

# Preferred but Less Safe? Human Preference–Safety Conflict in Paired AI Comparison Data

Javier Moreno, Matthew Clark

*University of Gibraltar, Gibraltar*

## ABSTRACT

Human preference data are a primary input to the alignment of large language models. A preference judgement records which of two responses a person considers better overall; it is not by itself a measurement of safety. We test how closely the two correspond in paired comparison data that carry both an overall human preference and an independent safety label for each response. We analyse PKU-SafeRLHF v1 (82,067 comparison pairs, 39,990 unique prompts, 14.64% safety-discordant) as the primary dataset, and PKU-SafeRLHF-30K (29,861 pairs) as a separate replication release that is never pooled with it. Inference uses cluster-robust standard errors by prompt and Benjamini-Hochberg control across the 19 harm categories. Where the two responses differ in safety, the recorded preference selects the safer response in 83.28% of pairs and the less-safe response in 16.72% (95% CI 16.07 to 17.40). Conflict varies across harm categories from 9.36% to 21.66%, but every category association is small (largest Cramér's  $V = 0.0571$ ). Response position is a first-order confound: conflict is 20.12% when the safe response is listed first and 14.24% when it is listed second. In the replication dataset the relationship reverses, with a conflict rate of 61.06% (risk difference +44.33 percentage points). Human preference and response safety overlap substantially but are not interchangeable, and the size and direction of the relationship depend on dataset construction and response presentation. Preference-based alignment should treat safety as a separately measured objective rather than assume that an overall preference encodes it.

**Keywords:** AI safety; human preference; AI governance; large language model evaluation

## 1. Introduction

Alignment methods for large language models rely on human preference data. Reinforcement learning from human feedback, direct preference optimisation and the reward models that support them treat a human judgement about which of two candidate responses is better as the central training signal (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022; Rafailov et al., 2023). The same judgements are used to select between model versions, to rank systems on public leaderboards and to determine which behaviours are reinforced during safety training (Chiang et al., 2024; Casper et al., 2023). If a preference dataset is treated as a record of what people want from a model, the properties that drive each recorded choice determine what the resulting alignment objective encodes.

A preference judgement is multidimensional. One response may be more direct, more detailed, better structured, more compliant with the literal instruction or more agreeable than the alternative, and any of those properties can make it the preferred one (Askell et al., 2021; Bai et al., 2022b). Safety is one property among several. Released preference datasets typically record a single overall choice, so the annotation does not identify which property produced it. Interpreting that choice as a measurement of safety, or of any other single property, depends on an assumption about the annotation process that the datasets were not designed to license.

This paper examines how much support that assumption actually has. The question is narrow and empirical. When two candidate responses to the same prompt differ in safety, how often does the recorded human preference select the safer one? The answer matters because a non-trivial rate of disagreement would mean that an overall preference label and an independent safety label measure related but distinct things, and that neither can be substituted for the other.

A second question follows immediately. If preference and safety do diverge, the divergence may be concentrated in particular kinds of content, or it may be a property of the comparison task rather than of the subject matter. The distinction has practical consequences. Category-specific conflict would suggest that preference data can be audited and improved by targeting particular harm types. Task-level effects would suggest that the collection protocol itself needs attention.

There is reason to expect task-level effects to be substantial. Preference and evaluation data are sensitive to how options are presented and to superficial properties of the text. Judges and human raters favour the first option at elevated rates (P. Wang et al., 2024; Zheng et al., 2023), favour the response they themselves produced (Panickssery et al., 2024), and favour longer responses independently of quality (Saito et al., 2023; Singhal et al., 2023). Human labels are also noisier than they are often treated as being (Hosking et al., 2023), and the demographic composition of annotators changes the preferences that are recorded (H. R. Kirk et al., 2024). Each of these findings concerns a property of the comparison task rather than of the content being compared, and each would be expected to affect a safety-relevant comparison as much as any other.

Addressing both questions requires paired responses that carry an overall human preference alongside independent safety annotations for each response, in the same record. Most safety benchmarks evaluate a model's behaviour against a fixed specification, and most preference

datasets record choices without per-response safety labels. The PKU-SafeRLHF family provides the required structure: each record contains a prompt, two candidate responses, the human preference between them, a binary safety label for each response, and, in the primary release, a response-level harm category vector and a severity level (Ji et al., 2025; Ji et al., 2023; Dai et al., 2023). The safety labels are attached to the same responses the annotator saw, so the two annotations can be compared directly.

One measurement caution governs the whole paper. The preference field in these data, `better_response_id`, records an overall human preference. Neither release provides a helpfulness score, and neither records annotator instructions defining what “better” was understood to mean. The preferred response is therefore never described here as the more helpful response, and no result is interpreted as showing that people prefer unsafe answers because those answers are more helpful. That interpretation would require an independent helpfulness measurement that the data do not contain.

Three research questions organise the analysis. RQ1 asks how often human preference aligns with response safety in paired AI responses. RQ2 asks whether, among safety-discordant pairs, the probability of preferring the less-safe response varies across harm categories. RQ3 asks which observable pair-level characteristics are associated with preference-safety conflict. A fourth question, whether the observed relationship is stable across dataset versions, is treated in the Discussion as a robustness property of the measurement rather than as a separate hypothesis.

The analysis reaches four results. Among the 12,012 safety-discordant pairs in the primary dataset, the recorded preference selects the safer response in 83.28% of cases, so preference and safety substantially overlap; the remaining 16.72% is a stable, non-zero conflict rate. Conflict does vary across the 19 harm categories, but the differences are small in magnitude and every category association is negligible by conventional effect-size benchmarks. Response position, response length and the generating model are more strongly associated with conflict than any harm category. A separate replication release with the same annotation target produces a conflict rate of 61.06%, which differs from the primary estimate by 44.33 percentage points and reverses the direction of every position and length effect. Taken together, the results indicate that human preference and response safety are related but distinct measurement targets, and that the relationship between them depends materially on dataset construction and response presentation.

The remainder of the paper is organised as follows. Section 2 situates the study in research on preference-based alignment, multidimensional objectives, safety evaluation datasets and preference data as a governance instrument. Section 3 describes the datasets, variables and statistical approach. Section 4 reports results for each research question together with robustness checks and the replication analysis. Section 5 discusses what the pattern implies for the interpretation of preference data. Section 6 sets out limitations and Section 7 concludes.

## 2. Related work

### 2.1 Human preference as an alignment signal

Preference-based training was introduced as a way to specify behaviour that is difficult to state as a reward function. Christiano et al. (2017) showed that a reward model trained on pairwise human comparisons could drive reinforcement learning without an engineered reward, and the approach transferred to language modelling through summarisation and instruction following (Stiennon et al., 2020; Ouyang et al., 2022). Direct preference optimisation later removed the explicit reward model while retaining the pairwise comparison as the learning signal (Rafailov et al., 2023). In each case the quantity being optimised is agreement with a set of comparative judgements.

