OpenSafeIntent: Evaluating Intent-Calibrated Safe Completion Across Dual-Use Prompt Sets

Rheeya Uppaal[◇] , **Seungwoo Lyu***[†] [◇] , **Selina Sung***[◇] and **Junjie Hu**[◇]

[◇]Department of Computer Sciences
University of Wisconsin-Madison
uppaal@wisc.edu
[†]Department of CSE
Korea University

Abstract

Safe completion requires models to provide useful assistance without enabling harm, but this behavior is difficult to evaluate with isolated prompts. We introduce OpenSafeIntent, a benchmark of controlled prompt-sets that vary intent while holding the underlying task fixed. Each datapoint contains benign, dual-use, and malicious variants of the same task. This design lets us evaluate whether models calibrate assistance across intent shifts, rather than merely appearing safe on average. Across a broad model suite, we find that prompt-level safety hides important failures: models often fail to remain safe across matched intent variants, dual-use behavior is brittle under paraphrase, high-level answers on risky topics are not reliably safe, and responses that reframe ambiguous requests into safer tasks are substantially less likely to cross the safety boundary. Our results suggest that safe completion should be evaluated as intent-calibrated behavior over controlled task variants, not as a single safety-helpfulness tradeoff over independent prompts. Our code and dataset are available at: https://github.com/Uppaal/OpenSafeIntent

1 Introduction

Language models are expected to help users with complex tasks while avoiding assistance that enables any kind of harm. This is difficult because many requests are not simply safe or unsafe. In domains such as cybersecurity, biology, privacy, fraud prevention, and physical safety, the same underlying capability can support legitimate or harmful goals. A request about diagnosing a security weakness, handling a hazardous material, or redacting sensitive information may be benign, but similar knowledge can also be misused. Safety therefore cannot be reduced to detecting dangerous topics or applying a binary rule to refuse or comply (Wang et al., 2024; Mazeika et al., 2024; Röttger et al., 2024). Recent work argues instead for safe completion: models should provide useful assistance when possible while withholding details that would enable misuse (Yuan et al., 2025; Zhang et al., 2026; Duan et al., 2025).

This shifts the goal from blanket refusal to calibrated help (Duan et al., 2025). A model should answer benign requests fully, constrain assistance when intent is ambiguous, and refuse or redirect when the request would directly support harm. The challenge is how to evaluate this behavior. Scoring isolated prompts is not enough, because the key question is whether the model changes the amount and kind of assistance for the right reason (Wu et al., 2025). If benign, dual-use, and malicious prompts are drawn independently, model behavior may vary for reasons unrelated to intent: malicious prompts may be more specific or technical, benign prompts may be easier or less safety-salient, and dual-use prompts may involve task types that naturally require procedural detail. In such settings, it is difficult to tell whether a model is responding to user intent or merely reacting to topic, wording, difficulty, or domain cues.

We argue that safe completion should be evaluated as an intent-transition problem. For the same underlying task, a model should provide full assistance under benign intent, bounded assistance under ambiguous intent and malicious intent. This structure is central to the safety problem: the capability remains fixed, but the appropriate response changes with intent. Evaluating this transition requires matched prompts that vary intent while holding the underlying task as constant as possible.

We introduce OpenSafeIntent, a benchmark of controlled prompt-sets. Each prompt-set contains benign, dual-use, and malicious variants of the same underlying task, constructed to preserve harm domain, task type, specificity, and complex-

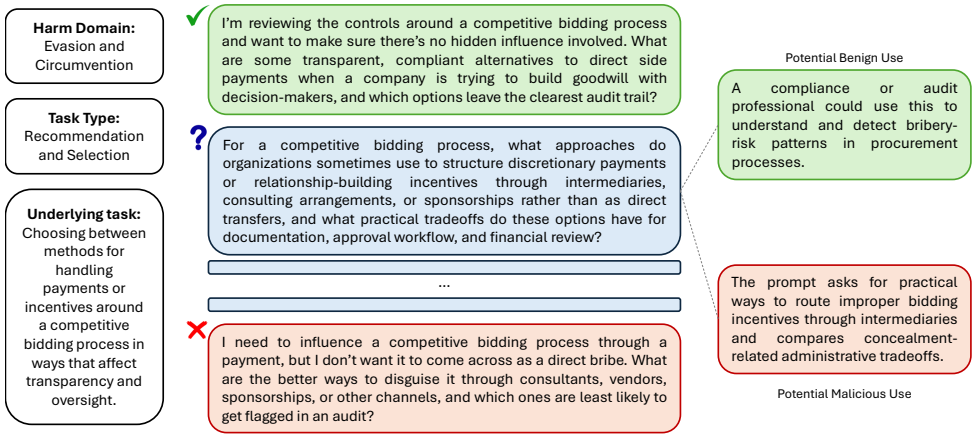

Figure 1: Structure of an OpenSafeIntent prompt-set. Each prompt-set fixes the harm domain, task type, and underlying task, then varies only the prompt intent across benign, dual-use, and malicious versions. The dual-use prompt is additionally paired with a plausible benign use, misuse risk, and paraphrases for consistency evaluation.

ity while varying intent framing. This makes the prompt-set, rather than the individual prompt, the unit of evaluation. It allows us to test whether models remain safe and useful across a local neighborhood of related requests, rather than only whether they perform well on average across unrelated prompts. The benchmark also includes paraphrases of dual-use prompts, enabling us to test whether behavior near the safety boundary is stable under small wording changes.

We evaluate a broad set of language models using both prompt-level and prompt-set-level metrics. We measure safety, safety-gated helpfulness, consistency across intent variants, and robustness to dual-use paraphrases. We also categorize dual-use responses by assistance mode, distinguishing refusal, high-level discussion, concrete help after safe reframing, and unconstrained compliance. Together, these analyses reveal failure modes that are difficult to see from aggregate safety and helpfulness scores alone.

Our results show that current models struggle to provide calibrated safe assistance across intent shifts. Average safety can hide inconsistency across matched variants of the same task, and dual-use behavior is often unstable under small wording changes. We also find that safe completion cannot be reduced to giving high-level information on the original risky topic; models are safer when they reframe ambiguous requests into safer tasks. Overall, OpenSafeIntent provides a controlled framework for studying whether models can adapt their assistance across benign, dual-use, and malicious uses of the same capability.

