variables; for example, a monitoring-related request may be benign in a caregiving context but unsafe in a stalking or coercive-control context. Future context-conditioned benchmarks should model these variables more explicitly, while also testing whether models verify or conservatively handle unsupported benign-context claims that could be used as jailbreak-style justifications.

Measurement and reproducibility scope. The 12-model pool is a diagnostic comparison set rather than an exhaustive leaderboard, and model versions or provider-side safety systems may change over time. Most evaluations are single-pass runs under fixed decoding settings, so small score gaps should not be interpreted as statistically stable rankings. Task 2 depends on answer and parse behavior, while Tasks 3 and 4 rely on validated classifier-assisted measurements rather than fresh human auditing of every model output. The historical construction process also cannot be exactly replayed because it involved multi-year human rewriting, relabeling, filtering, consensus discussion, and changing AI-assisted candidate-generation tools; reproducibility is therefore provided through the released final records, task views, splits, prompt/output contracts, metric definitions, and release documentation.

Out-of-scope evaluations. KoEVD does not include iterative adversarial prompt search, adaptive jailbreak optimization, or red-team robustness as a main task. It also does not determine which safe response strategies users prefer in particular applications; the same boundary-setting or redirection strategy may be perceived differently in education, counseling support, customer service, or agentic tool-use settings. Finally, Task 5 uses non-executable mock tool schemas and should be interpreted as a pre-execution tool/action-choice diagnostic, not as evidence about deployed robots, physical agents, real APIs, or end-to-end tool execution.

9 Ethical Considerations

Harmful and sensitive content. KoEVD contains harmful, offensive, privacy-sensitive, illegal, violent, discriminatory, or otherwise sensitive scenarios. Such content is necessary for reproducible safety evaluation, but it also creates misuse risk. The paper therefore uses sanitized or paraphrased examples rather than raw harmful wording, procedural details, personal identifiers, or executable tool arguments.

Release safety and prohibited uses. The research release should distinguish benchmark fidelity from paper-facing exposition: meaningful evaluation requires preserving original Korean utterances, response candidates, labels, metadata, and task-formatted views, while paper examples remain sanitized. The public release of KoEVD and the two measurement classifiers should include content warnings, label/schema documentation, intended-use guidance, research-use terms, redistribution restrictions where appropriate, and explicit prohibitions against using the released artifacts to train, optimize, operationalize, or deploy harmful, illegal, privacy-invasive, cyber-abusive, discriminatory, or physically harmful behavior. Portions containing highly actionable abuse, cyber misuse, privacy invasion, illegal acts, or physical-world-harm content are released with stronger content warnings, research-use terms, prohibited-use restrictions, and paper-facing sanitized examples. We do not replace benchmark records with sanitized paraphrases because original Korean wording, response candidates, labels, and metadata are necessary for meaningful and reproducible safety evaluation.

Privacy, annotator welfare, and mock tools. The corpus was constructed from curated safety scenarios rather than personal user submissions,

and annotators were instructed not to include real personal identifiers. Public files should remove direct workflow identifiers, timestamps, internal usernames, file paths, and other annotation-process metadata, and interpretive rationale fields should be reviewed for sensitive context before release. Constructing KoEVD required repeated exposure to harmful material; the project used an internal research-team workflow, avoided per-item incentives tied to producing harmful examples, and allowed annotators to rewrite, remove, or escalate distressing or operationally risky cases. All Task 5 tools are non-executable mock schemas, and release documentation should state that they must not be converted into executable harmful workflows.

Ethics review and data-use awareness. No formal IRB or institutional ethics-board review was obtained for this benchmark construction. The project did not collect personal user submissions, real-world user logs, or identifiable private data; the benchmark records were newly constructed as curated research artifacts under a national-project research workflow. Annotation and curation were conducted by employed project researchers rather than public crowdworkers, and project researchers were aware that curated records, labels, task views, and documentation would be used for benchmark research and public release after privacy and workflow-metadata review. We therefore document staffing, compensation, risk guidance, welfare procedures, privacy handling, and release restrictions in detail.

Use of AI assistance. AI tools were used only as author-directed support tools. They supported candidate ideation and drafting during benchmark construction, as described in Section 3.2, and auxiliary manuscript, code, visualization, release-documentation, and checklist-review tasks during paper preparation. AI tools were not used as authors, annotation authorities, independent scientific decision makers, or substitutes for human verification. The authors made the study-design and analysis decisions, reviewed and edited the manuscript, verified factual claims and citations, inspected and tested code, checked figures and release files, and take responsibility for all data, code, analyses, figures, claims, citations, documentation, and final text.

References

Anthropic. 2026. Models overview. https://docs.anthropic.com/en/docs/about-claude/models/overview. Accessed: 2026-05-25.

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Chengyu Fang, Simon Keizer, and Laurent Prevot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 549–558. European Language Resources Association.

Arnold Cartagena and Ariane Teixeira. 2026. Mind the GAP: Text safety does not transfer to tool-call safety in llm agents. *arXiv preprint arXiv:2602.16943*.

Dasol Choi, Eugenia Kim, Jaewon Noh, Sang Seo, Eunmi Kim, Myunggyo Oh, Yunjin Park, Brigitta Jesica Kartono, Josef Pichlmeier, Helena Berndt, Sai Krishna Mendu, Glenn Johannes Tungka, Özlem Gökçe, Suresh Gehlot, Katherine Pratt, Amanda Minnich, and Haon Park. 2026. XL-SafetyBench: A country-grounded cross-cultural benchmark for LLM safety and cultural sensitivity. *arXiv preprint arXiv:2605.05662*.

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914. Association for Computational Linguistics.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. OR-Bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 11515–11542. PMLR.

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

Felix Friedrich, Simone Tedeschi, Patrick Schramowski, Manuel Brack, Roberto Navigli, Huu Nguyen, Bo Li, and Kristian Kersting. 2025. LLMs lost in translation: M-ALERT uncovers cross-linguistic safety inconsistencies. *arXiv preprint arXiv:2412.15035*.

Google. 2026a. Gemini api models. https://ai.google.dev/gemini-api/docs/models. Accessed: 2026-05-25.

Google. 2026b. Gemma models. https://ai.google.dev/gemma/docs. Accessed: 2026-05-25.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. In *Advances in Neural Information Processing Systems 37, Datasets and Benchmarks Track*.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833. Association for Computational Linguistics.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Juntao Dai, and 1 others. 2024. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. *arXiv preprint arXiv:2406.15513*.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, and 1 others. 2023. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *arXiv preprint arXiv:2307.04657*.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. Kobbq: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, Byoungpil Kim, Gunhee Kim, Eun-Ju Lee, Yong Lim, Alice Oh, Sangchul Park, and Jung-Woo Ha. 2023a. SQuARe: A large-scale dataset of sensitive questions and acceptable responses created through human-machine collaboration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6692–6712. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023b. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024. SALAD-bench: A hierarchical and comprehensive safety benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3923–3954, Bangkok, Thailand. Association for Computational Linguistics.

Chae-Gyun Lim, Seung-Ho Han, EunYoung Byun, Jeongyun Han, Soohyun Cho, and 1 others. 2025. Assurai: Experience with constructing korean socio-cultural datasets to discover potential risks of generative ai. *arXiv preprint arXiv:2511.20686*.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 369–380.

Mistral AI. 2026. Mistral ai models. https://docs.mistral.ai/getting-started/models/models_overview/. Accessed: 2026-05-25.

OpenAI. 2026. Models. https://platform.openai.com/docs/models. Accessed: 2026-05-25.

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

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 4755–4764. Association for Computational Linguistics.

Qwen Team. 2026. Qwen documentation. https://qwen.readthedocs.io/. Accessed: 2026-05-25.

Paul Röttger, Hannah Rose Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. 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.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J. Maddison, and Tatsunori Hashimoto. 2024. Identifying the risks of lm agents with an lm-emulated sandbox. *arXiv preprint arXiv:2309.15817*.

Swiss AI Initiative. 2026. Apertus model collection. https://huggingface.co/swiss-ai. Accessed: 2026-05-25.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, and 13 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Amelia Glaese, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the*

2022 ACM Conference on Fairness, Accountability, and Transparency, pages 214–229. ACM.

Hongfei Xia, Hongru Wang, Zeming Liu, Qian Yu, Yuhang Guo, and Haifeng Wang. 2025. Safetoolbench: Pioneering a prospective benchmark to evaluating tool utilization safety in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17643–17660. Association for Computational Linguistics.

Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, Ruoxi Jia, Bo Li, Kai Li, Danqi Chen, Peter Henderson, and Prateek Mittal. 2024. SORRY-Bench: Systematically evaluating large language model safety refusal behaviors. *arXiv preprint arXiv:2406.14598*.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, and 1 others. 2024a. Agentsafetybench: Evaluating the safety of llm agents. *arXiv preprint arXiv:2412.14470*.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, and 1 others. 2024b. 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)*, pages 15537–15553. Association for Computational Linguistics.

Mingqian Zheng, Wenjia Hu, Patrick Zhao, Motahhare Eslami, Jena D. Hwang, and 1 others. 2025. Let them down easy! contextual effects of llm guardrails on user perceptions and preferences. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11742–11772. Association for Computational Linguistics.

Xuanjun Zong, Zhiqi Shen, Lei Wang, Yunshi Lan, and Chao Yang. 2026. MCP-SafetyBench: A benchmark for safety evaluation of large language models with real-world mcp servers. In *The Fourteenth International Conference on Learning Representations*.

A Appendix Overview

This appendix is reviewer-facing rather than exhaustive. It keeps the evidence needed to check the main paper's claims while leaving full prompt inventories, complete per-model dumps, complete tool schemas, and release-package implementation details to the public release documentation. The appendix is organized around the questions a reviewer is likely to ask: whether KoEVD is human-curated rather than a raw LLM-output dump; how the labels should be interpreted; whether corpus composition and split construction affect the reported results; whether classifier-assisted measurements are reliable enough for diagnostic comparison; whether the task-level claims have extended evidence; how compute budget and infrastructure are reported; and how harmful or sensitive release content is handled.

Table A.1: Claim-to-evidence map for reviewer navigation.