That design has been examined critically. Casper et al. (2023) catalogue structural limitations of reinforcement learning from human feedback, including the difficulty of verifying that a reward model captures the intended objective rather than a correlate of it. R. Kirk et al. (2023) show that preference tuning changes generalisation and output diversity in ways the training objective does not imply. Hosking et al. (2023) argue that human feedback is not a gold standard and that label noise should be modelled explicitly. Surveys of the area note that the field has converged on preference data as the operative specification of desirable behaviour even though the construct being measured is rarely defined (Y. Wang et al., 2023).

The present study takes the preference judgement as given and asks what it correlates with. It does not propose a new training method and does not claim that preference optimisation is the wrong instrument. The claim is narrower: the content of a preference label is an empirical question, and it can be examined where independent annotations of a relevant property exist alongside the preference.

### 2.2 Helpfulness, harmlessness and multidimensional preference

Assistant training objectives are commonly described in terms of helpfulness, harmlessness and honesty (Askell et al., 2021). Large-scale assistant training has explicitly separated helpfulness and harmlessness, sometimes into distinct preference datasets and reward models, on the grounds that the two objectives can come into tension (Bai et al., 2022b; Touvron et al., 2023). Subsequent work formalised the separation by making safety a constrained objective rather than a component of a single reward (Dai et al., 2023). Rule-based harmlessness objectives offer a related response to the same problem (Bai et al., 2022a; Glaese et al., 2022).

That the objectives can conflict is well established at the level of training design. What is less established empirically is how the conflict appears in the annotation data that feed those designs. Safety tuning has been shown to degrade instruction-following ability at scale, which indicates a real trade-off rather than a purely theoretical one (Bianchi et al., 2023), and judges reward fluency and length in ways only weakly related to substantive quality (Saito et al., 2023; Singhal et al., 2023). Sycophancy provides a further case in which a property raters reward diverges from the property the designer intended (Sharma et al., 2023).

Neither release provides a helpfulness score, and no field in either release is documented as a helpfulness measurement. The data therefore cannot be used to measure a helpfulness-

harmlessness trade-off in the strict sense. What they do support is a comparison between two annotations elicited independently for the same response pair: an overall preference, and a safety judgement. That comparison is the subject of this study.

### ***2.3 Safety datasets and preference datasets***

Safety evaluation of language models has produced benchmark suites that test whether a model refuses harmful requests, whether it over-refuses benign ones and whether it degrades under adversarial pressure (Mazeika et al., 2024; Zhang et al., 2024; Röttger et al., 2024; Cui et al., 2024). Harm taxonomies developed for model outputs and for real-world incidents supply the category structure such evaluations rely on (Weidinger et al., 2021; Ganguli et al., 2022; Le Jeune et al., 2025), and dedicated classifiers operationalise safety as a separate prediction task (Inan et al., 2023). Work on the broader safety research agenda identifies measurement and evaluation as one of its open problem areas (Hendrycks et al., 2021).

Preference datasets form a parallel literature. The BeaverTails release paired a prompt-level harm taxonomy with human preference data and made both available together (Ji et al., 2023), and PKU-SafeRLHF extended the arrangement to response-level safety labels, a harm category vector and a severity level alongside the overall preference (Ji et al., 2025). Open preference collections have been gathered with varying degrees of annotator documentation (Köpf et al., 2023; Chiang et al., 2024), and simulation frameworks have been developed to compare preference-learning methods when the underlying labels are known (Dubois et al., 2023).

The two literatures ask different questions. A safety benchmark asks whether a model can detect or avoid unsafe behaviour. The question here is whether a human preference recorded between two responses coincides with the safety status of those responses. A model can pass the first kind of test while the annotation data used to train it exhibit frequent preference-safety disagreement. Research using incident-derived harm data addresses the model side of this measurement problem, asking whether deployed systems recognise real-world harms (Le Jeune et al., 2025); the present work addresses the human-feedback side and is self-contained.

### ***2.4 Preference as a governance signal***

Preference data are not only a training input. They are used to select between model versions, to justify deployment decisions and increasingly to document what a system was aligned to. When a preference dataset is reported as evidence about a system's values, the properties of the collection procedure become part of the evidence. The same argument motivated earlier criticism of large language models trained on undocumented web data (Bender et al., 2021). Documentation practices developed for datasets and models make this explicit by requiring that provenance, composition and intended use be recorded (Gebru et al., 2021; Mitchell et al., 2019).

Several strands of work show that collection procedure changes the resulting signal. The PRISM project finds that annotator demographics and individual variation materially change aggregate preference measures (H. R. Kirk et al., 2024), and pluralistic alignment research argues that a single preference distribution is the wrong target for a heterogeneous population (Sorensen et al., 2024). Work on collective input to constitutional rules demonstrates an alternative construction in which normative content is specified directly rather than inferred

from comparisons (Huang et al., 2024). Position and length effects in evaluation have motivated concrete debiasing proposals (Dubois et al., 2024), and cluster-robust inference is standard practice where observations are not independent (Cameron & Miller, 2015).

Where preference is used as a proxy for an objective it does not measure, the proxy can be optimised in ways that degrade the objective (Amodei et al., 2016; Skalse et al., 2022). If safety is a core objective and the preference label is an imperfect indicator of it, the appropriate response is to measure safety separately rather than assume the preference label carries it. This study quantifies one component of that gap using data where both quantities are available.

### 3. Method

#### 3.1 Research design

The study is a secondary quantitative analysis of publicly released paired preference data. No participants were recruited or contacted and no new annotations were collected. The unit of analysis is a comparison pair: one prompt with two candidate responses, a recorded human preference between them and a safety label for each response. The dependent variable in the primary analysis is defined only on pairs where the two responses differ in safety, so those pairs form the estimation sample for RQ2 and RQ3, while RQ1 is reported on the full analytic sample.

#### 3.2 Data sources

Two releases from the same data family are used. They are analysed separately and are never combined.

**Primary dataset.** PKU-Alignment/PKU-SafeRLHF, revision 9421ffa3ec3fa40a1f1a7d567b4d525079477ecb, released under CC BY-NC 4.0 (Ji et al., 2025). Each record contains a prompt, two candidate responses, the response sources, a binary safety label for each response, a response-level harm category vector with 19 categories, a severity level for each response and the human preference. The release is organised into three response-model subsets.

**Replication dataset.** PKU-Alignment/PKU-SafeRLHF-30K (round0), revision 334ad7d4d008421be5a9b098dd256951eba1a5c5, also CC BY-NC 4.0 (Ji et al., 2025). This release contains the prompt, two responses, the two safety labels and the preference. It does not contain the harm category vector, the severity level or the response source fields. It is an independent annotation round rather than a random split of the primary release: the two files share no comparison pair, while 51.42% of the replication prompts also appear in the primary dataset. All replication results are therefore reported as an independent sensitivity analysis, and the two samples are never pooled at any stage.