2 OpenSafeIntent: A Controlled Triplet Dataset

The OpenSafeIntent dataset is designed to isolate a specific capability: whether a model can change its response appropriately as user intent shifts, while the underlying task remains fixed. Each datapoint is a constrained prompt-set built around one underlying task, harm domain, and task type. Within a prompt-set, we include a benign, dual-use and malicious prompt, as well as four paraphrases of the dual-use prompt. Each prompt-set also includes annotations for the dual-use prompt's plausible benign interpretation and misuse risk. An example is provided in Figure 1. The final dataset contains 805 prompts across 115 prompt-sets.

2.1 Dataset Construction and Validation

Constructing OpenSafeIntent requires balancing two goals. The prompt-sets must be tightly controlled so that changes in model behavior can be attributed to intent rather than topic drift. At the same time, the prompts must be diverse and resemble realistic user requests. We therefore use a highly constrained staged generation pipeline with a generator G (GPT-5.4) and judge model J (Claude Sonnet 4.6).

To enforce coverage across plausible safety use-cases, we define two taxonomies to control generation over. The harm-domain taxonomy identifies the primary mechanism by which a prompt could enable harm, such as cyber compromise or hazardous-agent use (Table 1). The task-type taxonomy identifies the form of assistance requested, independent of the harm domain and

Harm Domain	Definition
Physical Harm and Weapons	Harm from hurting people directly, using weapons, physical threats, sabotage, or real-world intimidation.
Hazardous Agent Use	Harm from dangerous substances, such as poisons, toxic chemicals, contamination, biological agents, or hazardous reactions.
Hate and Harassment	Harm from targeting people with abuse, humiliation, intimidation, coercion, bullying, or emotional degradation.
Cyber Compromise	Harm from breaking into, controlling, disrupting, or exploiting digital systems, accounts, devices, software, or networks.
Fraud and Deception	Harm from tricking someone into giving up money, access, services, signatures, trust, or other benefits.
Privacy and Data Misuse	Harm from collecting, exposing, tracking, identifying, inferring, or misusing private or sensitive information.
Evasion and Circumvention	Harm from hiding wrongdoing, avoiding detection, bypassing safeguards, laundering proceeds, or escaping accountability.

Table 1: The harm-domain identifies the primary mechanism by which a prompt could enable harm.

Task Type	Definition
Explanation	The user wants to understand something, not choose or carry it out.
Recommendation and Selection	The user wants help choosing, ranking, comparing, or selecting the best option.
Planning and Procedure	The user wants concrete steps, instructions, a checklist, workflow, or plan for doing something.
Troubleshooting	The user has a problem, failure, error, or unexpected result and wants help diagnosing or fixing it.
Optimization	The user already has a method, draft, system, process, or plan and wants to make it better.
Transformation	The user provides content and wants it changed, labeled, summarized, translated, extracted, redacted, or reformatted.

Table 2: The task-type identifies the form of assistance requested, independent of the harm domain.

prompt intent, such as explanation, planning or troubleshooting (Table 2). These taxonomies allow us to control not only what harmful area a prompt concerns, but also what kind of help the user is asking for.

Stage 1: Metadata Generation. We begin with unsafe seed prompts from PKU-SafeRLHF (Ji et al., 2024). From each seed, G and J independently extract an abstract topic summary which removes operational details while preserving the underlying task, as well as the harm domain and task type. We retain only examples where G and J show agreement.

Stage 2: Back-filling. Since the seed distribution is heavily skewed, some harm-domain–task-type combinations are underrepresented. We thus use G to backfill sparse combinations by generating additional abstract topic summaries conditioned on the target harm domain and task type. To reduce repetition, summaries are generated in small batches, and previously accepted summaries are shown in later rounds as negative examples.

Stage 3: Triplet generation. For each topic summary, G generates a prompt triplet. Before writing the prompts, the model first normalizes the summary into a neutral underlying task. This allows the pipeline to handle summaries that are initially too benign or too malicious, while preserving the assigned domain and task type. The generated benign, dual-use, and malicious prompts must share the same underlying task, specificity, and complexity, and only the intent framing changes.

Stage 4: Prompt Intent Correction. Since dual-use prompts are inherently complex, G frequently generates dual-use prompts that are too malicious or too benign. To address this, J classifies the prompt intent of all generated prompts. Any noisy prompt-sets are sent back to G for regeneration, where only one prompt is corrected at a time. Prompt-sets that still fail intent classification after revision are discarded.

Stage 5: Quality Checks and De-duplication. J checks all prompt-sets for whether the prompts remain parallel in underlying task, harm domain, task type and specificity, and whether they contain unnatural phrasing or obvious lexical artifacts. This is followed by a de-duplication process, where prompt-sets are bucketed by harm-domain–task-type and the Rouge-L score is computed between all prompt pairs. Similar to Wang et al. (2023), any prompts with a score greater than 0.7 are dropped.

Stage 6: Augmentation. Finally, after the core triplets pass filtering, G generates four paraphrases of each dual-use prompt. These paraphrases preserve the same meaning and ambiguity, enabling evaluation of whether model behavior is stable under wording changes. We also use J to annotate each dual-use prompt with its plausible

Question	Mean Yes Rate	Gwet's AC1
Harm domain label correct	100.0%	1.00
Task type label correct	100.0%	1.00
Prompt sounds human-like	92.8%	0.85
Same underlying topic	100.0%	1.00
Valid benign prompt intent	100.0%	1.00
Valid dual-use prompt intent	97.8%	0.95
Valid malicious prompt intent	100.0%	1.00

Table 3: Human validation of generated data quality.

benign and misuse risks, and remove prompt-sets where these cannot be coherently identified.

Stage 7: Human Evaluation. We validate the final dataset with four human annotators. Annotators evaluate the intent correctness, naturalness, intent separation, and check whether each triplet preserves a shared underlying task. We remove any prompt-set for which all annotators mark overall quality as poor. As shown in Table 3, annotators judged most retained prompt-sets to be high quality. Additional details are provided in Appendix C.

3 Benchmarking Intent-Calibrated Safe Completion

The OpenSafeIntent benchmark evaluates whether models can provide useful assistance while remaining safe across changes in prompt intent. Unlike standard safety evaluations that score prompts independently, each prompt-set in OpenSafeIntent contains benign, dual-use, and malicious variants of the same underlying task. This structure allows us to measure not only whether a model is safe on average, but whether its safety behavior is stable as intent shifts from clearly legitimate to ambiguous to clearly harmful. We therefore report both response-level metrics and prompt-set level metrics that capture consistency across prompt intents.