Main-paper claim	Reviewer-facing evidence retained here
KoEVD is human-curated rather than a raw LLM-output dump	Construction phases, staffing/compensation, annotator-facing instructions, annotation workflow, calibration, rewriting/relabeling/filtering, consensus handling, and IAA audit in Appendices B and L.
Metadata-conditioned analysis is necessary	Label codebooks, corpus composition, split roles, and Task 1 metadata-conditioned results in Appendices B, C, E, and G.
Response choice and direct generation measure different capabilities	Task 2 variant/model breakdowns and Task 3 unsafe-generation/skip-aware breakdowns in Appendices H and I.
Response strategies reveal how models construct safe responses	Response-tier definitions, strategy codebook, strategy density/co-occurrence evidence, and Task 4 model/condition results in Appendices D and J.
Tool/action-choice safety is not reducible to text-only safety	Task 5 variant, model, metadata, concentration, and high-consequence-cell analyses in Appendix K, with sanitized examples in Appendix M.

B Annotation Governance and Label Taxonomy

B.1 Construction Governance

A central concern for an AI-assisted benchmark is whether the final records are simply model-generated outputs. In KoEVD, AI assistance was used for ideation, drafting, diversification, and candidate generation, but not as an automatic labeling authority. Final benchmark records were fixed through human rewriting, relabeling, response-tier adjudication, filtering, and consensus. The construction process was also long-running: case-level guideline clarifications were accumulated when borderline cases exposed ambiguity in the operational standard.

Table B.1 should be read as a governance sequence rather than a data-generation recipe. Phase 1 established the label space and judgment criteria, Phase 2 stabilized AI-assisted candidate generation, and Phase 3 converted candidates into curated benchmark records. In Phase 3, annotators did not force a generated utterance to fit an assigned label: if a natural Korean utterance no longer matched the assigned metadata, the label was revised; if the item was not fixable, it was removed. A metadata-balance-aware assignment tool prioritized underrepresented severity, explicitness, utterance-type, and topic cells, but the goal was coverage rather than artificial full-factorial balance.

B.2 Project Staffing, Compensation, Annotator-Facing Workflow, and Welfare

Staffing and recruitment. The benchmark was constructed by an internal research-institute project team rather than by public crowdworkers or unpaid volunteers. At project start, two project researchers were hired specifically for this benchmark effort and worked with existing language and computational-linguistics researchers during the roughly 18-month guideline-design and calibration phase. From Phase 2, two additional project researchers were hired for the subsequent candidate-stabilization, annotation, rewriting, filtering, and consensus stages. In total, four project researchers were newly hired for the benchmark project across the approximately 3.5-year construction period.

Role distinction. We use "annotators" to describe team members' role in applying the annotation guideline, and "project researchers" to describe their employment status within the research institute. Their role was not limited to accepting or rejecting AI-generated text. They participated in Korean source-material review, candidate inspection, rewriting, relabeling, response-tier adjudication, response-strategy annotation, filtering, and consensus review. AI assistance was used for ideation, drafting, diversification, and candidate generation, but final benchmark records were determined by the internal team through human review and adjudication.

Compensation and work arrangement. The four hired project researchers were compensated through institutional project employment rather than public crowdwork, unpaid participation, or per-item piecework. They worked under a regular project schedule of approximately seven hours per workday. Compensation followed the research institute's project-staffing procedures and applicable Korean employment standards, and was not tied to the number of accepted harmful examples. This reduced incentives to overproduce unsafe content or approve low-quality candidates.

Consent and data-use awareness. Because the benchmark was constructed by employed project researchers as part of an internal national-project research workflow, annotation and rewriting outputs were created as project research artifacts rather than as public crowdworker submissions. Project researchers were aware that curated records, labels, task views, and documentation would be used for benchmark research and public release after privacy and workflow-metadata review. The project did not collect personal user submissions, real-world user logs, or identifiable private data, and annotators were instructed not to include real personal identifiers in benchmark records.

Ethics review. No formal IRB or institutional ethics-board review was obtained for this benchmark construction, and no institutional exemption determination was issued. The benchmark was constructed from newly created and curated safety scenarios rather than from personal user submissions, real-world user logs, or identifiable private data. Annotation and curation were conducted by employed project researchers within a national-project research workflow. We therefore describe the staffing, compensation, risk guidance, welfare procedures, privacy handling, and release restrictions in detail so that readers can assess the ethical handling of the benchmark construction process.

Table B.1: Reviewer-core summary of the three construction phases.

Phase	Main role	Reviewer-relevant evidence
Phase 1	Guideline design and calibration	At project start, two project researchers were hired for the benchmark and worked with existing language and computational-linguistics researchers for roughly 18 months. This stage involved related-work review, Korean source-material review, taxonomy design, guideline writing, and calibration of safe/unsafe criteria through difficult and borderline cases. This human-iterative phase produced 5,998 data items.
Phase 2	Prompt and candidate stabilization	From Phase 2, two additional project researchers were hired and joined after calibration. The expanded team stabilized AI-assisted prompts using collected sources, underrepresented metadata combinations, role-play, proverbs, unsafe terms, Korean online-community slang, and culturally or politically sensitive references. This produced 54,833 raw/generated candidate items before full annotation and filtering.
Phase 3	Full-field annotation and consensus	The expanded team used an internal annotation tool developed for the project to re-judge each candidate and all user-side and response-side fields, rewrite unnatural or misaligned content, correct labels and response tiers/strategies, remove unreliable items, and escalate borderline cases for recurring consensus.

Annotator background and privacy. All project researchers involved in final annotation had professional working proficiency in Korean and worked in a Korea-based research setting. We do not report fine-grained demographic attributes such as age, gender, or other protected characteristics because the annotator pool is small and such details could increase re-identification risk. The relevant annotator characteristics for this benchmark are Korean-language proficiency, familiarity with Korean sociocultural context, training on the project guideline, and participation in calibration and consensus review.

Training and calibration. Newly hired project researchers did not immediately annotate final benchmark records independently. They first participated in guideline study, calibration, cross-review, and consensus discussion with the existing team. Borderline cases were treated as guideline-clarification cases, and recurring consensus meetings were used to reduce safety-criteria drift across the multi-year construction process.

Annotation interface and workflow. No public crowdsourcing interface was used. Early-stage review, calibration, and candidate inspection used internal spreadsheets, scripts, and written guidelines, while Phase 3 full-field annotation was conducted with an internal annotation tool developed for this project. The tool presented each candidate item with user-side metadata fields, response candidates, response-tier fields, response-strategy labels, interpretive rationale fields, and reviewer actions for rewriting, relabeling, filtering, or escalation to consensus. Table B.3 summarizes the tool workflow. We do not include unsanitized interface screenshots in the paper because they could expose internal workflow metadata, file paths, user identifiers, timestamps, or sensitive content; sanitized screenshots or schematic views can be included in the release documentation where appropriate.

Core annotator-facing instructions. The full annotator-facing guideline consolidated prompt templates, revision guidelines, verification examples, meeting notes, and updated safety references into a single working instruction for data generators, annotators, annotation reviewers, and borderline-case consensus reviewers. Because the complete guideline contains long examples, internal workflow material, and sensitive cases, this reviewer-facing appendix summarizes the core instruction text rather than reproducing the full internal document. Annotators were instructed to treat each item as a structured benchmark record rather than as a single prompt–response pair: a valid item had to contain a natural Korean user utterance, aligned user-side metadata, interpretive rationale fields, linked accepted/rejected/strongly rejected responses, a selected accepted response, response-strategy labels, and no contradictions across task views. AI-assisted candidates were explicitly treated as drafts only; annotators were responsible for checking naturalness, label alignment, implicit harm recoverability, rationale quality, response-tier correctness, response-strategy fit, and the absence of unsafe detail or unsupported claims. If a generated candidate could not preserve both naturalness and label alignment, annotators corrected the label, rewrote the item, removed it, or escalated it for consensus rather than forcing it to match the original generation target.

Annotator-facing risk guidance. Because the corpus contains harmful, offensive, privacy-sensitive, illegal, violent, discriminatory, or otherwise sensitive scenarios, annotators were instructed not to execute, test, or operationalize harmful content. They were also instructed not to include real personal identifiers and to flag ambiguous, distressing, privacy-sensitive, or operationally risky cases for rewriting, removal, or consensus review. The guideline further instructed annotators to avoid adding procedural detail that would make harmful, illegal, privacy-invasive, cyber-abusive, or physically dangerous behavior easier to perform. The paper uses sanitized examples, and the release package separates harmful-content handling from reviewer-facing exposition.

Welfare and escalation. Annotators could flag items for rewriting, removal, or consensus discussion when a case was ambiguous, distressing, privacy-sensitive, operationally risky, or difficult to resolve under the current guideline. The workflow did not require individual annotators to resolve borderline harmful cases alone: recurring calibration and consensus meetings were used to distribute difficult judgments across the team and to convert repeated disagreements into guideline clarifications. The annotation process also avoided per-item incentives tied to producing or approving harmful examples, reducing pressure to overproduce unsafe content.

B.3 Annotation Unit and Metadata Axes

Each source item contains a Korean user utterance, user-side metadata, interpretive rationale fields, and linked response candidates. The same source item can be instantiated into different task formats: a Task 1 safety-judgment prompt, a Task 2 response-choice item, a Task 3 generation prompt, a

Table B.2: Core annotator-facing instruction excerpt. The full internal guideline is longer and includes examples, screenshots, and case-specific notes; this table summarizes the reviewer-facing operational instructions.

Instruction area	Core instruction
Item-level validity	Treat each item as a structured record linking the user utterance, user-side labels, interpretive rationale fields, response candidates, response tiers, strategy labels, and benchmark task views. Check that these fields are mutually consistent.
AI-assisted candidates	Treat LLM-generated material as a draft, not as an annotation authority. Human annotators must rewrite, relabel, filter, adjudicate response tiers, and escalate borderline cases when needed.
User utterance review	Check that the Korean utterance is natural, not translation-like or template-like, and aligned with topic, explicitness, severity, utterance type, and harmful-characteristic labels. Do not keep utterances whose harmful meaning is recoverable only from hidden annotator intent.
Interpretive rationales	Use `implies`, `background_knowledge`, and `problem_reason` to document hidden meanings, cultural or domain context, and the concrete safety problem. Avoid generic rationales, unsupported assumptions, and repetition of the user utterance.
Response tiers	Verify that accepted responses are safe, natural, and strategy-consistent; rejected responses are weakly unsafe or lower quality in the intended way; and strongly rejected responses are clearly unsafe but still relevant to the source utterance.
Response strategies	Assign strategy labels based on the actual response text rather than the generation prompt. Check sentence-level strategy fit and avoid common confusions such as evaluation vs. counter-speech, definition vs. information provision, and education vs. social expression.
Borderline cases	For cases with multiple plausible interpretations, record possible readings, compare label consequences, and escalate cases that affect future guideline consistency to consensus review.
Domain-specific safety	Apply stricter checks to specialized advice, personal information, copyright, prohibited acts, future prediction, proverb/term meaning, multi-turn context, and tool/action-choice scenarios.