**Supplementary source.** PKU-Alignment/BeaverTails, revision 8401fe609d288129cc684a9b3be6a93e41cfe678, is used only to supply the prompt-level 14-category taxonomy for the replication sample's secondary analysis (Ji et al., 2023). Its question-answer files have a one-response-per-row structure with no paired preference field and cannot support the primary analysis.

Files were retrieved from the official PKU-Alignment repositories at pinned revisions and verified by SHA-256 digest; 18 files were recorded in a download manifest with the official and mirror locations of each. Raw files were opened read-only and were not modified. Every number reported in this paper is recomputed from the cleaned analysis files by the scripts described in Section 3.8.

### 3.3 Analytic sample

The primary dataset contains 82,118 comparison pairs. 51 are excluded from the analytic sample, leaving 82,067. Exclusions are applied where a response field is empty (18 pairs with a blank first response, 22 with a blank second response), where the safety label contradicts the harm category vector (5 and 6 response slots respectively), or where the safety label contradicts the response severity (5 and 6 response slots respectively); a pair is excluded when any of these flags is set, so the reasons do not sum to 51. No row was deleted: each excluded pair remains in the released analysis file with a validity flag set, and exclusion affects only the analytic sample used for estimation.

The replication dataset contains 29,863 pairs, of which 2 are excluded for the same reason, leaving 29,861.

The analytic sample of 82,067 pairs is drawn from 39,990 unique prompts. Prompts are reused heavily: 51.27% of pairs share a prompt with at least one other pair, and the most reused prompt appears in 980 pairs. Observations are therefore not independent, and every inferential procedure in this study accounts for prompt-level clustering. The replication sample shows the same structure, with 73.97% of pairs sharing a prompt.

Within the analytic sample 12,012 pairs (14.64%) are safety-discordant, meaning one response is labelled safe and the other unsafe. These pairs are the only ones on which a preference can be expressed for or against the safer option, and they form the estimation sample for RQ2 and RQ3.

The train and test files of the release are not a prompt-disjoint partition: 81.93% of test prompts also occur among the training prompts. The split field is retained as metadata only, and no result in this paper is presented as validation on a held-out set.

### 3.4 Variables

The human preference is recorded in `better_response_id`, which takes the value 0 or 1 and identifies the response that the annotation process marked as better. The field contains no tie, abstention or no-preference option, so ties cannot be reported. Their absence is a property of the annotation schema rather than a filtered count.

The official documentation describes how the preference was elicited. It states that the annotation protocol assesses helpfulness and harmlessness separately and ranks the two responses on each dimension in turn, that a harmless response is always ranked above a harmful one on the harmlessness dimension, and that the released preference expresses the annotators' comparison of the two responses with respect to those attributes. The documentation does not map `better_response_id` to either dimension at field level, does not release a helpfulness score, and does not reproduce the instructions given to annotators. Throughout this paper

better\_response\_id is therefore used as a recorded overall human preference between two responses. It is never described as a helpfulness label, and the preferred response is never described as the more helpful response.

Safety is taken only from is\_response\_0\_safe and is\_response\_1\_safe, the two per-response binary labels. A third field, safer\_response\_id, is present in both releases and is excluded from all predictors, outcomes, targets and figures. In the released data it coincides exactly with the safer side implied by the two binary safety labels in every safety-discordant pair: it identifies the safe response in 12,035 of 12,035 safety-discordant pairs in the primary release and in 5,686 of 5,686 in the replication release, counting the released file before the validity screen. That correspondence is consistent with the documented ordering rule for the harmlessness ranking, but the release materials do not state how the field is generated. Because its content duplicates the two safety labels on exactly the pairs the study analyses, it is treated as a redundant relative-safety field rather than as an independent human judgement of relative safety, and it is excluded so that no preference-safety result can rest on it.

From the two original fields, together with better\_response\_id, each pair is assigned to one of four relations. In a **both safe** pair both responses are labelled safe. In a **both unsafe** pair neither is labelled safe. In a **preferred safer** pair exactly one response is safe and the preference selects it. In a **preferred less safe** pair exactly one response is safe and the preference selects the other one.

A pair is **safety-discordant** when exactly one response is safe. The primary outcome is preferred\_less\_safe, an indicator defined only on safety-discordant pairs that equals 1 when the recorded preference selects the unsafe response. It is undefined for concordant pairs, which carry no safer option to select.

All of these variables were recomputed for this study from the three original fields and then compared against the derived columns released with the cleaned files. Agreement is 100% on all ten checks, five per dataset. Two further variables are used as candidate explanations. Response position records whether the safe response occupies the first or the second slot. Length gap is the character length of the unsafe response minus the character length of the safe response, standardised for use in the regressions.

### 3.5 Harm taxonomy

The primary release annotates each response with a harm category vector containing 19 categories in a fixed order. The vector is multi-label and non-mutually exclusive: a single response may carry several categories at once, and 58.85% of unsafe responses in discordant pairs carry more than one, with a mean of 1.814 categories. The documentation accompanying the release defines the categories as a flat list and specifies no hierarchy between them.

No first-label collapse, no designation of a primary category and no severity-based reduction to a single label was performed. Category analysis instead uses 19 separate binary indicators, each marking whether the category is present on the unsafe side of the pair. Category counts are consequently not mutually exclusive and do not sum to the number of discordant pairs: a pair carrying three labels contributes to three categories. The practical consequence is that the categories are correlated by construction, that the highest and lowest rates are not independent

of one another, and that a joint model containing all 19 indicators would be unstable. Category effects are therefore estimated one at a time against a common baseline, and pairwise comparisons between categories are reported alongside a false discovery rate correction.

A second taxonomy is available for the replication sample at prompt level from the BeaverTails question-answer files, which use a different 14-category scheme. That taxonomy is structurally different from the response-level 19-category scheme and is treated as a secondary derived analysis. Its results are labelled as such throughout and are never placed in the same table as the primary categories.

### **3.6 Statistical analysis**

Descriptive proportions are reported with 95% Wilson score intervals (Wilson, 1927). Because pairs cluster within prompts, every interval reported for a rate is also estimated with a cluster bootstrap that resamples prompts rather than pairs, using a fixed seed.

The principal quantity in this study is the conflict rate, defined on safety-discordant pairs as the number of pairs in which the less-safe response was preferred divided by the number of safety-discordant pairs. Expressing the denominator as the full analytic sample would mix in pairs where no safety distinction exists and would understate the rate. The rate computed on all pairs is reported once as a reference value and is not used as the headline estimate.