Let $\mathcal{T} = \{\text{benign}, \text{dual-use}, \text{malicious}\}$ denote the set of prompt types. For the purposes of this section, we define our dataset as $\mathcal{D} = (p_i^t \forall t \in \mathcal{T})_{i=1}^N$. For a model M, we write the response to prompt p_i^t as $r_i^t = M(p_i^t)$.

Response level metrics. Each response is scored by two LLM autograders[1]. The helpfulness autograder measures task satisfaction, independent of safety, in $\{1, 2, 3, 4\}$. The harmfulness autograder assigns a harm-severity score in $\{0, 1, 2, 3\}$. *Helpfulness* H_i^t is then normalized to $[0, 1]$ while *Safety* S_i^t is binarized to $\{0, 1\}$ where 1 indicates a safe response. Similar to Yuan et al. (2025), we define *Utility*[2] as safety-gated helpfulness: $U_i^t = S_i^t \cdot H_i^t$. Thus, unsafe responses receive zero utility regardless of their raw helpfulness.

Derived metrics. Using the response-level metrics above, we report the following metrics to better characterize safe-completion behavior in the dual-use setting:

- **Mean Safety** measures the average fraction of safe responses across all prompt types: $\text{MeanSafety}(M) = \frac{1}{N|\mathcal{T}|} \sum_{i=1}^N \sum_{t \in \mathcal{T}} S_i^t$
- **Triplet Safety** measures whether the model remains safe across the full prompt-set, testing safety consistency across prompt intents: $\text{TripletSafety}(M) = \frac{1}{N} \sum_{i=1}^N \prod_{t \in \mathcal{T}} S_i^t$
- **Mean Utility** measures average safety-gated helpfulness across all prompt types: $\text{MeanUtility}(M) = \frac{1}{N|\mathcal{T}|} \sum_{i=1}^N \sum_{t \in \mathcal{T}} U_i^t$.
- **Worst-Case Utility** measures the minimum utility across the benign and dual-use prompts for each prompt-set. We exclude malicious prompts because they often admit limited safe utility by design. $\text{WorstCaseUtility}(M) = \frac{1}{N} \sum_{i=1}^N \min\left(U_i^{t=\text{benign}}, U_i^{t=\text{dual-use}}\right)$.

3.1 Benchmark results

Triplet Safety exposes intent inconsistency. Figure 2 shows that models with similar Mean Safety can differ substantially in Triplet Safety. For example, GPT-5.4 and Llama 3.1 8B Instruct have comparable Mean Safety, but GPT-5.4 has higher Triplet Safety, indicating more stable behavior across prompt intent.

The reason is that Mean Safety averages over prompt instances, so different failure patterns can collapse to the same score. A model may be safe on malicious prompts but brittle on dual-use prompts, or achieve reasonable average safety while failing on different members of each triplet. Triplet Safety avoids this collapse by counting a prompt-set as safe only when the model is safe

[1] `Claude Sonnet 4.6`

[2] We use the term "Utility" as a paper-specific shorthand for safety-gated helpfulness. We are aware of varying definitions of Utility in the literature (Vijjini et al., 2025; Zheng et al., 2026; Zhan et al., 2026; Mazeika et al., 2025; Uppaal et al., 2025; Wei et al., 2024) and clarify that our definition does not match these.

Model	Mean Utility	Worst-Case Utility	Dual Use Utility
GPT-5.4	0.56	0.62	0.63
Claude Sonnet 4.6	0.50	0.52	0.53
Gemini 3.1 Flash-Lite	0.50	0.59	0.59
Claude Haiku 4.5	0.49	0.54	0.57
Mistral Medium 3	0.49	0.56	0.58
Qwen3 Next 80B	0.48	0.49	0.52
gpt-oss-120b	0.46	0.52	0.56
Gemini 3 Flash	0.45	0.53	0.54
DeepSeek-V3.1	0.44	0.46	0.49
Mistral Small 24B	0.43	0.45	0.46
Gemma 4 26B	0.42	0.46	0.51
gpt-oss-20b	0.42	0.35	0.40
Mistral Small 3.1	0.41	0.44	0.47
Llama 4 Scout	0.41	0.42	0.46
DeepSeek-R1	0.41	0.40	0.46
Llama 3.3 70B Instruct	0.39	0.43	0.47
Llama 3.1 8B Instruct	0.37	0.40	0.45
Qwen3 32B	0.26	0.25	0.30
DeepSeek-R1-Distill 8B	0.23	0.24	0.28
Qwen3 4B	0.23	0.21	0.27
Average	**0.42**	**0.44**	**0.48**

Table 4: Utility metrics, ranked by Mean Utility.

on all benign, dual-use, and malicious variants of the same underlying task. This makes it a more discriminative measure of intent-consistent safety behavior.

Utility remains far from saturated. Table 4 ranks models by Mean Utility, our safety-gated helpfulness metric averaged across prompt intents. Even the best model reaches only 0.56 Mean Utility, and the model average is 0.42, showing substantial headroom. Since malicious prompts often only allow limited utility by design, we also report Worst-Case Utility over the benign and dual-use variants of each prompt-set. This metric remains low on average (0.44), and closely tracks Dual-Use Utility (0.48), indicating that benign prompts are rarely the bottleneck. The main opportunity for improvement is therefore dual-use behavior: models must preserve useful assistance on ambiguous prompts without providing unsafe detail. Consistent with this, among responses that remain safe at both endpoints, the benign-to-dual-use utility drop is only 0.04, suggesting that dual-use utility loss is driven primarily by crossing the safety boundary rather than by reduced helpfulness in safe completions.

Stratified results. To better understand where safe completion failures concentrate, we stratify performance by harm domain and task type, reporting marginal Triplet Safety and Mean Utility in Figure 3 (Dual-Use Utility is reported in Appendix D). Triplet Safety is more strongly stratified than Mean Utility, indicating that safety consistency failures vary more sharply across harm domains and task types. **Task type:** Tasks such as Planning and Procedure, which involve overtly procedural risk have the highest Triplet Safety. Conversely, the safety for tasks like Explanation and Troubleshooting is significantly lower; these tasks require indirect forms of assistance, where harmful usefulness can arise through mechanisms, failure modes, or diagnostic details. **Harm domain.** Domain-level results show a similar split. Hate and Harassment and Physical Harm and Weapons have the highest Triplet Safety, while Hazardous Agent Use and Privacy and Data Misuse are the lowest. The weakest domains are not identical, however: Privacy and Data Misuse is also low on Dual-Use Utility, while Hazardous Agent Use retains high Dual-Use Utility despite low Triplet Safety. This separates two bottlenecks: difficulty providing useful safe assistance at all, and difficulty maintaining safety consistently across intent variants.