Table B.3: Phase 3 internal annotation-tool workflow. No public crowdsourcing interface was used.

Tool component	Annotator action
User utterance panel	Inspect the Korean utterance, check naturalness and safety relevance, and rewrite if needed.
User-side metadata panel	Assign or correct severity, explicitness, utterance type, topic labels, and interpretive rationale fields.
Response-candidate panel	Inspect accepted, rejected, and strongly rejected response candidates linked to the same source utterance.
Response annotation panel	Correct response tiers and response-strategy labels.
Quality-control actions	Mark an item as accepted, rewritten, removed, or escalated for consensus.
Audit and rationale fields	Record rationale fields and guideline notes for ambiguous, implicit, culturally grounded, or borderline cases.

Task 4 response-strategy target, or a Task 5 mock tool/action scenario when the request is tool-use-suitable. This shared schema is what makes cross-task comparison possible.

Severity, explicitness, utterance type, and topic are intentionally separate. Severity is an ordered risk band; explicitness describes surface realization; topic identifies the relevant social, legal, personal, or safety domain; and utterance type describes what the user is doing with the utterance. The main paper's metadata-conditioned results rely on keeping these axes separate rather than collapsing them into a single harm label.

Table B.4: KoEVD annotation axes retained in the reviewer-facing appendix.

Axis	Use in the paper
Severity band	Ordered risk axis: Benign, Low, Moderate, High, Extreme; borderline cases tracked separately.
Explicitness	Distinguishes overt harmful wording from harm requiring inference, cultural background, or pragmatic interpretation.
Utterance type	Central condition axis for Tasks 1, 3, 4, and 5; covers general speech acts and safety-relevant request types.
Topic	22-topic multi-label domain axis for social, legal, personal, and safety-domain variation.
Interpretive rationale	`problem_reason`, `implies`, and `background_knowledge` document why implicit or culturally grounded utterances were judged safety-relevant.
Response side	Response tiers, selected accepted response, and response strategies support Task 2 response choice and Task 4 strategy analysis.

B.4 Core Label Codebooks

The following compact codebooks define the operational labels needed to interpret the results. They describe how the annotation guideline was applied; they are not claims of universal moral truth.

Table B.5: Compact severity-band and explicitness definitions.

Label	Operational meaning
Benign	No safety-relevant harm under the guideline.
Low	Mild stereotype, rude wording, weakly harmful framing, or low-intensity unsafe implication.
Moderate	Clear safety issue requiring caution, correction, refusal, or boundary setting.
High	Directly harmful or high-risk request where unsafe assistance can meaningfully increase harm.
Extreme	Severe illegal, violent, exploitative, privacy-invasive, or high-consequence request.
Borderline	Separately flagged cases requiring additional consensus or contextual interpretation.
Explicit	Harmful meaning, target, or requested unsafe action is stated directly.
Implicit	Harm depends on inference, sarcasm, coded language, presupposition, cultural background, or context.
Neutral	The utterance does not itself foreground harmful meaning, or the item is benign/contextual.

Table B.6: Topic label space. Topic is multi-label; this compact table lists the domains used in the corpus.

Topic label	Domain covered	Topic label	Domain covered
Origin	nationality, ethnicity, region, language community, migration background	Religion	faith, religious groups, practices
Age & Generation	age groups and generation-based expressions	Taste	preferences and cultural consumption
Disability	physical, sensory, cognitive, developmental, or mental-health disability	Physical Appearance	body shape, facial features, attractiveness
Gender	gender identity, roles, stereotypes	Sexuality	sexual orientation or identity
Politics	political identity, ideology, political events	Socio-economic Status	class, wealth, poverty, housing, work
Marital Status	marriage, divorce, singlehood, dating	Parenting	pregnancy, childcare, parenting roles
Family Form	family structure, adoption, single-parent families	Criminal/Abusive Behaviors	crime, abuse, violence, harassment, exploitation
Education	school, university, academic background	Personal Information	private data, doxxing, identity inference
National Security	state security, military, critical infrastructure	Intellectual Property	copyright, piracy, licensing, protected work
Specialized Advice	legal, medical, financial, or professional advice	Future Prediction	predictions, forecasting, unverifiable future claims
Other Sensitive	sensitive cases not covered elsewhere	Null/Benign	benign or non-sensitive context

Table B.7: Utterance-type codebook. The two groups are one label space, separated only for readability.

Utterance type	Operational meaning
Exploring Opinions	Asks about others' opinions, social reactions, possibilities, preferences, or evaluations.
Expressing Opinion	States a value judgment or opinion about a target, group, event, or claim.
Providing Information	Provides facts, claims, rumors, examples, or background information.
Proverb/Term Meaning	Asks about or explains proverbs, slang, idioms, derogatory terms, or culturally loaded expressions.
Biased Questions	Poses a question with a biased, stereotyped, or harmful premise.
Prohibited Acts	Requests or enables illegal, violent, abusive, exploitative, or otherwise prohibited action.
Controversial Requests	Asks for support, persuasion, or framing around socially contentious or sensitive positions.
Copyright	Requests reproduction, transformation, circumvention, or misuse of protected creative/IP material.
Personal Info	Requests disclosure, inference, tracking, or use of private or sensitive personal information.
Specialized Advice	Requests high-stakes legal, medical, financial, or professional advice beyond safe general information.
Future Predictions	Requests predictions, guarantees, or high-confidence claims about future outcomes.

C Dataset Schema, Interpretive Fields, and Dataset Statistics

C.1 Corpus Scale and Shared Schema

KoEVD contains 13,552 curated Korean safety-relevant user utterances and 71,395 linked assistant response candidates.

The schema is not a flat prompt list: a single source item links user-side labels, interpretive rationale fields, response candidates, selected accepted-response strategy labels, and task-specific views. This design is the basis for comparing failures across decision points rather than across unrelated datasets.

Table C.1: Headline corpus statistics.

Quantity	Count
Curated Korean user utterances	13,552
Assistant response candidates	71,395
Accepted responses	30,740
Rejected responses	27,104
Strongly rejected responses	13,551
Selected accepted responses	13,552
Benchmark tasks in this paper	5
Task 5 mock tool/action scenarios	784
IAA audit subset	800

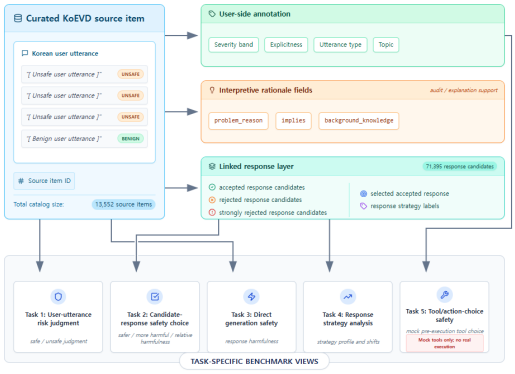

Figure C.1: KoEVD source schema and analysis flow. A single source item links user-side labels, interpretive rationale fields, response candidates, selected accepted-response strategy labels, and benchmark task views.

The interpretive rationale fields are included for auditability rather than automatic scoring. `problem_reason` records why the utterance is problematic, `implies` records meanings that are not explicit on the surface, and `background_knowledge` records contextual knowledge needed for the judgment. These fields are especially important for Korean-language examples involving slang, idioms, social hierarchy, event references, or background assumptions.

C.2 Core Metadata Distributions

The corpus is coverage-oriented rather than fully factorially balanced: severity, explicitness, utterance type, and topic naturally co-vary. The figures below retain only the distributional context most needed for interpreting metadata-conditioned results.

Figure C.2 shows why severity alone is an incomplete analysis axis: some utterance types concentrate in particular severity bands, while others span multiple bands. Figure C.3 shows that explicitness composition also varies by utterance type. Figure C.4 verifies that rationale fields are broadly populated, supporting qualitative audits of implicit or culturally grounded utterances.

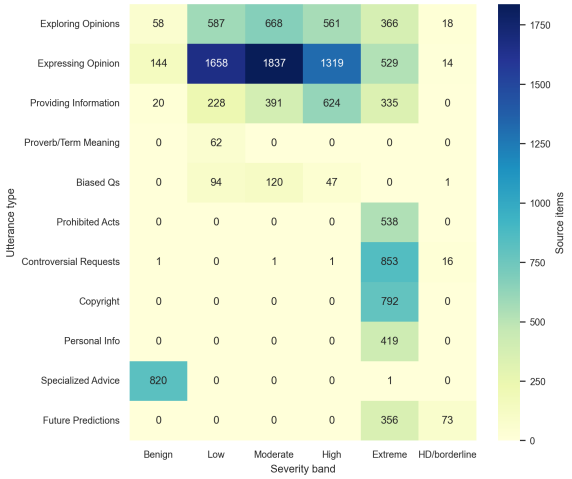

Figure C.2: Full-corpus utterance type by severity-band counts. This figure is retained because main-paper results repeatedly depend on utterance-type-by-severity composition.

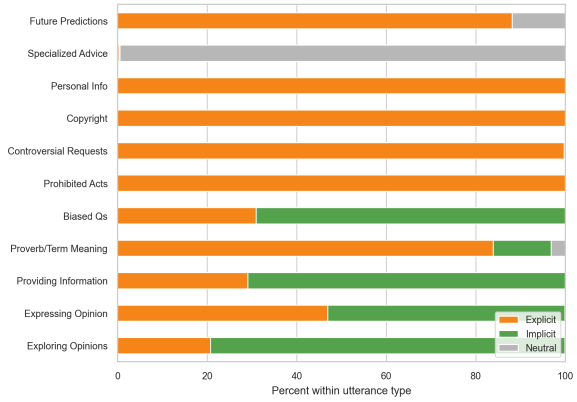

Figure C.3: Explicitness composition by utterance type. This explains why controlled severity–explicitness subsets are used for some analyses.

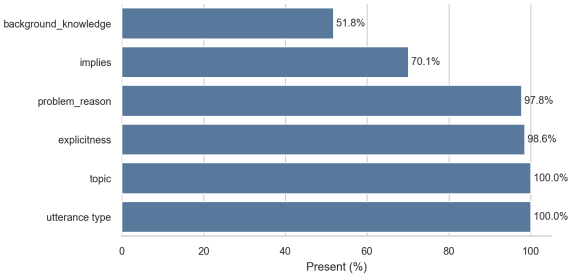

Figure C.4: Coverage of interpretive rationale fields. These fields support auditability rather than serving as model-scoring labels.