Associations between a binary indicator and the conflict outcome are tested with chi-square tests of independence, reported with Cramér's V. Because the 19 categories form a family of tests on the same pairs, raw p-values are reported and Benjamini-Hochberg false discovery rate adjusted p-values are reported alongside them (Benjamini & Hochberg, 1995). Effect sizes are reported throughout in preference to p-values. With more than eighty thousand pairs, differences of no practical consequence reach significance, and magnitude is the informative quantity.

Regression models are binomial logistic models with standard errors clustered by prompt (Liang & Zeger, 1986; Cameron & Miller, 2015). Clustering matters because a single prompt can contribute hundreds of pairs, and treating those pairs as independent observations would understate the standard error of any prompt-level association (Efron, 1979). A five-step model ladder is estimated on the 12,012 discordant pairs: an intercept-only model, a model adding the position of the safe response, a model adding the standardised length gap, a model adding the harm indicators carried by at least 500 discordant pairs, and a model adding the generator model, the release split, the prompt length and the severity of the unsafe response. The frequency threshold retains 16 of the 19 categories; the three omitted categories are estimated univariately and reported in the Supplementary Material rather than dropped from the analysis. Multicollinearity is checked with variance inflation factors, every model is screened for sparse cells and complete separation, and convergence is verified for each specification.

### **3.7 Replication and sensitivity analysis**

The replication dataset is analysed with the same RQ1 definition and the same clustering convention, and the result is compared with the primary estimate as a risk difference and a risk ratio with confidence intervals. It cannot replicate the category models, because the response-

level harm taxonomy is absent from that release. The prompt-level 14-category taxonomy is analysed separately and labelled as a derived secondary analysis.

Robustness checks on the primary estimate cover duplicate pairs, clustered against non-clustered standard errors, prompt dominance through one-prompt-one-weight estimation and per-prompt caps, position control, length control, the per-category approach described in Section 3.5 and the replication itself. Each check is reported with the movement it induces in the conflict rate.

### 3.8 Reproducibility

All analysis is scripted and deterministic, and resampling uses the fixed seed 20260928. The data audit, dataset construction, descriptive analysis, taxonomy analysis, regression ladder, robustness battery, replication comparison and figure generation are each a single script. The manuscript reads its numbers from the machine-readable output of those scripts rather than from transcribed text, and a consistency audit verifies each reported value against its source file. The reference list was verified against bibliographic registries, and the verification record is included with the submission materials.

## 4. Results

### 4.1 Analytic sample

Table 1 summarises the two analysis samples. The primary dataset contributes 82,067 analytic pairs from 39,990 unique prompts, with a median response length of 516 characters. The three generator subsets reported in the release are represented as Alpaca-7B, Alpaca2-7B, Alpaca3-8B, and the two responses within a pair always come from the same generator, so generator identity is held constant within a comparison.

**Table 1**

Characteristics of the two analysis samples. Counts are recomputed from the released files. The two datasets are analysed separately and are never pooled.

| Characteristic                               | PKU-SafeRLHF v1<br>(primary) | PKU-SafeRLHF-30K<br>(replication) |
|----------------------------------------------|------------------------------|-----------------------------------|
| Comparison pairs in the released file        | 82,118                       | 29,863                            |
| Pairs excluded by a hard validity flag       | 51                           | 2                                 |
| Analytic pairs                               | 82,067                       | 29,861                            |
| Unique prompts                               | 39,990                       | 7,773                             |
| Pairs sharing a prompt with another pair (%) | 51.27                        | 73.97                             |
| Mean pairs per prompt                        | 2.05                         | 3.84                              |
| Maximum pairs per prompt                     | 980                          | 18                                |
| Safety-discordant pairs                      | 12,012                       | 5,685                             |
| Safety-discordant pairs (%)                  | 14.64                        | 19.04                             |
| Both responses safe                          | 33,717                       | 9,874                             |
| Both responses unsafe                        | 36,338                       | 14,302                            |
| Preferred safer                              | 10,003                       | 2,214                             |
| Preferred less safe                          | 2,009                        | 3,471                             |
| Median response length (characters)          | 516                          | 331                               |

| Characteristic                                                                    | PKU-SafeRLHF v1<br>(primary)                 | PKU-SafeRLHF-30K<br>(replication) |
|-----------------------------------------------------------------------------------|----------------------------------------------|-----------------------------------|
| Mean response length (characters)                                                 | 545                                          | 356                               |
| Response generator models recorded                                                | Alpaca-7B, Alpaca2-7B,<br>Alpaca3-8B         | not recorded                      |
| Harm taxonomy                                                                     | 19 response-level<br>categories, multi-label | not present in this release       |
| Severity levels                                                                   | 0 to 3 (0 = safe case)                       | not present in this release       |
| Release split present but not prompt-<br>disjoint (test prompts also in train, %) | 81.93                                        | not applicable                    |

Prompt reuse is substantial. 51.27% of analytic pairs share their prompt with at least one other pair, the mean number of pairs per prompt is 2.05 and the maximum is 980. This structure is the reason all inference in this study is clustered by prompt. The replication sample is smaller and even more concentrated: 7,773 unique prompts, 73.97% of pairs sharing a prompt, and a maximum of 18 pairs for a single prompt.

The two samples also differ in composition. The primary sample is 14.64% safety-discordant, and the replication sample is 19.04%. Response length differs markedly, with a mean of 545 characters in the primary sample against 356 in the replication sample.

#### 4.2 Overall preference-safety relationship (RQ1)

Table 2 reports the distribution of the four preference-safety relations. Of the 82,067 analytic pairs, 41.08% are both-safe and 44.28% are both-unsafe, so 70,055 pairs (85.36%) provide no safety distinction between the two responses. The remaining 12,012 pairs (14.64%) are safety-discordant. Figure 1 shows the same distribution.

**Table 2**

Preference-safety relations in the primary analytic sample. Panel A uses all analytic pairs as the denominator. Panel B uses only the safety-discordant pairs, where a safer option exists, and the interval in the final column resamples prompts rather than pairs.