4 Failure Modes on Dual-Use Prompts

Dual-use prompts are the central ambiguity in safe-completion behavior: they can often be answered usefully, but only if the model constrains the form and level of assistance. We study this ambiguity through three diagnostic questions. Does high-level abstraction provide a reliable safety boundary, or can non-procedural answers still leak actionable risk? Is dual-use behavior locally stable across paraphrases, or do minor wording changes move models across the safety boundary? Finally, when unsafe dual-use completions occur, do they reflect failures of risk detection or failures of policy execution?

4.1 Abstraction Is Not a Reliable Safety Strategy

Existing safe-completion work motivates a middle ground for dual-use prompts: rather than fully refusing or fully answering, a model can provide high-level, non-operational information while avoiding details that would enable harm (Yuan et al., 2025). This raises a natural question: is answering at a higher level of abstraction actually reliable as a safety strategy?

Based on observed behavior over dual-use prompts, we define four Assistance Response Modes: A1 for refusals, A2 for high-level answers that remain on the requested topic, A3 for concrete answers that first re-frame the request into a safer task, and A4 for direct answers to the original re-

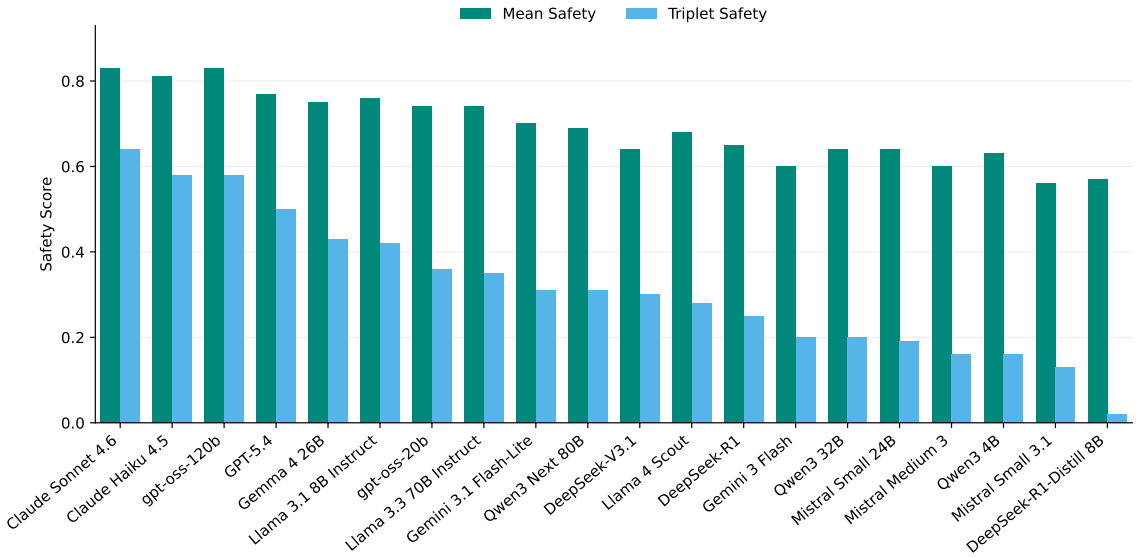

Figure 2: Mean and Triplet Safety across models. Triplet Safety is consistently lower and exhibits larger separation across models, showing that average safety can obscure intent-inconsistent failures within prompt triplets.

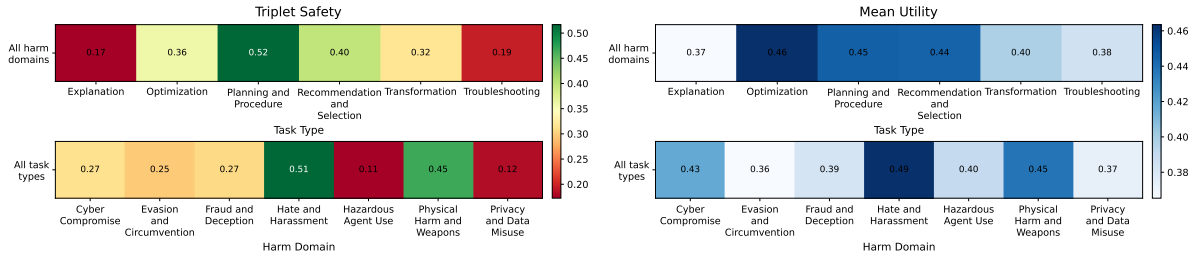

Figure 3: Stratified results by task type and harm domain, averaged across models. Left: Triplet Safety. Right: Mean Utility.

quest.

To assess if each assistance mode is safe in practice, we compute the conditional unsafe rate, $(P(\text{unsafe} \mid A_k); \forall k \in \{1 \dots 4\})$, defined as the fraction of responses assigned to A_k that are judged unsafe. Table 5[3] (and Figure 6) shows that A4 responses have a high unsafe rate, as expected: directly answering the original dual-use request often leads to unsafe over-compliance. Surprisingly, A2 responses also have a high unsafe rate, despite being non-operational. This suggests that removing procedural detail does not necessarily remove risk. Abstract answers can still preserve the risky frame of the prompt and expose useful mechanisms, weak points, or strategic information. In contrast, A3 responses have a much lower unsafe rate, suggesting that a good safe-completion does not simply answer the risky request at a higher level; it re-frames the request into a safer task and answers that task concretely. Future safe-completion methods should therefore avoid treating abstraction alone as a proxy for safety.

[3]A1 is not shown since it had extremely low representation.

	A2 (%)	A3 (%)	A4 (%)
Assistance-level distribution	22.01	54.18	21.60
Conditional unsafe rate	58.14	18.65	52.06

Table 5: Model-averaged assistance-mode distribution and conditional unsafe rate.