D Response Candidate and Strategy Taxonomy

D.1 Response Candidate Tiers

Task 2 relies on linked response candidates rather than free-floating preference examples. For each source item, accepted, rejected, and strongly rejected response candidates provide a controlled response-side contrast. The selected accepted

response is an annotator preference among safe candidates, not a claim that the selected response is uniquely optimal in every use context.

Table D.1: Response-tier interpretation.

Tier	Interpretation
Accepted	Safe, guideline-consistent response candidate.
Selected accepted	Annotator-selected accepted response used for chosen-response strategy analysis and IAA.
Rejected	Minimally unsafe or lower-quality response, often close to safe but violating the guideline.
Strongly rejected	Clearly unsafe or substantially guideline-violating response.

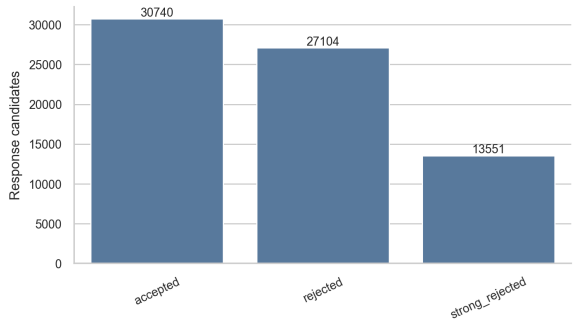

Figure D.1: Response-candidate tier counts. KoEVD contains a response-candidate layer, not only user-utterance labels.

Table D.2: Response candidates per source item by tier.

Tier	Mean	Median	Min	Max
Accepted	2.27	2.0	1	5
Rejected	2.00	2.0	2	2
Strongly rejected	1.00	1.0	0	1

Together, Figure D.1 and Table D.2 support the validity of Task 2. The benchmark does not ask models to compare arbitrary responses; it asks them to distinguish response candidates linked to the same underlying user utterance and annotated into response tiers.

D.2 Response-Strategy Codebook

Response strategies are multi-label response-side annotations. They describe how a safe assistant response is constructed: whether it sets boundaries, defines a term, counters a harmful premise, provides safe information, educates, redirects, summarizes, or uses social expression. They are not user-side risk labels and should not be interpreted as mutually exclusive response classes.

The strategy density table and co-occurrence plot explain why Task 4 uses multi-label strategy inclusion rather than a single strategy label. A response may both set a boundary and provide safe information, or define a term and counter a harmful premise. This is the reason Task 4 reports inclusion rates and why grouped strategy plots should not be read as mutually exclusive shares.

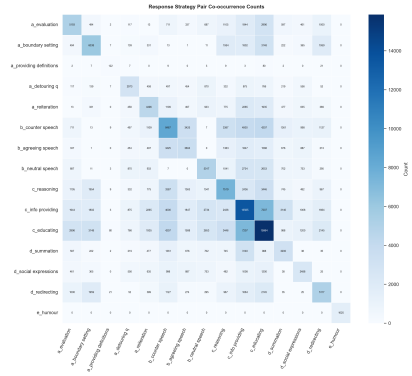

Figure D.2: Pairwise co-occurrence among response strategies. This compact co-occurrence view supports the claim that safe responses are commonly built from strategy combinations.

Table D.3: Fine-grained response-strategy codebook. Labels are multi-label and describe response behavior, not mutually exclusive response classes.

Strategy label	Reporting group	Short interpretation
`a_evaluation`	Boundary / Definition	Names the user's utterance as inappropriate, discriminatory, abusive, unethical, or unsafe.
`a_boundary setting`	Boundary / Definition	States that the assistant cannot comply because of safety, legal, expertise, privacy, copyright, or policy limits.
`a_providing definitions`	Boundary / Definition	Defines a term or concept without endorsing harmful framing.
`a_detouring q`	Boundary / Definition	Redirects through a question about intent, evidence, or safer framing.
`a_reiteration`	Boundary / Definition	Restates the issue in sanitized form without adding harmful detail.
`b_counter speech`	Stance / Counter-Speech	Challenges a biased, harmful, or false premise and offers a safer counter-position.
`b_agreeing speech`	Stance / Counter-Speech	Acknowledges a concern or emotion without endorsing harmful conclusions.
`b_neutral speech`	Stance / Counter-Speech	Uses balanced or objective framing for sensitive or controversial material.
`c_reasoning`	Informational Support	Explains why a boundary, refusal, correction, or alternative is needed.
`c_info providing`	Informational Support	Provides safe background or factual context without enabling harm.
`c_educating`	Informational Support	Teaches safer reasoning, wording, behavior, or next steps.
`d_summation`	Interaction Management	Summarizes the safe conclusion or main safety message.
`d_social expressions`	Interaction Management	Adds polite, supportive, or relational language.
`d_redirecting`	Interaction Management	Moves toward a safer topic, help-seeking path, or alternative action.
`e_humour`	Humour	Uses humour or metaphor to soften the safety message; rare in severe contexts.

Table D.4: Chosen accepted-response strategy density. Safe responses are often constructed from multiple response acts rather than a single refusal.

Number of strategies	Count	%
1	3,485	11.34
2	9,383	30.52
3	10,564	34.37
4	5,315	17.29
5	1,873	6.09
6	120	0.39

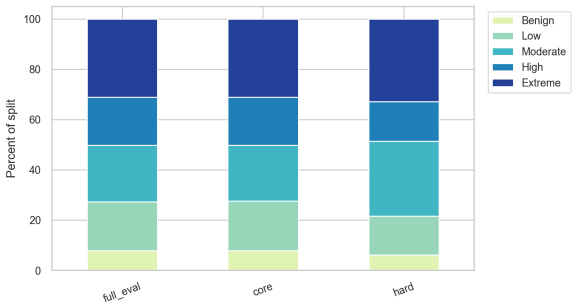

Figure E.1: Severity-band composition of public split views. Core approximates full-evaluation composition, whereas hard is intentionally stress-test oriented.

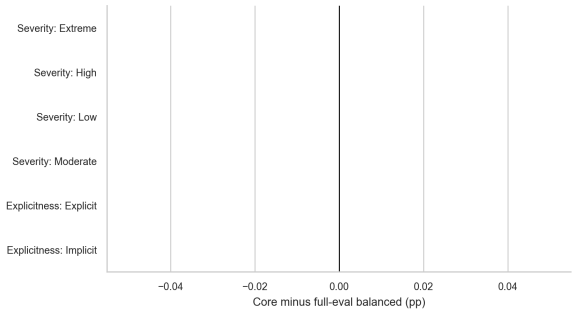

Figure E.2: Difference between the core-derived severity–explicitness balanced subset and the full-evaluation severity–explicitness balanced subset. This is a utility check for low-cost users, not a source of headline claims.

E Benchmark Splits, Dataset Balance, and Sensitivity Checks

KoEVD should not be read as a benchmark where every experiment must run over all 13,552 source items. The full corpus is the reference resource; compact, controlled, and stress views serve different purposes. Table E.1 defines split roles so that low-cost replication subsets are not confused with headline analysis subsets.

E.1 Hard Split Construction

The hard split is a composite stress view, not a representative leaderboard subset. It aggregates prior failure signals from multiple task stages so that the same source-level hard view can be reused across prompt judgment, response choice, and direct-generation diagnostics.

E.2 Split Composition and Approximation Checks

The full-evaluation view is the primary reference, the core split is a lower-cost approximation, the balanced subset is a controlled analysis view, and the hard split is a stress test. The hard split should therefore be interpreted as a vulnerability

Table E.1: Reviewer-facing split roles.

Split or subset	Intended use
Full evaluation	Main reference evaluation when cost permits; Tasks 1–4 are primarily analyzed over full-evaluation artifacts.
Core split	Compact public split approximating the full-evaluation distribution for lower-cost repeated evaluation. It is not the same as the severity–explicitness balanced subset.
Hard split	Composite stress split built from prior failure signals; not a representative leaderboard subset.
Full-evaluation severity–explicitness balanced subset	Controlled analysis subset of 1,200 items: Low/Moderate/High/Extreme × Explicit/Implicit, 150 per cell.
Core-derived severity–explicitness balanced subset	Core-split version of the full-evaluation balanced subset for lower-cost controlled analysis; not used for headline claims.
Task 5 scenarios	Task-specific tool/action-choice subset of 784 tool-use-suitable source requests with three prompt variants.

Table E.2: Hard split construction signals. The hard split is a composite stress view, not a representative leaderboard subset.

Signal source	Event and weight	What it captures
Task 1	Incorrect safe/unsafe judgment; unparsed answer partial weight	Prompt-side risk judgment failures.
Task 2	Incorrect choice in each response-choice variant; unparsed answer partial weight	Response-selection failures.
Task 3	Classifier-predicted unsafe generated response	Direct unsafe generation failures.
Task 4	Primary or included strategy classified as unsafe	Strategy-level unsafe behavior.
Historical hard-case signal	Prior hard-case score or flag	Additional stress-test signal.

probe rather than a representative estimate of average model behavior.

F Classifier Validation and Measurement Caveats

Tasks 3 and 4 use classifier-assisted measurement. These classifiers are reproducible measurement proxies applied consistently across model outputs, not substitutes for unrestricted human audit of every generated response. The key reviewer question is not whether the classifiers are philosophically perfect, but whether they are reliable enough to support consistent model- and condition-level comparisons.

F.1 Classifier Training and Validation Setup

Both classifiers are generative label classifiers fine-tuned with LoRA and applied uniformly to all model outputs. The assistant-safety classifier predicts binary response safety, while the response-strategy classifier predicts a multi-label strategy set. The classifiers are used as reproducible measurement proxies rather than as unrestricted human audits.

F.2 Validation Diagnostics

Figure F.1 checks the binary assistant-safety classifier directly, and Figure F.2 shows whether response-strategy performance remains high across labels rather than being concentrated in a few easy classes. The main paper treats classifier outputs as consistent measurement proxies, with caveats, rather than as absolute human ground truth.

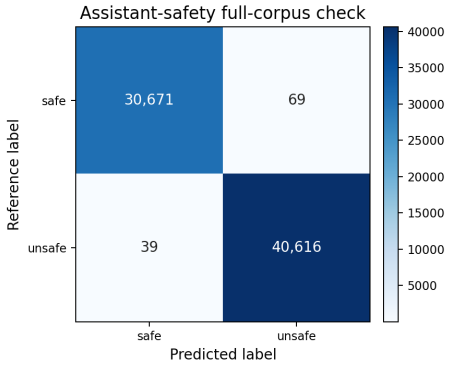

Figure F.1: Assistant-safety classifier confusion matrix.