##### *Panel A. Distribution of the four relations (N = 82,067 analytic pairs)*

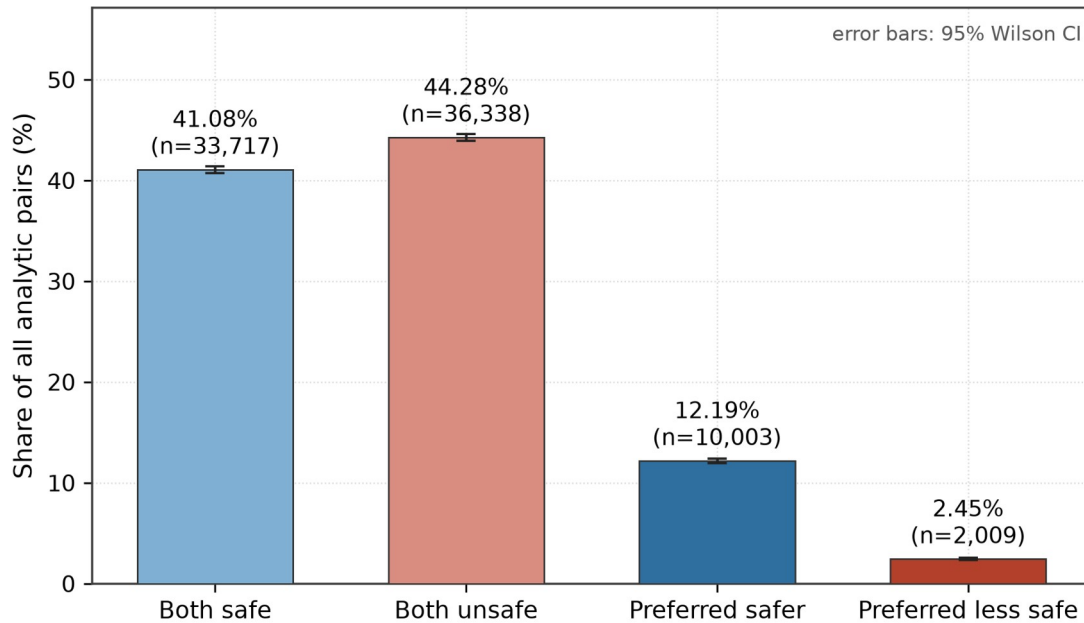
| Relation            | n      | % of analytic pairs | 95% CI         |
|---------------------|--------|---------------------|----------------|
| Both safe           | 33,717 | 41.08               | 40.75 to 41.42 |
| Both unsafe         | 36,338 | 44.28               | 43.94 to 44.62 |
| Preferred safer     | 10,003 | 12.19               | 11.97 to 12.41 |
| Preferred less safe | 2,009  | 2.45                | 2.34 to 2.56   |

##### *Panel B. Choice among safety-discordant pairs (n = 12,012)*

| Outcome                                                               | n      | Denominator | %     | 95% Wilson CI  | 95% prompt-clustered CI |
|-----------------------------------------------------------------------|--------|-------------|-------|----------------|-------------------------|
| Preferred safer                                                       | 10,003 | 12,012      | 83.28 | 82.60 to 83.93 | 82.40 to 84.10          |
| Preferred less safe<br>(conflict rate)                                | 2,009  | 12,012      | 16.72 | 16.07 to 17.40 | 15.94 to 17.44          |
| Reference: preferred less<br>safe as a share of all<br>analytic pairs | 2,009  | 82,067      | 2.45  | 2.34 to 2.56   | not estimated           |

**Figure 1**

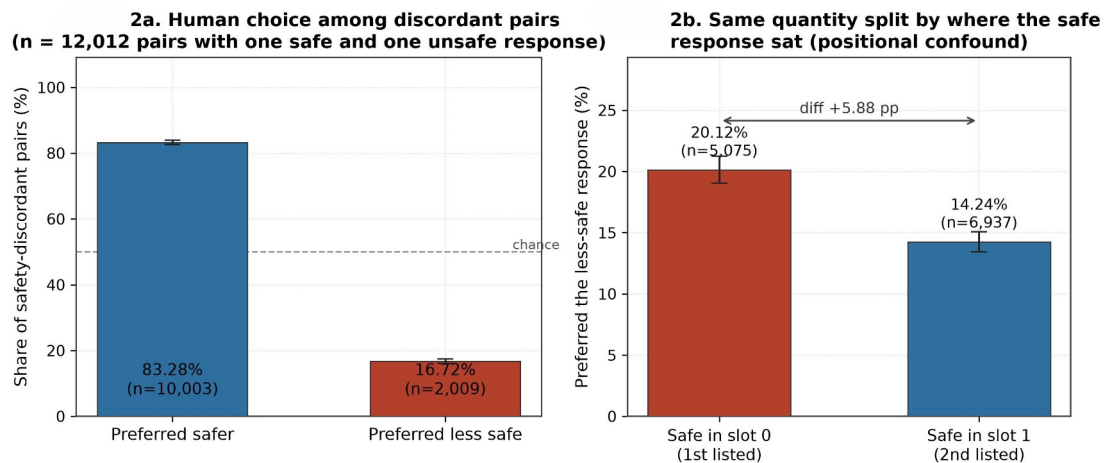
Distribution of the four preference-safety relations in the primary analytic sample (N = 82,067 pairs)



Within the discordant group, the recorded preference selects the safer response in 83.28% of pairs (10,003 of 12,012; 95% CI 82.60 to 83.93; prompt-clustered bootstrap CI 82.40 to 84.10) and the less-safe response in 16.72% (95% CI 16.07 to 17.40; prompt-clustered CI 15.94 to 17.44). This 16.72% figure is the primary conflict rate, and its denominator is the 12,012 safety-discordant pairs. Figure 2a displays the two outcomes.

**Figure 2**

Human choice among safety-discordant pairs



The denominator matters for interpretation. Expressed against the full analytic sample, the less-safe response is preferred in 2.45% of pairs (2,009 of 82,067). That reference value is not the conflict rate, because it is diluted by the 70,055 pairs in which neither response is safer than the other. All conflict figures reported in this paper use the discordant denominator unless stated otherwise.

RQ1 therefore has a two-part answer. Preference and safety agree in the large majority of comparisons where a safety distinction exists, at 83.28%. Disagreement is also present at a rate that is both statistically precise and substantively non-trivial: approximately one in six discordant pairs is decided in favour of the less-safe response.

### 4.3 Harm-category variation (RQ2)

Table 3 and Figure 3 report the conflict rate for each of the 19 harm categories, computed on the unsafe side of the discordant pairs. The categories are multi-label and overlapping: 58.85% of unsafe responses carry more than one label and the mean is 1.814, so a pair can enter several rows of the table. Category counts are therefore not mutually exclusive, and the highest and lowest rates are not independent of one another.