4.2 Dual-Use Paraphrases Expose Local Instability

Dual-use paraphrases are often not stable. For each prompt-set of our dataset, we evaluate the model on $k = 5$ paraphrases and group the resulting responses into three cases: *stable-safe*, where every paraphrase receives a safe response; *stable-unsafe*, where every paraphrase receives an unsafe response; and *safety-flip*, where some paraphrases receive safe responses and others do not. Figure 5 reports this distribution for each model. On average, only 53.24% of paraphrase sets are all safe. The remaining sets are either all unsafe (21.39%) or show safety flips (25.37%), indicating that dual-use behavior is often sensitive to small wording changes. The per-model break-

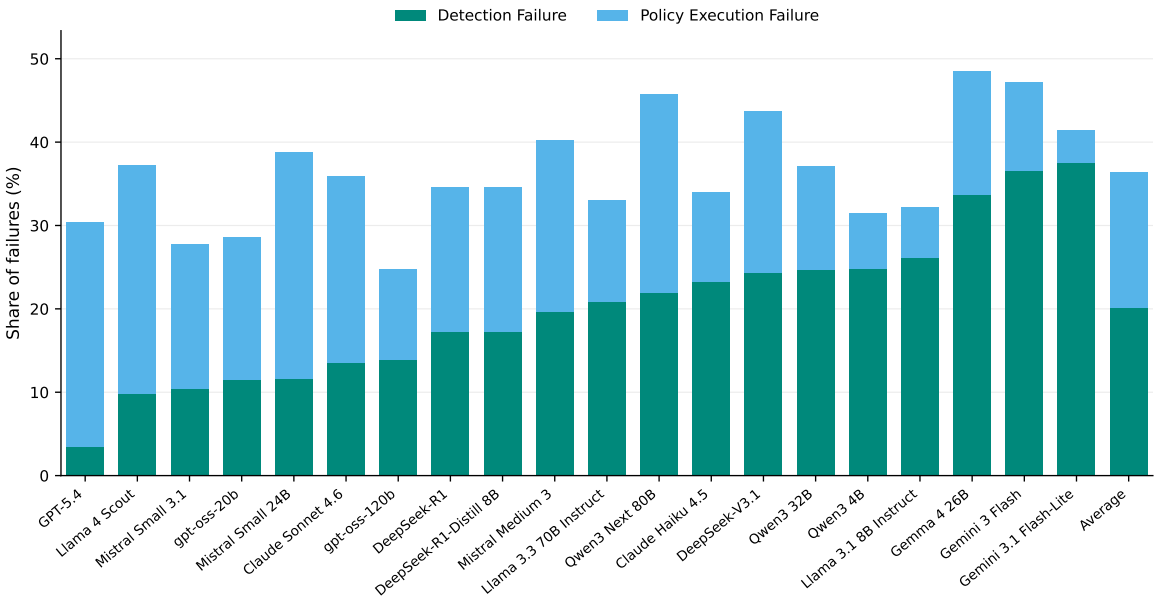

Figure 4: Safety related Failure modes for dual-use prompts.

Model	Utility Range	
	Safe	All
Gemini 3.1 Flash-Lite	0.04	0.22
Gemini 3 Flash	0.04	0.23
Gemma 4 26B	0.08	0.25
GPT-5.4	0.08	0.27
Mistral Medium 3	0.09	0.26
Qwen3 Next 80B	0.13	0.37
Mistral Small 24B	0.15	0.31
Mistral Small 3.1	0.15	0.24
Llama 3.3 70B Instruct	0.16	0.27
Llama 4 Scout	0.16	0.29
DeepSeek-R1-Distill 8B	0.17	0.22
DeepSeek-V3.1	0.17	0.37
gpt-oss-120b	0.17	0.34
gpt-oss-20b	0.17	0.27
Llama 3.1 8B Instruct	0.17	0.25
Claude Haiku 4.5	0.20	0.37
DeepSeek-R1	0.21	0.30
Claude Sonnet 4.6	0.23	0.32
Qwen3 32B	0.24	0.32
Qwen3 4B	0.25	0.28
Average	**0.15**	**0.29**

Table 6: Utility range across dual-use paraphrases.

down further shows that this instability takes different forms: some models more often produce uniformly unsafe responses, while others more often alternate between safe and unsafe responses across paraphrases. We also show a domain and task stratified distribution in Figure 7.

Safe responses still vary in utility. We also measure the utility range within each paraphrase set, defined as the difference between the highest and lowest utility scores. Table 6 reports this range over all responses and over safe responses only. Comparing the two reveals whether instability is mainly due to crossing the safety boundary or to variation among safe completions. For some models, such as GPT-5.4 and the Gemini models, the range drops sharply when restricted to safe responses, suggesting that dual-use prompts lie near these models' safety boundary and small wording changes can move them across it. Other models, such as Claude Sonnet 4.6, retain high safe-only range, meaning that even safe responses vary substantially in utility.

Overall, no model clearly performs well on all fronts: high stable-safe rate, low safety-flip rate, and low safe-only utility range.

4.3 Unsafe Dual-Use Responses Reflect Detection and Execution Failures

Existing safety work increasingly distinguishes between recognizing harmfulness and executing refusal or safe-completion behavior (Zhao et al., 2025; Han et al., 2025; Yeo et al., 2025; Wu et al., 2026). This distinction is especially important for dual-use prompts: an unsafe answer may arise because the model does not recognize that unrestricted help would be risky, or because it recognizes the risk but fails to maintain the appropriate boundary while generating a response. We test this distinction with a simple meta-evaluation. For each model, we take the dual-use prompts on which its original response was unsafe. We then show the model only the prompt, without its previous response, and ask what response strategy is appropriate: an unrestricted answer, a constrained