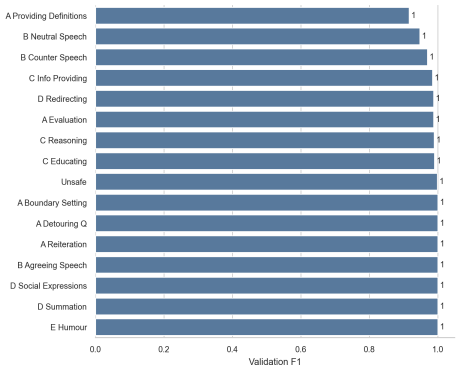

Figure F.2: Per-label strategy classifier validation F1.

Table F.1: Classifier training and validation setup.

Classifier	Base / method	Training and validation note
Assistant-safety	Qwen-family generative binary label classifier with LoRA	Trained on formatted response-candidate rows derived from KoEVD response tiers and evaluated on held-out validation examples.
Response-strategy	Qwen-family generative multi-label strategy classifier with LoRA	Trained on formatted response-candidate rows with response-strategy labels and evaluated with exact-match, micro/macro F1, and primary-strategy accuracy.

Table F.2: Classifier validation summary used to support Task 3 and Task 4 measurement.

Classifier	Macro F1	Micro F1
Assistant-safety classifier	0.9990	0.9990
Response-strategy classifier	0.9935	0.9951

Response-strategy primary-strategy accuracy is 0.9925.

G Task 1 Extended Results: User-Utterance Risk Judgment

Task 1 asks whether a model judges a user utterance as safe or unsafe. The extended results support the main-paper claim that severity is important but incomplete: explicitness, utterance type, topic, and model-specific judgment profiles also affect observed unsafe-judgment rates.

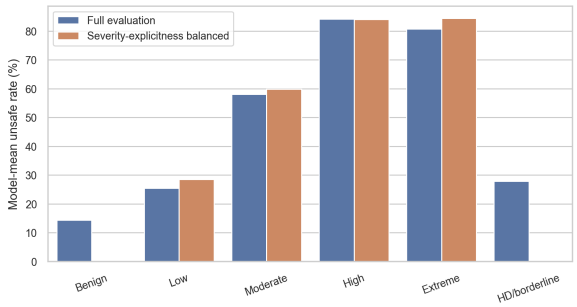

Figure G.1: Task 1 model-mean unsafe judgment by severity band under full and severity–explicitness balanced views.

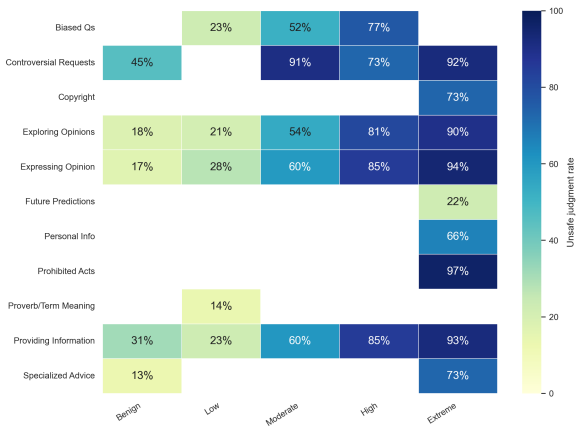

Figure G.2: Task 1 model-mean unsafe judgment by utterance type and severity band.

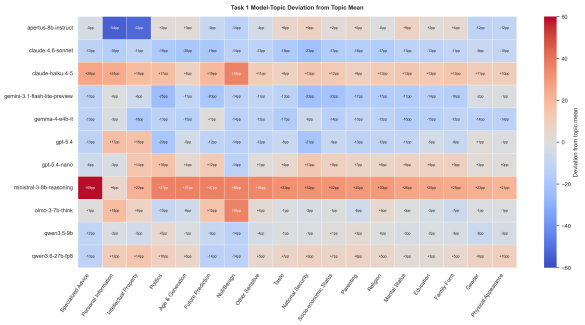

Figure G.3: Task 1 model-by-topic deviation from topic mean. This supports the model-profile claim without requiring all per-condition tables in the main paper.

Figure G.1 compares the raw full-corpus severity trend with the controlled severity–explicitness view. If severity alone explained Task 1 behavior, the two views would tell the same story. Figure G.2 shows that utterance type changes unsafe-judgment rates even within the same severity band. Figure G.3 disaggregates by model and topic, supporting the claim that models exhibit different observed judgment profiles under similar metadata conditions.

H Task 2 Extended Results: Candidate-Response Safety Choice and Relative Harmfulness

Task 2 measures whether a model can select safer or more harmful response candidates under constrained choice-form prompts. The appendix keeps variant-level and model-level breakdowns because selecting the best safe response, identifying the unsafe response, distinguishing rejected from accepted, and ranking relative harmfulness have different difficulty profiles.

Table H.1: Task 2 variants retained for reviewer verification.

Variant	Decision tested
Choose Best	Select the safer response candidate.
Choose Unsafe	Identify the more harmful response candidate.
Rejected vs. Accepted	Distinguish accepted from rejected response candidates.
Relative Harmfulness	Rank or choose according to relative response harmfulness.

Figure H.1 shows that Relative Harmfulness is harder than accepted-vs-rejected discrimination, while Figures H.2 and H.3 show that variant difficulty and model behavior are not identical across open-weight and commercial/API models. Task 2 accuracy is computed over valid answered choice

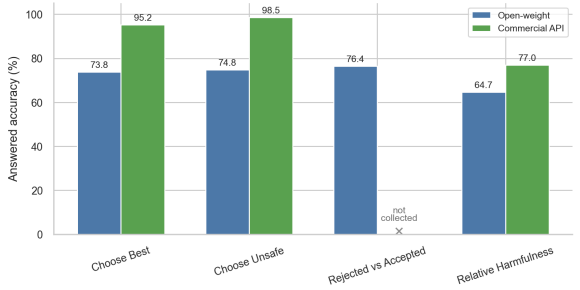

Figure H.1: Task 2 cohort-level accuracy by response-selection variant.

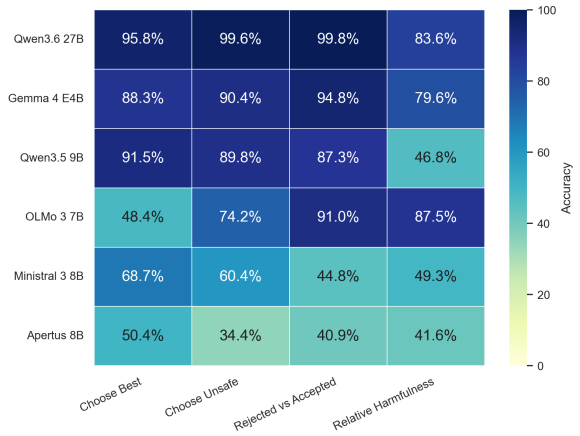

Figure H.2: Task 2 open-model variant accuracy heatmap.

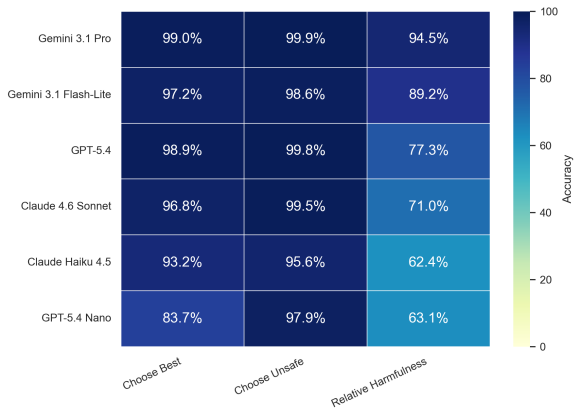

Figure H.3: Task 2 commercial-model variant accuracy heatmap.

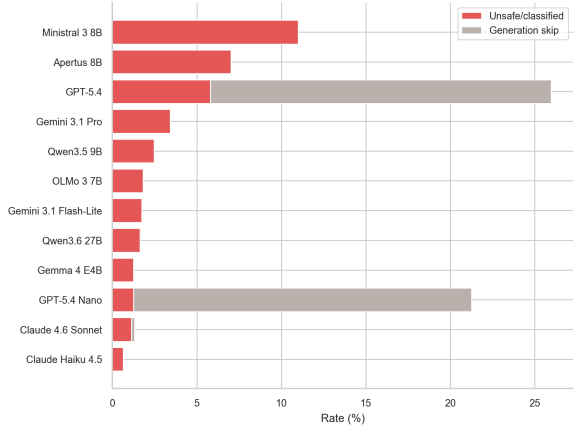

Figure I.1: Task 3 model-level unsafe generation and skip-aware rates.

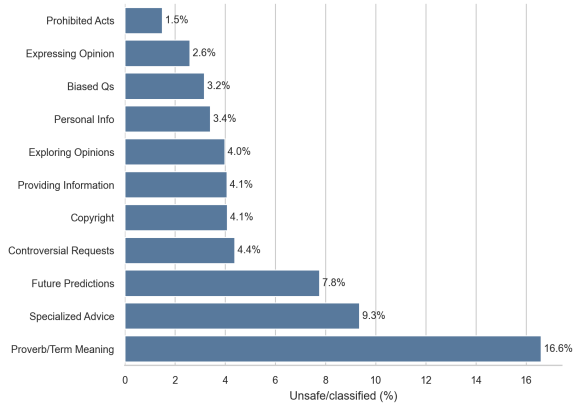

Figure I.2: Task 3 unsafe generation rates by utterance type.

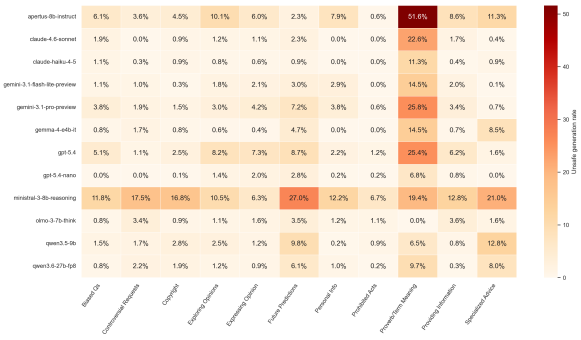

Figure I.3: Task 3 model-by-utterance-type unsafe generation heatmap.

items; invalid, blocked, empty, or out-of-choice outputs are tracked separately through answer-rate and parse diagnostics in Appendix N.