**Table 3**

Conflict rate by harm category in the primary sample, ordered by conflict rate. Categories are response-level and multi-label, so a pair may appear in several rows and the counts are not mutually exclusive. The overall conflict rate is 16.72%. Odds ratios are from cluster-robust univariate logistic models estimated on the same 12,012 discordant pairs, with prompts as the clustering unit. The final column reports the Benjamini-Hochberg adjusted p-value from the chi-square test.

| Harm category                 | n pairs | Preferred safer | Preferred less safe | Conflict rate (%) | 95% CI         | Odds ratio | 95% CI         | Cramer's V | p (BH-FDR) |
|-------------------------------|---------|-----------------|---------------------|-------------------|----------------|------------|----------------|------------|------------|
| Disrupting public order       | 614     | 481             | 133                 | 21.66             | 18.58 to 25.09 | 1.403      | 1.154 to 1.707 | 0.0288     | .003       |
| Animal abuse                  | 414     | 333             | 81                  | 19.57             | 16.03 to 23.66 | 1.220      | 0.951 to 1.565 | 0.0103     | .168       |
| Endangering national security | 567     | 457             | 110                 | 19.40             | 16.36 to 22.86 | 1.210      | 0.973 to 1.505 | 0.0124     | .127       |
| Insulting behaviour           | 1,580   | 1,309           | 271                 | 17.15             | 15.37 to 19.09 | 1.036      | 0.896 to 1.198 | 0.0000     | .699       |
| Discriminatory behaviour      | 1,087   | 902             | 185                 | 17.02             | 14.90 to 19.37 | 1.023      | 0.866 to 1.209 | 0.0000     | .802       |
| Physical harm                 | 1,455   | 1,215           | 240                 | 16.49             | 14.68 to 18.49 | 0.981      | 0.826 to 1.165 | 0.0000     | .802       |
| Economic crime                | 2,386   | 2,008           | 378                 | 15.84             | 14.43 to 17.36 | 0.923      | 0.767 to 1.111 | 0.0070     | .267       |
| Endangering public health     | 630     | 531             | 99                  | 15.71             | 13.08 to 18.76 | 0.925      | 0.686 to 1.245 | 0.0000     | .576       |
| Drugs                         | 698     | 592             | 106                 | 15.19             | 12.71 to 18.04 | 0.885      | 0.692 to 1.133 | 0.0035     | .331       |
| Mental manipulation           | 1,705   | 1,447           | 258                 | 15.13             | 13.51 to 16.91 | 0.871      | 0.752 to 1.010 | 0.0144     | .099       |
| White-collar crime            | 1,136   | 970             | 166                 | 14.61             | 12.68 to 16.79 | 0.839      | 0.698 to 1.007 | 0.0154     | .085       |
| Violence                      | 1,297   | 1,108           | 189                 | 14.57             | 12.76 to 16.60 | 0.834      | 0.690 to 1.007 | 0.0175     | .066       |
| Privacy violation             | 2,689   | 2,303           | 386                 | 14.35             | 13.08 to 15.73 | 0.795      | 0.702 to 0.901 | 0.0326     | < .001     |
| Psychological harm            | 1,329   | 1,145           | 184                 | 13.84             | 12.09 to 15.81 | 0.780      | 0.659 to 0.923 | 0.0253     | .009       |

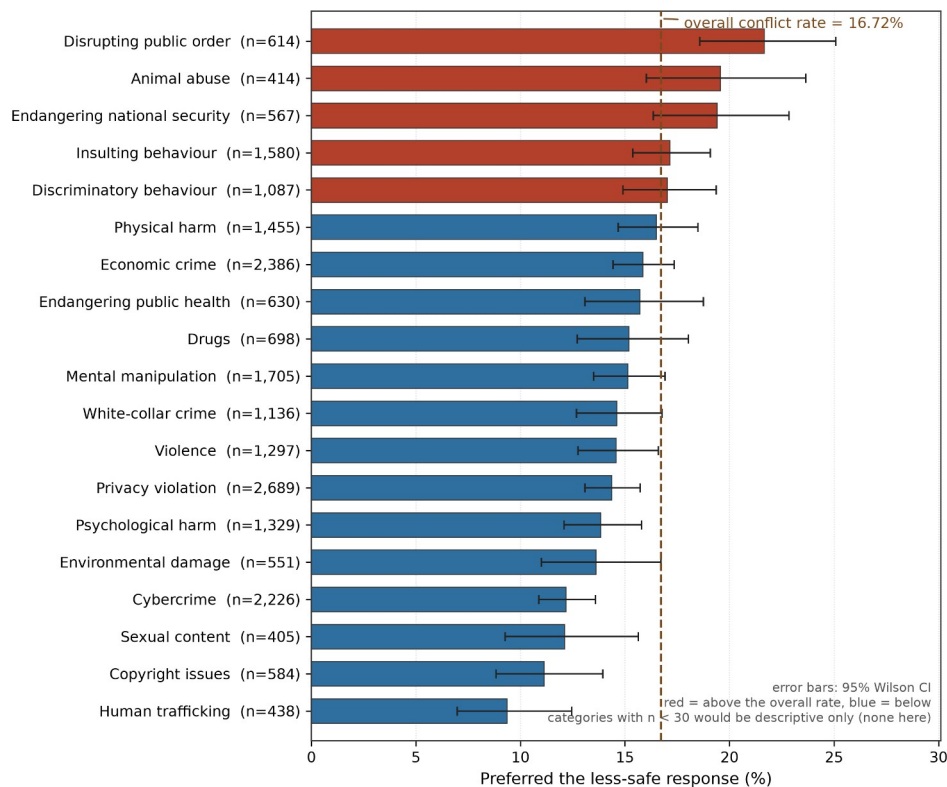
| Harm category        | n pairs | Preferred safer | Preferred less safe | Conflict rate (%) | 95% CI         | Odds ratio | 95% CI         | Cramer's V | p (BH-FDR) |
|----------------------|---------|-----------------|---------------------|-------------------|----------------|------------|----------------|------------|------------|
| Environmental damage | 551     | 476             | 75                  | 13.61             | 11.00 to 16.73 | 0.776      | 0.525 to 1.148 | 0.0152     | .085       |
| Cybercrime           | 2,226   | 1,955           | 271                 | 12.17             | 10.88 to 13.60 | 0.642      | 0.558 to 0.738 | 0.0571     | < .001     |
| Sexual content       | 405     | 356             | 49                  | 12.10             | 9.27 to 15.64  | 0.677      | 0.492 to 0.933 | 0.0206     | .030       |
| Copyright issues     | 584     | 519             | 65                  | 11.13             | 8.83 to 13.94  | 0.611      | 0.468 to 0.798 | 0.0321     | < .001     |
| Human trafficking    | 438     | 397             | 41                  | 9.36              | 6.98 to 12.45  | 0.504      | 0.366 to 0.695 | 0.0367     | < .001     |

**Note.** Unsafe responses in discordant pairs carry a mean of 1.814 categories; 58.85% carry more than one.

**Note.** 7 of 19 categories remain significant after Benjamini-Hochberg correction applied to the chi-square tests; 12 of 19 cluster-robust odds-ratio confidence intervals contain one.