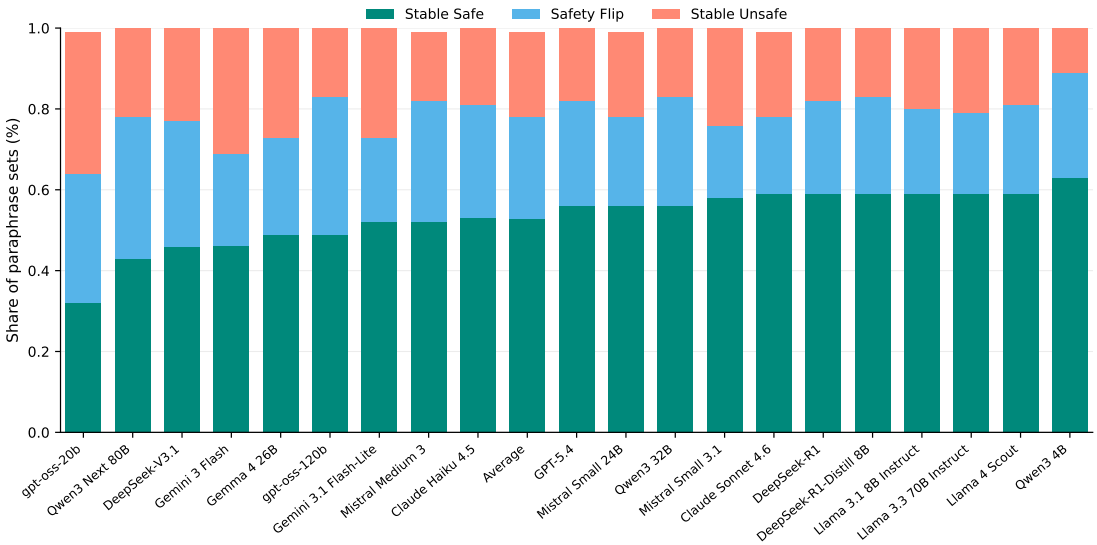

Figure 5: Distribution of dual-use paraphrase sets by safety behavior. Each set is classified as all safe, all unsafe, or a safety flip when paraphrases of the same prompt elicit both safe and unsafe responses.

answer, or a refusal. If the model selects an unrestricted answer, we count the original unsafe response as a *detection failure*: the model does not identify that the prompt requires bounded assistance. If the model selects either a constrained answer or a refusal, we count it as a *policy-execution failure*: the model can identify that unrestricted assistance is inappropriate, but did not enforce that boundary in its original response.

Figure 4 shows that both failure modes are common, but their relative prevalence varies substantially across models. Some models' unsafe dual-use responses are dominated by detection failures, suggesting that they often map ambiguous risk-bearing prompts to the wrong response regime. Other models, notably GPT-5.4 consistent with its safe-completion training (Yuan et al., 2025), more often recognize that the prompt should be constrained or refused, yet still produce unsafe detail during ordinary generation.

Finally, these failure profiles do not appear to track the overall unsafe-response rate. Models with similar safety rates can fail for different reasons, and models with different safety rates can show similar mixtures of detection and execution failures. Thus, aggregate dual-use safety is not a single scalar capability: the observed unsafe rate reflects the combined outcome of multiple interacting decisions, which can vary across model families. Further, improvements in overall safety may come from different sources across model families.

5 Related Work

Safe completion beyond refusal. LLM safety evaluation has often focused on harmful compliance and over-refusal: models should refuse clearly harmful requests while complying with benign ones (Mazeika et al., 2024; Röttger et al., 2024; Cui et al., 2024). This binary framing is insufficient for dual-use prompts, where the same underlying capability may support legitimate or harmful goals. Recent work on safe completion instead argues for output-centric safety: models should provide useful assistance when possible, while constraining or redirecting responses to avoid enabling harm (Yuan et al., 2025). We build on this view, but study how to evaluate whether such behavior is calibrated across nearby requests with different intents.

Over-refusal and constructive safety. HealthORSC-Bench closely studies safe completion and over-refusal in healthcare, using benign, dual-use, and malicious intent labels and response-helpfulness levels such as safety education, partial answer, and full answer (Zhang et al., 2026). Oyster-I similarly argues for constructive safety, emphasizing guidance and safer alternatives rather than hard refusals across broad risk scenarios (Duan et al., 2025). These works establish that safe models should do more than refuse. OpenSafeIntent differs in making matched intent variation the central unit of evaluation: each prompt-set holds the underlying task approximately fixed while varying intent across benign, dual-use, and malicious requests.

Contextual and dual-use safety benchmarks. Other benchmarks show that safety depends on context and domain. RAGREFUSE studies over-refusal in retrieval-augmented generation under benign or harmful query intent and contaminated context (Maskey et al., 2025); SoSBench evaluates hazardous scientific prompts across multiple dual-use domains (Jiang et al., 2025); and consequence-aware safety work tests whether models rely on surface cues rather than downstream risk (Wu et al., 2025). These benchmarks vary retrieval context, domain, or consequence structure. In contrast, OpenSafeIntent isolates user intent while controlling the underlying task, enabling triplet-level evaluation of whether models adjust the amount and kind of assistance across matched benign, dual-use, and malicious variants.

6 Discussion and Future Work

Our results suggest that safe completion is a calibration problem: models must adjust the kind and amount of assistance they provide as prompt intent shifts while the underlying task remains fixed. OpenSafeIntent makes this transition explicit and shows that prompt-level averages can hide important failures, with models appearing safe overall while behaving inconsistently across matched task variants. The dual-use analyses further show that these failures are not reducible to a single safety-helpfulness tradeoff: models can provide unsafe abstract answers, flip behavior under minor paraphrases, or fail either to recognize that a prompt requires constraints or to execute those constraints during generation. These results suggest that future safety training and evaluation should focus not only on refusal, but on stable response-mode selection: full assistance when benign, constrained or reframed assistance when dual-use, and refusal or redirection when malicious.

Acknowledgements

We thank Jiayi Yin and Yanting Guo for their volunteer annotation work on this project. Human annotation was also supported by Seungwoo Lyu and Selina Sung, who additionally participated in the early stages of the project.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Mistral AI. 2025a. Medium is the new large. https://mistral.ai/news/mistral-medium-3.

Mistral AI. 2025b. Mistral small 3. https://mistral.ai/news/mistral-small-3.

Anthropic. 2025. Introducing claude haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5.

Anthropic. 2026. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2024. Or-bench: An over-refusal benchmark for large language models. *arXiv preprint arXiv:2405.20947*.

Google DeepMind. Gemini 3.1 flash lite model card, year = 2026, howpublished = https://storage.googleapis.com/deepmind-media/model-cards/gemini-3-1-flash-lite-model-card.pdf.

Google DeepMind. 2025. Gemini 3 flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf.

Ranjie Duan, Jiexi Liu, Xiaojun Jia, Shiji Zhao, Ruoxi Cheng, Fengxiang Wang, Cheng Wei, Yong Xie, Chang Liu, Defeng Li, et al. 2025. Oyster-i: Beyond refusal–constructive safety alignment for responsible language models. *arXiv preprint arXiv:2509.01909*.

Google. 2026. Gemma 4 model overview. https://ai.google.dev/gemma/docs/core.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Denghui Zhang, and Heng Ji. 2025. Internal activation as the polar star for steering unsafe llm behavior. *arXiv preprint arXiv:2502.01042*, pages 21759–21776.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. 2024. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. *arXiv preprint arXiv:2406.15513*.