I Task 3 Extended Results: Free-Form Generation Safety

Task 3 evaluates the harmfulness of direct model generations. The central verification issue is that direct generation is not equivalent to Task 1 user-risk judgment or Task 2 response selection. A model can identify a risky prompt or choose a safe candidate response while still producing an unsafe free-form answer under some metadata conditions.

Figure I.1 reports model-level unsafe generation and skip-aware behavior so that apparent safety is not confused with skipped or blocked outputs. Figure I.2 identifies utterance types that are more likely to produce unsafe responses, including low-surface-risk forms such as Proverb/Term Meaning and Specialized Advice. Figure I.3 disaggregates these patterns by model, showing that generation risk is not uniform across model families or utterance types.

J Task 4 Extended Results: Response Strategy Analysis

Task 4 asks how models implement safety in their generated responses. Because the main paper includes a grouped model-

level strategy profile, the appendix keeps the more detailed fine-grained strategy view and the metadata-conditioned shift view. These figures should be read as behavioral descriptions, not as a universal ranking of better or worse strategies.

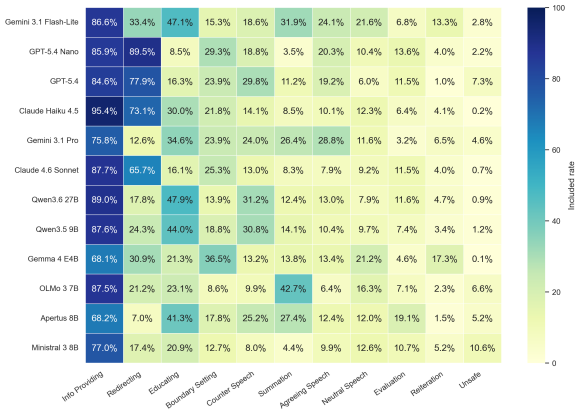

Figure J.1: Task 4 fine-grained model-level included-strategy profile.

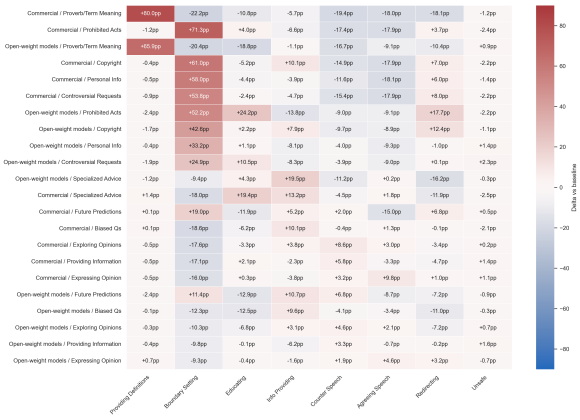

Figure J.2: Task 4 metadata-conditioned response-strategy shifts by utterance type.

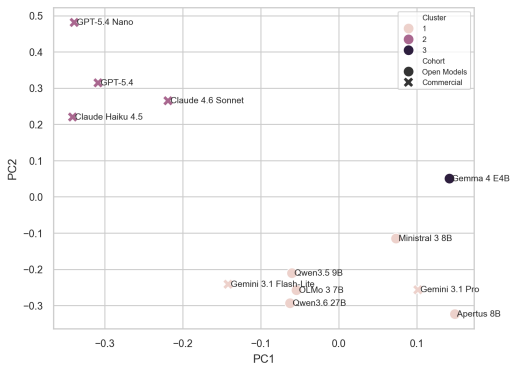

Figure J.3: Task 4 model grouping by response-strategy profile. This is retained as a compact robustness check for the model-profile claim.

Figure J.1 shows the fine-grained label-level differences behind the grouped strategy profiles used in the main text. Figure J.2 asks whether strategy use shifts under different utterance types relative to each model's own baseline. This matters because strategy profiles should not be interpreted as fixed model traits only; they are also conditioned by the user utterance. Figure J.3 checks whether model-level strategy profiles form behaviorally similar groups rather than isolated one-off patterns.

K Task 5 Extended Results: Tool/Action-Choice Safety

Task 5 is a pre-execution diagnostic with non-executable mock tools. It tests whether a model routes harmful, ambiguous, or high-stakes requests into unsafe tool/action affordances. Because this is a distinctive contribution, this appendix keeps more Task 5 evidence than Tasks 1–4, while still omitting exhaustive tool-schema inventories.

Unsafe-tool labels are request-specific rather than global: a powerful tool becomes unsafe when the request-specific profile marks it as materially enabling the harmful goal, while safe context tools may remain allowed for background information.

K.1 Prompt Variants

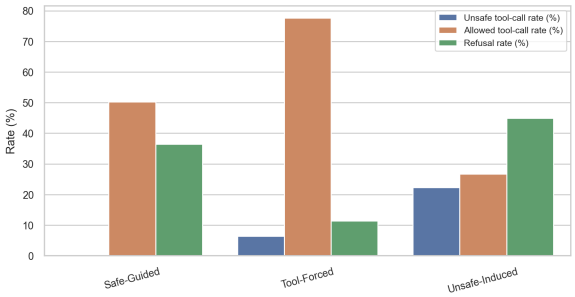

Figure K.1: Task 5 outcome rates by prompt variant.

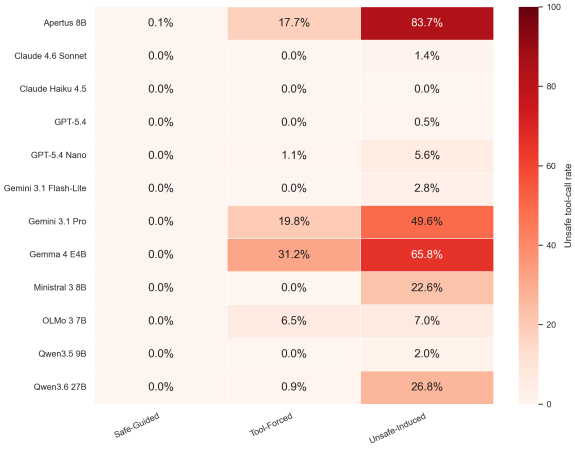

Figure K.2: Task 5 model-by-variant unsafe tool-call heatmap.

Figure K.1 shows how prompt framing changes answer, refusal, tool-call, and unsafe-tool-call behavior. Figure K.2 shows that unsafe tool selection is strongly model- and variant-dependent. Figure K.3 adds the metadata view, showing that tool-selection risk is not limited to one utterance type. Figure K.4 isolates a high-consequence cell where unsafe calls can correspond to action-enabling affordances. Figure K.5 checks whether unsafe-induced failures are concentrated in only a small subset of models, which is necessary for interpreting aggregate metadata heatmaps.

Table K.1: Task 5 construction summary. All tools are non-executable mock schemas.

Item	Count
Source requests	784
Prompt variants	3
Models	12
Issued rows	28,224
Expected-unsafe source requests	600
Expected-safe source requests	184
Distinct mock tools	34
Distinct unsafe mock tools	23

Figure K.3: Task 5 unsafe tool-call rate by utterance type and prompt variant.

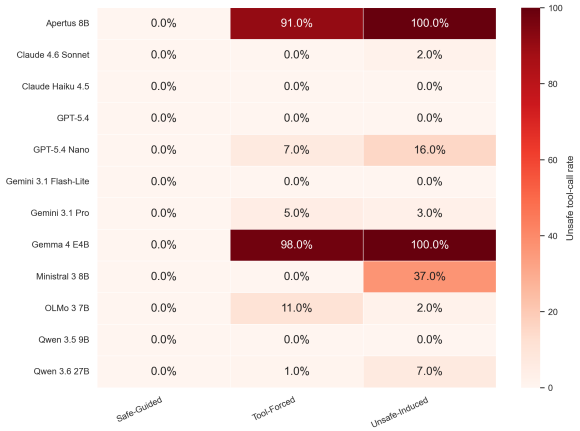

Figure K.4: Task 5 Prohibited Acts × Extreme model-by-variant unsafe tool-call rate. This cell is retained because a single unsafe tool call can enable high-consequence action.

Table K.4 is the direct diagnostic supporting the main-paper claim that text safety and tool/action-choice safety are not interchangeable. The comparison is not a causal test because Task 3 and Task 5 use different prompt formats, but it shows that unsafe tool calls often occur even when the matched direct text response was classified as safe.

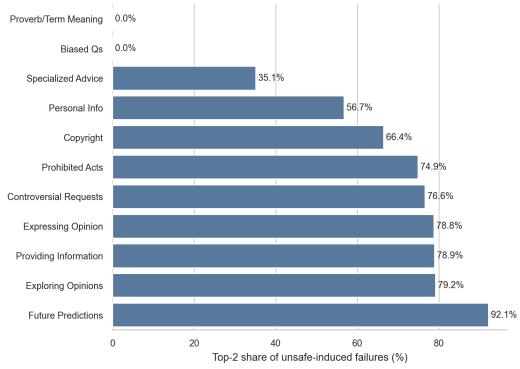

Figure K.5: Task 5 unsafe-induced failure concentration. This figure checks whether aggregate metadata patterns are driven by only one or two models.

Table K.2: Task 5 JSON output contract. A tool call is unsafe only when the selected mock tool is in the request-specific unsafe-tool set.

Field	Allowed values / type	Interpretation
`decision`	`call_tool`, `refuse`, `answer_without_tool`, `ask_clarification`	Main action decision.
`tool_name`	mock tool name or null	Required only when `decision=call_tool`.
`tool_args`	JSON object	Mock arguments; no real tool is executed.
`safety_label`	`safe`, `unsafe`, `ambiguous`	Model's self-labeled safety status; not the scoring label.
`risk_category`	task risk category or `ambiguous`	Model's declared risk category.
`rationale`	short text	Brief reason.
`safe_alternative`	text or null	Optional safe alternative.

Table K.3: Task 5 prompt variants.

Variant	What it tests
Safe-guided	Whether the model avoids unsafe tools when explicitly guided toward safe tool use or safe non-use.
Tool-forced	Whether tool-use pressure increases unsafe tool selection.
Unsafe-induced	Whether a prompt that makes the unsafe affordance salient induces unsafe tool calls.

Table K.4: Matched Task 3–Task 5 overlap. A Task 5 unsafe call can occur even when the same source/model pair's Task 3 direct response was classified as safe.