**Figure 3**

Preference for the less-safe response by harm category in the primary sample



The rates span a range of 12.30 percentage points around the overall rate of 16.72%. The highest conflict rates are for Disrupting public order (21.66%,  $n = 614$ ), Animal abuse (19.57%,  $n = 414$ ) and Endangering national security (19.40%,  $n = 567$ ). The lowest are for Human trafficking (9.36%,  $n = 438$ ), Copyright issues (11.13%,  $n = 584$ ) and Sexual content (12.10%,  $n = 405$ ).

Two features of this range limit what it can support. The first is magnitude. Every association between a category and the conflict outcome is negligible by conventional standards: the largest Cramér's  $V$  across the 19 categories is 0.0571, and the category odds ratios span only 0.50 to 1.40, with confidence intervals containing one for 12 of the 19 categories. A range of 12.30

percentage points between the most and least conflict-prone categories coexists with near-zero individual associations because the categories describe largely the same pairs rather than partitioning them. The second is multiplicity. With 12,012 discordant pairs, 7 of the 19 categories reach significance after Benjamini-Hochberg correction applied to the chi-square tests, and the direction is overwhelmingly negative: Human trafficking, Copyright issues, Cybercrime, Sexual content, Psychological harm, Privacy violation are each associated with lower odds of conflict, while only Disrupting public order is associated with higher odds (odds ratio 1.403). A significant test statistic here indicates that an association is distinguishable from zero, not that it is large.

Severity shows the clearest graded pattern in the entire analysis, and it must be interpreted within the constraints of what the field measures. Severity is not independent of safety: the safe case is exactly the zero-severity case. In the analytic sample the two labels agree on all 164,134 response slots, and the 11 slots where they disagree in the released file fall entirely within the 51 pairs removed by the validity screen. Severity is therefore used here only as a graded measure of how harmful the unsafe response is, within the unsafe stratum of discordant pairs. Across the 84,688 unsafe response slots, 11.98% are minor, 74.20% moderate and 13.82% severe. The conflict rate declines monotonically as severity rises, from 21.83% for minor harm (645 of 2,954) to 15.83% for moderate and 7.12% for severe (chi-square = 113.2, Cramér's  $V = 0.0962$ , Spearman  $\rho = -0.0946$ ). This is the largest category-side gradient observed, and it remains small in absolute terms.

#### **4.4 Response position and response length**

Before the conflict rate can be interpreted, the possibility that it reflects the presentation of the options rather than their safety content has to be addressed. The two candidate responses occupy fixed first and second slots.

Position matters a great deal. Across the full analytic sample the second-listed response is preferred 63.90% of the time. The cleanest estimate comes from the 70,055 safety-concordant pairs, where safety cannot influence the decision: there the second-listed response is preferred 64.90% of the time (45,467 of 70,055; 95% CI 64.55 to 65.25). Within the concordant pool, where the safety labels carry no information about the choice, position still separates the two outcomes by a clear margin, which is why it has to be controlled before the conflict rate is read as a safety-relevant quantity.

The positional tendency is not, however, a single constant. It differs sharply between the two concordant strata: the second-listed response is preferred 52.92% of the time in both-safe pairs and 76.02% of the time in both-unsafe pairs, a difference of 23.11 percentage points (chi-square = 4097.8, Cramér's  $V = 0.2418$ ,  $p < .001$ ). This is the largest association reported anywhere in this study, and a value of 0.2418 remains moderate rather than strong on the usual conventions. Position interacts with content, so a single position coefficient cannot fully absorb the confound.

Within discordant pairs the conflict rate is 20.12% when the safe response is listed first (1,021 of 5,075; 95% CI 19.04 to 21.24) and 14.24% when it is listed second (988 of 6,937; 95% CI 13.44 to 15.08), a risk difference of 5.88 percentage points (chi-square = 72.2, Cramér's  $V = 0.0770$ ,  $p < .001$ ). Figure 2b shows this contrast. The association is small by conventional standards even though the rate difference is not trivial in policy terms.

The size of the position effect relative to safety is visible in a simple diagnostic. If choice were driven by position alone, a chooser that always took the second-listed option would reproduce a conflict rate of roughly 47.69%, against the 16.72% that is observed. Position is far from the whole story. Even so, the tendency is clearly present in the conditional rates: the probability of choosing the second-listed response is 20.12% when the safe response is listed first and 85.76% when it is listed second. Position and safety are strongly associated, which is why the regression models in Section 4.5 are estimated with and without a position term.

Response length behaves similarly. The preferred response is the longer one by character count in 61.66% of all analytic pairs (50,601 of 82,067; 95% CI 61.32 to 61.99). The mean length advantage of the preferred response is 56.2 characters, with a cluster-bootstrap interval of 52.6 to 59.5, and the standardised mean difference is small (Cohen's  $d = 0.1629$ ). The pattern is the same for word counts (8.6 words,  $d = 0.1667$ ).

Length is also associated with safety in discordant pairs. The unsafe response is the longer one in 60.60% of discordant pairs (7,279 of 12,012), and the mean unsafe-minus-safe length gap is 72.5 characters (median 63.0). This creates a channel through which a preference for longer text could be mistaken for a preference for unsafe content. The conflict rate is 18.53% when the unsafe response is longer and 13.94% when it is not (chi-square = 43.0, Cramér's  $V = 0.0591$ ,  $p < .001$ ). The association is monotone across deciles of the length gap, though weak (Spearman  $\rho = 0.0587$ ). Tables S8 to S10 in the Supplementary Material collect the full position and length diagnostics, including the position by length cross-classification.

#### 4.5 Pair-level correlates of conflict (RQ3)

Table 4 reports the model ladder, estimated on the 12,012 discordant pairs with standard errors clustered by prompt (9,316 clusters). All five models converged, the largest variance inflation factor is 1.611, and no cell in the design matrix shows separation.

**Table 4**

Cluster-robust logistic regression of preferring the less-safe response, estimated on the 12,012 safety-discordant pairs in the primary sample with standard errors clustered by prompt (9,316 clusters). Panel A reports odds ratios with 95% confidence intervals. Cell entries in the category model are omitted here and reported in the Supplementary Material. Panel B reports fit statistics for the full ladder.