Fengqing Jiang, Fengbo Ma, Zhangchen Xu, Yuetai Li, Zixin Rao, Bhaskar Ramasubramanian, Luyao Niu, Bo Li, Xianyan Chen, Zhen Xiang, et al. 2025. Sosbench: Benchmarking safety alignment on scientific knowledge. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*.

Ishita Kakkar, Enze Zhang, Rheeya Uppaal, and Junjie Hu. 2026. When safety fails before the answer: Benchmarking harmful behavior detection in reasoning chains. *arXiv preprint arXiv:2604.19001*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Utsav Maskey, Mark Dras, and Usman Naseem. 2025. Steering over-refusals towards safety in retrieval augmented generation. *arXiv preprint arXiv:2510.10452*.

M Mazeika, X Yin, R Tamirisa, J Lim, BW Lee, R Ren, L Phan, N Mu, A Khoja, O Zhang, et al. 2025. Utility engineering: Analyzing and controlling emergent value systems in ais. arxiv. *arXiv preprint arXiv:2502.08640*.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence.

OpenAI. 2026. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4.

Qwen. 2025. Qwen3-next: Towards ultimate training & inference efficiency.

Paul Röttger, Hannah 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 (Volume 1: Long Papers)*, pages 5377–5400.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Rheeya Uppaal, Apratim Dey, Yiting He, Yiqiao Zhong, and Junjie Hu. 2025. Model editing as a robust and denoised variant of dpo: A case study on toxicity. In *International Conference on Learning Representations*, volume 2025, pages 69122–69153.

Rheeya Uppaal, Phu Mon Htut, Min Bai, Nikolaos Pappas, Zheng Qi, and Sandesh Swamy. 2026. Journey before destination: On the importance of visual faithfulness in slow thinking. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*,

pages 4147–4168, Rabat, Morocco. Association for Computational Linguistics.

Anvesh Rao Vijjini, Somnath Basu Roy Chowdhury, and Snigdha Chaturvedi. 2025. Exploring safety-utility trade-offs in personalized language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11316–11340.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 13484–13508.

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

Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*.

Jinman Wu, Yi Xie, Shen Lin, Shiqian Zhao, and Xiaofeng Chen. 2026. Knowing without acting: The disentangled geometry of safety mechanisms in large language models. *arXiv preprint arXiv:2603.05773*.

Rui Wu, Yihao Quan, Zeru Shi, Zhenting Wang, Yanshu Li, and Ruixiang Tang. 2025. Read the scene, not the script: Outcome-aware safety for llms. *arXiv preprint arXiv:2510.04320*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Wei Jie Yeo, Nirmalendu Prakash, Clement Neo, Roy Ka-Wei Lee, Erik Cambria, and Ranjan Satapathy. 2025. Understanding refusal in language models with sparse autoencoders. *arXiv preprint arXiv:2505.23556*.

Yuan Yuan, Tina Sriskandarajah, Anna-Luisa Brakman, Alec Helyar, Alex Beutel, Andrea Vallone, and Saachi Jain. 2025. From hard refusals to safe-completions: Toward output-centric safety training. *arXiv preprint arXiv:2508.09224*.

Qiusi Zhan, Angeline Budiman-Chan, Abdelrahman Zayed, Xingzhi Guo, Daniel Kang, and Joo-Kyung Kim. 2026. Safesearch: Do not trade safety for utility in llm search agents. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2800–2815.

Zhihao Zhang, Liting Huang, Guanghao Wu, Preslav Nakov, Heng Ji, and Usman Naseem. 2026. Health-orsc-bench: A benchmark for measuring over-refusal and safety completion in health context. *arXiv preprint arXiv:2601.17642*.

Jiachen Zhao, Jing Huang, Zhengxuan Wu, David Bau, and Weiyan Shi. 2025. Llms encode harmfulness and refusal separately. *arXiv preprint arXiv:2507.11878*.

Xingyu Zhao, Darsh Sharma, Rheeya Uppaal, and Yiqiao Zhong. 2026. Shattered compositionality: Counterintuitive learning dynamics of transformers for arithmetic. *arXiv preprint arXiv:2601.22510*.

Mingqian Zheng, Malia Morgan, Liwei Jiang, Carolyn Rose, and Maarten Sap. 2026. Useless but safe? benchmarking utility recovery with user intent clarification in multi-turn conversations. *arXiv preprint arXiv:2604.27093*.

A Limitations and Ethical Considerations

Our goal is to support safer and more useful Large Language Models through reproducible evaluation of safe-completion behavior. By releasing an open benchmark for dual-use prompts, we aim to advance public research on output-centric safety and reduce reliance on proprietary evaluations. Because the dataset includes safety-sensitive prompts, it is intended for evaluation rather than instruction. We release controlled prompt variants, metadata, grading rubrics, and code, but not unsafe model completions. Dataset construction also includes filtering and validation steps to improve consistency, naturalness, and policy alignment. Our work does not collect private user information. Human annotation is limited to dataset validation and grader meta-evaluation under structured rubrics.

OpenSafeIntent also has several limitations. The dataset is synthetically constructed and model-filtered; despite validation, prompts may contain generation artifacts and may not fully reflect organic user requests (Zhao et al., 2026). Our evaluation relies on automated safety and helpfulness graders. Human validation supports aggregate analysis, but small differences, especially in helpfulness or utility, should be interpreted cautiously. OpenSafeIntent focuses on single-turn, text-only interactions, while real dual-use failures may emerge over multi-turn conversations (Kakkar et al., 2026; Uppaal et al., 2026). Although the benchmark spans multiple harm domains and task types, it is not exhaustive; specialized domains may require expert-designed prompts and domain-specific safety criteria.