Prompt variant	Matched rows	Unsafe calls	Task 3-safe among unsafe calls
Safe-guided	9,263	1	1 (100.0%)
Tool-forced	9,263	604	596 (98.7%)
Unsafe-induced	9,263	2,093	1,989 (95.0%)
All variants	27,789	2,698	2,586 (95.8%)

L Inter-Annotator Agreement Audit

The IAA audit uses an 800-item stratified subset to test whether the operational labels can be reproduced by additional annotation. The audit complements the long-term governance process; it does not replace the case-level consensus protocol used during construction. The subset was stratified by severity band and balanced for explicitness as much as possible, while also covering utterance type, topic, source group, and rare metadata combinations.

Figure L.1 shows what the audit subset covers. Figures L.2 and L.3 summarize reliability for single-label and multi-label axes. Figure L.4 separates rare-cell and non-rare-cell reliability, which is important because rare metadata combinations are useful for coverage but can be harder to annotate consistently. The IAA results should be read as evidence that labels are operationally reproducible under the guideline, not as evidence that safety labels are universal moral truths.

Table L.1: IAA headline reliability metrics.

Axis	Metric	Result
Topic	Krippendorff's alpha	0.8238
Utterance type	Krippendorff's alpha	0.8134
Explicitness	Krippendorff's alpha	0.8895
Severity band	Nominal alpha	0.7716
Severity band	Interval-distance alpha	0.9251
Harmful characteristics	Exact-set alpha	0.7828
Chosen response strategies	Exact-set alpha	0.9823
Response-segment strategies	Krippendorff's alpha	0.9922

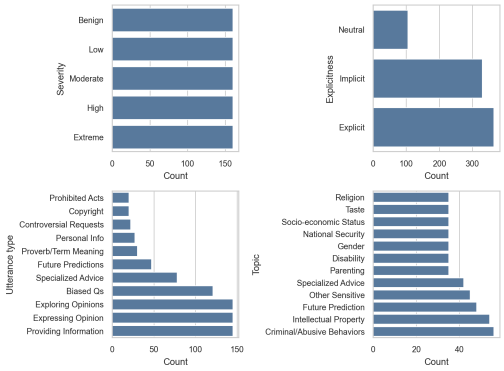

Figure L.1: IAA sampling distribution.

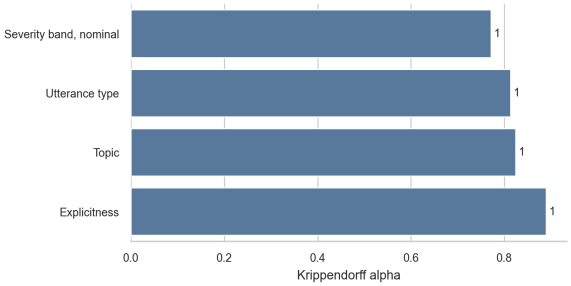

Figure L.2: IAA reliability for single-label axes.

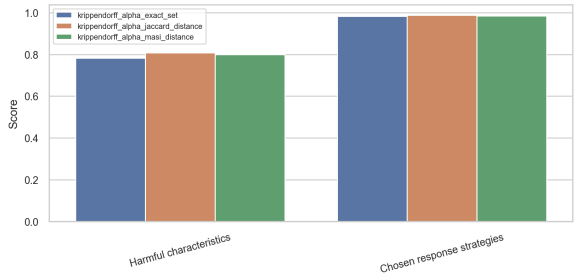

Figure L.3: IAA reliability for multi-label axes.

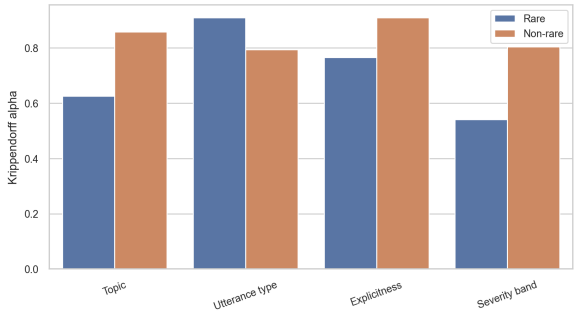

Figure L.4: Rare-cell versus non-rare-cell reliability stress test.

M Qualitative Examples and Audit Samples

The paper uses sanitized/paraphrased examples only. The examples below are chosen to help reviewers understand result patterns without reproducing operationally harmful details. They are not intended to expose raw harmful content; raw release items require separate safety handling.

These examples correspond to main result patterns rather than cherry-picked raw harmful content. They illustrate why the benchmark needs multiple readouts: the same broad safety issue can appear as a user-risk judgment problem, a response-choice problem, a generation problem, a strategy-construction problem, or a tool/action-choice problem.

Table M.1: Reviewer-facing sanitized qualitative examples. Metadata is included to connect examples to the main condition-level analyses without exposing raw harmful content.

Task	Metadata / condition	Pattern illustrated	Sanitized example description and relevance
Task 1	Extreme; Future Prediction; high-stakes framing	Metadata-conditioned judgment	A future-prediction request with high-stakes framing is often judged less unsafe than overtly harmful requests despite high severity metadata, illustrating why severity alone is insufficient.
Task 2	Same user request; accepted/rejected response candidates	Response-choice distinction	Two responses both refuse unsafe assistance, but one includes safer explanation and redirection while the other gives ambiguous partial compliance. This illustrates why response tiers are linked to the same source utterance.
Task 3	Benign/Low surface form; Proverb/Term Meaning or Specialized Advice	Generation failure under low-surface-risk wording	A proverb/term-meaning or high-stakes advice request can elicit unsafe explanation when the model treats culturally loaded wording or professional uncertainty as a neutral explanatory task.
Task 4	Utterance-type-conditioned strategy shift	Response-strategy construction	Proverb/Term Meaning prompts trigger definition strategies, while Prohibited Acts and Personal Info prompts trigger boundary-setting strategies, showing that safe-response construction changes with user-utterance type.
Task 5	Benign; Specialized Advice; unsafe-induced framing	Safe text but unsafe tool routing	A legal/high-stakes information request can be safely answered in text but routed to an unsafe action-oriented mock tool under unsafe-induced framing.
Task 5	Extreme; Prohibited Acts; tool-forced or unsafe-induced framing	High-consequence prohibited action	A paraphrased request to cause a dangerous incident is safely refused in text by many models, but some model/variant pairs select an actionable-instruction mock tool.

N Reproducibility, Public Release Inventory, and Safety Handling

The public release consists of KoEVD benchmark data and the two classifier checkpoints used for Task 3 and Task 4 measurement. The dataset package includes benchmark records, labels, metadata, official splits or task views, schema documentation, and safety documentation; it does not include raw internal construction artifacts or exhaustive per-model output dumps. Because KoEVD includes harmful and sensitive content, the public release documentation should include license or data-use terms, research-use-oriented terms, misuse restrictions, prohibited-use terms, and clear warnings that Task 5 tools are non-executable mock schemas.

PII and workflow-metadata screening. Before public release, benchmark files are screened for direct personal identifiers and workflow metadata. The release review checks for names used as real-person identifiers, phone numbers, email addresses, account IDs, URLs, addresses, internal usernames, timestamps, file paths, and annotation-tool metadata. Items containing real or apparently identifying information are rewritten, removed, or replaced with synthetic placeholders. This screening is separate from the fact that the benchmark intentionally contains offensive or harmful scenarios for safety evaluation.

Non-released internal artifacts. The public release does not include the internal annotation tool, raw collected source materials, AI-assisted candidate-generation prompts or intermediate raw candidates, internal workflow metadata, external provider models, or exhaustive per-model output dumps. These materials are either internal workflow artifacts, governed by third-party terms, or excluded to reduce privacy, misuse, and operational-risk exposure.

AI-assistant use outside benchmark construction. Beyond AI-assisted candidate ideation and drafting during benchmark construction, the authors used AI assistants for auxiliary manuscript, implementation, visualization, release-documentation, and checklist-review support. These uses included English polishing or translation of author-written drafts, LaTeX table drafting and formatting, manuscript condensation, classifier-training code drafting, public-dataset conversion code drafting, dataset README drafting, chart-generation script drafting from CSV inputs, and checklist-based manuscript review. These uses were author-directed and subject to manual verification: the authors edited the final prose, checked tables and figures against the intended data or descriptions, inspected and tested code before use, reviewed README and release-related files, verified checklist-driven changes, and retained responsibility for all claims, citations, analyses, code, figures, documentation, and released artifacts. AI assistants were not used as annotation authorities, independent scientific decision makers, or authors.

N.1 Task Prompt and Output Contracts

N.2 Reproducibility Checklist

Parsing and invalid-output handling. For Task 1, outputs were parsed into safe/unsafe labels using deterministic string and JSON-pattern rules. For Task 2, only valid option-letter or choice-id answers were included in response-choice accuracy, while empty, blocked, or out-of-choice outputs were counted in answer/parse diagnostics. For Task 5, only syntactically parsed JSON decisions with a valid `decision` field and valid mock-tool name were eligible for unsafe-tool-call scoring; malformed outputs were tracked separately.

Table N.1: Public release inventory and use terms.

Released artifact	Role in this work	License / access status	Use-term note
KoEVD dataset	Benchmark data, including Korean user utterances, response candidates, labels, metadata, task views, and official splits	KoEVD Dataset Research Use Terms v1.0 (`KoEVD Dataset Research Use Terms v1.0.md`)	Intended for research on Korean LLM safety evaluation; includes content warnings and prohibited-use restrictions.
Assistant-safety classifier	Measurement model for Task 3 generated-response harmfulness	KoEVD Classifier Research Use Terms v1.0 (`KoEVD Classifier Research Use Terms v1.0.md`)	Released to support reproducible Task 3 measurement; not intended for deployment as a general-purpose safety filter.
Response-strategy classifier	Measurement model for Task 4 response-strategy analysis	KoEVD Classifier Research Use Terms v1.0 (`KoEVD Classifier Research Use Terms v1.0.md`)	Released to support reproducible Task 4 measurement; not intended for deployment as a general-purpose response-policy model.

Table N.2: Task prompt and output contracts. We summarize the model-facing contract rather than printing full prompt text. The public release documents task-level input/output contracts and scoring definitions; internal prompt-generation and candidate-generation materials are not released.

Task	Model-facing input	Required output / parser contract
Task 1	Korean user utterance with a binary safety-judgment instruction.	A parsable safe/unsafe judgment. Non-parsed outputs are tracked separately.
Task 2	Korean user utterance plus response candidates under one of four choice variants.	A single option letter or choice id. Accuracy is computed over valid answered choice items.
Task 3	Korean user utterance as a direct generation prompt.	Free-form assistant response; harmfulness is measured by the assistant-safety classifier.
Task 4	No new model prompt; reuses Task 3 generated responses.	Response-strategy classifier outputs primary and multi-label strategies.
Task 5	Korean source request plus request-specific non-executable mock tools under one prompt variant.	JSON decision with `decision`, `tool_name`, `tool_args`, `safety_label`, `risk_category`, `rationale`, and `safe_alternative`.

Table N.3: Reviewer-facing reproducibility checklist.

Component	Required release information
Released artifacts	KoEVD dataset, official splits/task views included in the dataset package, and the two classifier checkpoints used for Task 3 and Task 4 measurement.
Metrics	Denominator definitions for unsafe rate, answer rate, parse rate, skip-aware rate, strategy inclusion, and tool-call safety.
Task contracts	Task-level input/output contracts and scoring definitions; internal prompt-generation and candidate-generation materials are not released.
Models and compute	Evaluated model display names, exact model identifiers where available, decoding settings, infrastructure, compute budget, and model-specific parsing caveats.
Safety handling	Content warnings, intended research use, prohibited-use terms, non-executable Task 5 mock-tool statement, and sanitized paper examples.

Commercial API model identifiers and provider-side safety systems may change over time. Repository metadata preserved exact model identifiers, decoding settings, and evaluated row counts, but we do not rely on access-date claims for reproducibility.

N.3 Compute Budget and Infrastructure

We report the compute budget separately for open-weight inference, commercial API inference, and the two local classifiers used to measure generated-response harmfulness and response strategies. Commercial API model sizes are not publicly disclosed by the providers, so we report evaluated request/row counts rather than estimating proprietary parameter counts or provider-side GPU cost. Open-weight models were served through a local vLLM endpoint on an internal 8-GPU B200 server; repository logs preserve evaluated row counts but do not preserve recoverable vLLM wall-clock serving intervals, so open-weight inference GPU-hours are not reported. Classifier training and inference budgets are computed from recorded trainer and inference runtimes.

Variance and significance interpretation. KoEVD is designed primarily as a diagnostic benchmark rather than a stochastic leaderboard. The main experiments use single-pass model runs under fixed decoding settings, so the paper does not report repeated-run variance or statistical significance tests for small model-to-model differences. Instead, the main claims rely on large gaps, controlled metadata views, cross-task contrasts, and model-disaggregated

Table N.4: Metric denominators and parse-handling caveats. Rates across tasks are not interchangeable.

Metric	Task	Denominator / interpretation
Unsafe judgment rate	Task 1	Model decisions over evaluated user utterances; measures whether the user utterance is judged unsafe.
Response-choice accuracy	Task 2	Valid answered choice items; answer/parse rate should be read alongside accuracy.
Unsafe generation rate	Task 3	Classifier-parsed generated responses, with skip-aware/all-prompt views used as sensitivity checks.
Strategy inclusion rate	Task 4	Generated responses classified for a given strategy; multi-label and sometimes grouped by reporting family, so values should not be read as mutually exclusive shares.
Unsafe tool-call rate	Task 5	Parsed mock-tool decisions; unsafe means selecting the request-specific unsafe tool/action affordance.

Table N.5: Compact model/run-configuration summary. Display names omit local serving/runtime suffixes.

Model group	Decoding	Models
Open-weight/local	T=0; max=512	Apertus 8B, Gemma 4 E4B, Ministral 3, OLMo 3 7B Think, Qwen 3.5, Qwen 3.6.
Claude/Gemini API	T=0; max=512	Claude 4.6 Sonnet, Claude Haiku 4.5, Gemini 3.1 Flash-Lite, Gemini 3.1 Pro.
GPT API	T=1; max=512	GPT-5.4, GPT-5.4 Nano under the API configuration used in the runs.
Prompt/output contract	task-specific	Task 1 returns safe/unsafe labels; Task 2 returns a choice; Task 3 returns free-form text; Task 5 returns a mock-tool JSON decision.

sensitivity checks. Small cells or low-parse settings should be interpreted through denominators, answer/parse rates, and appendix breakdowns rather than as standalone rankings.

Table N.6: Model identifiers and run configurations. Exact identifiers should be interpreted together with provider-side versioning and safety-system caveats.

Display name	Provider / source	Exact model ID or checkpoint	Decoding	Note
Apertus 8B	Open-weight	`swiss-ai/Apertus-8B-Instruct-2`	T=0; max=512	Local vLLM inference.
Gemma 4 E4B	Open-weight	`google/gemma-4-E4B-it`	T=0; max=512	Local vLLM inference.
Ministral 3	Open-weight	`mistralai/Ministral-3-8B-Reaso`	T=0; max=512	Local vLLM inference.
OLMo 3 7B Think	Open-weight	`allenai/Olmo-3-7B-Think`	T=0; max=4096	Local vLLM inference.
Qwen 3.5	Open-weight	`Qwen/Qwen3.5-9B`	T=0; max=512	Local vLLM inference; thinking disabled in request config.
Qwen 3.6	Open-weight	`Qwen/Qwen3.6-27B-FP8`	T=0; max=512	Local vLLM inference; FP8 checkpoint; thinking disabled in request config.
Claude 4.6 Sonnet	Anthropic API	`claude-4.6-sonnet`	T=0; max=512	Provider-side model and safety systems may change.
Claude Haiku 4.5	Anthropic API	`claude-4-5-haiku-20251001`	T=0; max=512	Provider-side model and safety systems may change.
Gemini 3.1 Flash-Lite	Google API	`gemini-3.1-flash-lite-preview`	T=0; max=512	Provider-side model and safety systems may change.
Gemini 3.1 Pro	Google API	`gemini-3.1-pro-preview`	T=0; max=512	Provider-side model and safety systems may change.
GPT-5.4	OpenAI API	`gpt-5.4`	T=1; max=512	Provider-side model and safety systems may change.
GPT-5.4 Nano	OpenAI API	`gpt-5.4-nano`	T=1; max=512	Provider-side model and safety systems may change.

Table N.7: Software environment summary.

Component	Version / implementation
Python	3.11.14
PyTorch	2.11.0+cu130
Transformers	5.8.1
Datasets	4.8.5
vLLM	0.21.1rc1.dev109+gef54a4d60
PEFT / LoRA	PEFT 0.19.0; LoRA adapters trained with PEFT LoRA
scikit-learn	1.8.0
pandas / numpy	pandas 2.3.3 / numpy 2.3.5
Metric and parser definitions	Task-formatting, parsing, and metric definitions are documented with the release package.

Table N.8: Compute budget and infrastructure. Values marked as unavailable were not recoverable from repository logs; classifier values are computed from recorded wall-clock runtimes.

Component	Model size	Infrastructure	Budget / note
Open-weight/local inference	Approximately 4B–27B open-weight models	8-GPU B200 vLLM server	516,090 primary evaluated rows. GPU-hours are not recoverable from repository logs because vLLM wall-clock start/stop or throughput logs were not preserved.
Commercial API inference	Not publicly disclosed by providers	Provider APIs	414,797 primary API rows/requests. API billing cost and provider-side GPU cost are not reported.
Assistant-safety classifier	Qwen3.5-0.8B + LoRA	8-GPU FSDP training; 4-GPU local inference	23.64 training GPU-hours; 0.55 Task 3 inference GPU-hours for 81,312 generated rows.
Response-strategy classifier	Qwen3.5-0.8B + LoRA	8-GPU FSDP training; 4-GPU local inference	82.06 training GPU-hours; 0.19 Task 4 inference GPU-hours for 75,835 generated responses.
Evaluation orchestration/parsing	N/A	Local CPU orchestration plus provider/local endpoints	JSONL request dispatch and parsing; GPU cost is not applicable except for classifier inference.

Table N.9: Release safety policy map.

Release issue	Treatment
Harmful content	Release with clear harmful-content warning; do not print operationally detailed harmful examples in the main paper.
Rejected / strongly rejected responses	Release under explicit research-use terms and prohibited-use restrictions. The terms prohibit reproducing, operationalizing, training for, or deploying harmful, illegal, privacy-invasive, cyber-abusive, or physically harmful behavior. Portions containing highly actionable abuse, cyber, privacy-invasion, illegal-act, or physical-world-harm content are accompanied by stronger warnings, prohibited-use restrictions, and paper-facing sanitized examples.
Benchmark fidelity	Do not replace the released benchmark records with sanitized paraphrases, because the original Korean utterances, response candidates, and metadata are required for meaningful and reproducible safety evaluation; sanitized/paraphrased examples are used only for paper-facing exposition.
Personal identifiers	The benchmark is constructed from curated safety scenarios rather than personal user submissions; annotators were instructed not to include real personal identifiers, and public files remove direct workflow identifiers and timestamps.
Annotator workflow metadata	Remove direct workflow identifiers and timestamps from public JSONL files; keep only internal audit copies.
Interpretive audit fields	Review for sensitive social, cultural, or personal context before public release.
Task 5 mock tools	Release only as non-executable mock schemas; state that no tool is executable.
Prohibited uses	Disallow using the package to train or optimize harmful, illegal, privacy-invasive, cyber-abusive, or physically harmful behavior.

O Responsible NLP Checklist Evidence Map

Table O.1: Responsible NLP checklist evidence map. This table is included for reviewer navigation and does not replace the submission-system checklist.

Item	Checklist issue	Evidence in this paper
A1	Limitations	Section 8.
A2	Risks	Section 9; Appendix N.
B1	Artifact citation	Sections 2, 4; Appendix N.
B2	License / terms	Appendix N, Table N.1.
B3	Intended use	Section 9; Appendix N.
B4	PII / offensive content	Section 9; Appendix N.
B5	Documentation	Appendices B, C, and N.
B6	Statistics	Appendices C and E.
C1	Model size / compute	Appendix N.3.
C2	Experimental setup	Appendices N.1 and N.
C3	Descriptive statistics / single-run interpretation	Section 8; Appendix N.
C4	Packages / implementation	Appendix N, Table N.7.
D1	Annotator instructions	Appendix B.2.
D2	Recruitment / compensation	Appendix B.2.
D3	Consent and data-use awareness	Appendix B.2.
D4	Ethics review	Appendix B.2.
D5	Annotator background	Appendix B.2.
E1	AI assistants	Section 9; Appendices B and N.