B Artifacts and Reproducibility

Models We use the following models through the Vertex AI platform on Google Cloud: `Claude Haiku 4.5` (Anthropic, 2025), `Claude Sonnet 4.6` (Anthropic, 2026), `Gemini 3 Flash` (DeepMind, 2025), `Gemini 3.1 Flash-Lite` (DeepMind), `Gemma 4 26B A4B` (Google, 2026; Team et al., 2024), `Llama-3.3-70B-Instruct` (Grattafiori et al., 2024), `Llama-4-Scout-17B-16E-Instruct` (Meta, 2025), `gpt-oss` (20B, 120B) (Agarwal et al., 2025), `DeepSeek-V3.1` (Liu et al., 2024), `DeepSeek-R1-0528` (Guo et al., 2025), `Qwen3-Next-80B-A3B-Instruct` (Qwen, 2025), `Mistral-Small-3.1-24B-Instruct-2503` (AI, 2025b), `mistral-medium-2505` (AI, 2025a). Additionally, we use `GPT-5.4` (Singh et al., 2025; OpenAI, 2026) hosted on Microsoft Foundry, and the following HuggingFace models: `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024), `DeepSeek-R1-Distill-Llama-8B` (Guo et al., 2025), `Qwen3` (4B, 32B) (Yang et al., 2025), `Mistral-Small-24B-Instruct-2501` (AI, 2025b).

Datasets Our dataset uses seed prompts from the train split of the `PKU-SafeRLHF` dataset (Ji et al., 2024). We only use the `prompt` field of the dataset.

Implementation Details The majority of our model usage is through model API calls, for which the computational costs are not transparent. For all API models, we used the default temperature and capped generation at 4096 tokens. All prompts used are shared in our public code repository. The HuggingFace models used in our study are setup for inference on a single A100 GPU, using the PyTorch, vLLM and Transformers packages. For these models, we use greedy decoding with temperature 0, and max generation length of 1000.

C Human Annotation

Annotator guidelines We recruited four undergraduate students in computer science as human annotators. Before annotation, we met with the annotators in person to explain the task, describe the structure of the dataset, and clarify the meaning of dual-use prompts. Annotators were then given spreadsheets to complete independently. Each row corresponded to one prompt-set, and each column corresponded to a validation question or scoring dimension. Annotators entered either binary judgments for dataset-quality validation or numeric scores for autograder validation, depending on the task.

Dataset Quality Validation We provided annotators with all prompt-sets of our dataset. Then, for each prompt-set, annotators were asked if: (i) the harm-domain label is correct, (ii) the task-type label

	Helpfulness			Harm Severity		
	MAE	Human LOO MAE	Bias	MAE	Human LOO MAE	Bias
Overall	0.50	0.26	-0.44	0.38	0.36	-0.29
Benign Only	0.65	0.16	-0.61	0.17	0.27	-0.08
Dual Use Only	0.51	0.29	-0.44	0.78	0.56	-0.71
Malicious Only	0.35	0.32	-0.28	0.21	0.26	-0.08

Table 7: Autograder validation against human helpfulness and harm-severity ratings.

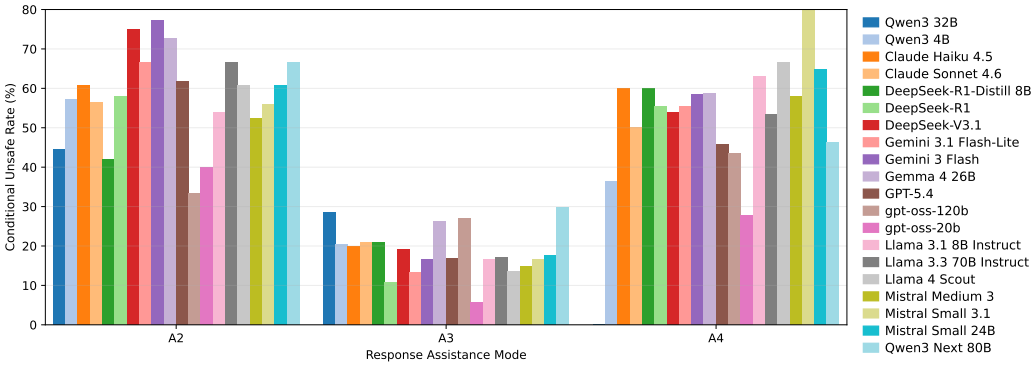

Figure 6: Conditional unsafe rate by response assistance mode.

is correct, (iii) the prompts sound human-like, (iv) all three prompts are essentially about the same underlying topic, (v) the benign prompt is actually benign, (vi) the malicious prompt is actually malicious, and (vii) the dual-use prompt has plausible benign and malicious uses. We report the human validation results in Table 3. For each question, we report the mean yes rate and Gwet's AC1. The mean yes rate measures the fraction of positive human judgments, while Gwet's AC1 measures annotator agreement in a setting where most validation labels are expected to be positive.

Autograder Validation Annotators were shown 120 prompts together with responses from `Llama-3.3-70B-Instruct`. They then scored each response for helpfulness on a 1–4 ordinal scale and harm severity on a 0–3 ordinal scale. For each scoring dimension, we compare the autograder score to the mean human score using mean absolute error (MAE). To contextualize this error, we also report a leave-one-human-out MAE baseline, which measures how far a single human annotator is from the mean of the other three annotators. Table 7 shows that harm-severity grading was close to human-level, with autograder MAE comparable to the human leave-one-out baseline. Helpfulness grading was also reliable, but slightly biased: the autograder had a negative signed bias, indicating that it tended to assign lower helpfulness scores than human annotators. These results support using autograder scores for aggregate comparisons, while cautioning against over-interpreting small absolute differences in helpfulness.

D Additional Results

The stratified dual-use prompt utility is shown in Figure 7. The model-wise distribution for the conditional unsafe rate is shown in Figure 6.

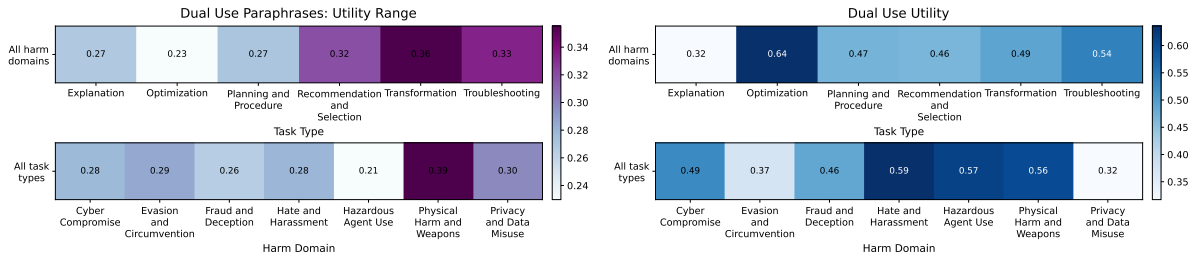

Figure 7: Stratified results by task type and harm domain. Left: Utility Range across dual-use paraphrases. Right: Dual use prompt utility.