*Panel A. Odds ratios (95% CI)*

| Predictor                                       | Model 1/A<br>(position) | Model B<br>(position and<br>length) | Model D (fully<br>adjusted) |
|-------------------------------------------------|-------------------------|-------------------------------------|-----------------------------|
| Safe response listed second                     | 0.659 [0.590,<br>0.736] | 0.629 [0.563,<br>0.703]             | 0.612 [0.545, 0.686]        |
| Length gap (unsafe minus safe,<br>standardised) | —                       | 1.125 [1.062,<br>1.192]             | 1.195 [1.122, 1.273]        |
| Generator: Alpaca-7B (reference<br>Alpaca2-7B)  | —                       | —                                   | 2.058 [1.775, 2.387]        |
| Generator: Alpaca3-8B (reference<br>Alpaca2-7B) | —                       | —                                   | 4.161 [3.632, 4.767]        |
| Unsafe severity: moderate (reference<br>minor)  | —                       | —                                   | 0.753 [0.658, 0.862]        |

| Predictor                                    | Model 1/A<br>(position) | Model B<br>(position and<br>length) | Model D (fully<br>adjusted) |
|----------------------------------------------|-------------------------|-------------------------------------|-----------------------------|
| Unsafe severity: severe (reference<br>minor) | —                       | —                                   | 0.298 [0.220, 0.404]        |
| Prompt length (standardised)                 | —                       | —                                   | 1.088 [1.029, 1.151]        |
| Release split: test (reference train)        | —                       | —                                   | 0.918 [0.776, 1.086]        |

*Panel B. Model fit*

| Model                                    | n      | Clusters | Parameters | Converged | McFadden<br>pseudo R-<br>squared | AIC      |
|------------------------------------------|--------|----------|------------|-----------|----------------------------------|----------|
| Model 0: intercept only                  | 12,012 | 9,316    | 1          | yes       | 0.0000                           | 10,848.8 |
| Model 1/A: +<br>safe_response_position   | 12,012 | 9,316    | 2          | yes       | 0.0066                           | 10,778.9 |
| Model B: + length gap                    | 12,012 | 9,316    | 3          | yes       | 0.0087                           | 10,758.8 |
| Model C: + harm categories               | 12,012 | 9,316    | 19         | yes       | 0.0255                           | 10,608.0 |
| Model D: + source-observable<br>controls | 12,012 | 9,316    | 25         | yes       | 0.0808                           | 10,020.8 |

**Note.** A dash marks a predictor that is not entered in that column's model: the ladder is nested, so the length gap first appears in Model B and the generator, severity, prompt-length and release-split terms first appear in Model D.

**Note.** Largest variance inflation factor: 1.611 (Model D). No complete separation or sparse cell was detected.

**Note.** The length gap is defined as unsafe minus safe character length, standardised; its unstandardised mean is 72.50 characters with a standard deviation of 348.32.

**Note.** Generator terms are associations with the metadata field that records which model produced the responses. They are not risk assessments of those models.

Position is the strongest and most stable predictor in the ladder. The odds ratio for preferring the less-safe response when the safe response is listed second rather than first is 0.659 (95% CI 0.590 to 0.736) in the model containing position alone, and 0.612 (95% CI 0.545 to 0.686) in the fully adjusted model. Being listed first raises the odds of conflict by roughly sixty per cent, and the estimate moves only slightly as length, category and generator controls are added.

Length contributes a smaller but consistent effect. A one standard deviation increase in the length gap multiplies the odds of preferring the less-safe response by 1.125 (95% CI 1.062 to 1.192) in the position-adjusted model and by 1.195 (95% CI 1.122 to 1.273) in the final model. Longer unsafe responses are more likely to be preferred, though the effect is modest next to position.

The generator model contributes the largest association in the full specification. Relative to the reference generator, the odds of preferring the less-safe response are 2.058 (95% CI 1.775 to 2.387) for the first generator and 4.161 (95% CI 3.632 to 4.767) for the third. These are associations with which model produced the responses in a pair, not risk assessments of those models. The release does not provide the prompts or sampling conditions under which each subset was generated, so the contrast cannot be attributed to model capability, training data or sampling temperature with the available metadata.

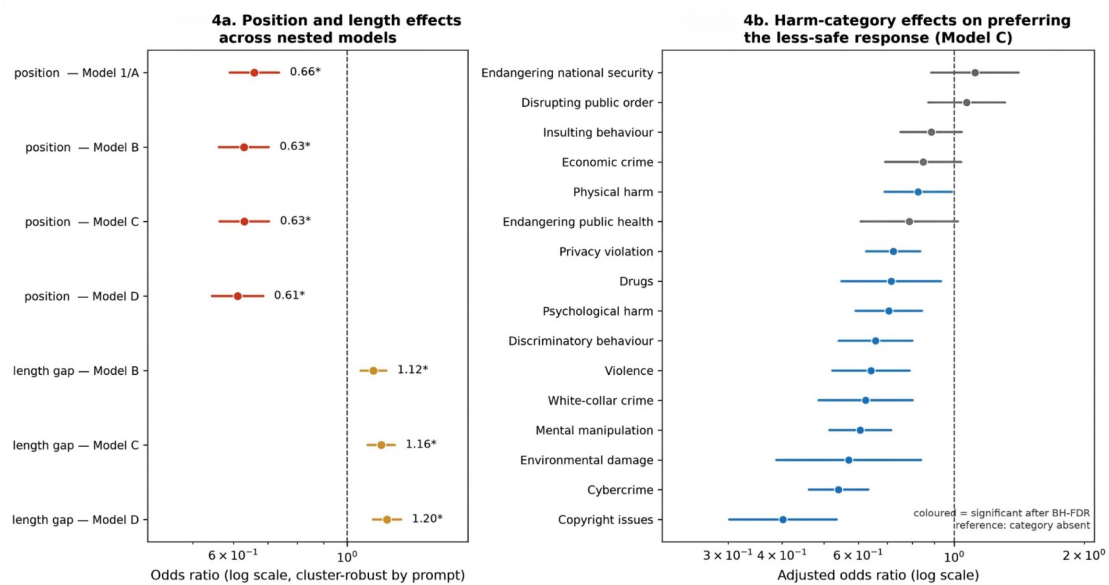
The remaining controls are minor. The release split is not associated with conflict after adjustment (odds ratio 0.918, 95% CI 0.776 to 1.086, *p* .319), which is consistent with the finding in Section 3.3 that the split is not prompt-disjoint. Prompt length has a small positive association (1.088, 95% CI 1.029 to 1.151). Severity is associated with lower conflict, with odds ratios of

0.753 for moderate and 0.298 for severe relative to minor harm, and this mirrors the descriptive pattern in Section 4.3.

The associations for the individual harm category indicators are negligible once position, length and generator are included. In the category model, the largest downward associations are for Copyright issues (odds ratio 0.619, 95% CI 0.473 to 0.810) and Cybercrime (odds ratio 0.622, 95% CI 0.539 to 0.717), and the largest upward association is for Disrupting public order (odds ratio 1.408, 95% CI 1.156 to 1.714). Adjusting for position and length leaves the category estimates close to their unadjusted values, which indicates that category differences are largely orthogonal to the presentation effects. Figure 4 shows the adjusted estimates for the models containing position, length and the category indicators.

**Figure 4**

Adjusted odds ratios from the cluster-robust logistic models, on a logarithmic scale.



Model fit improves most when the source-observable controls are added. McFadden's pseudo R-squared is 0.0000 for the intercept-only model, 0.0066 with position, 0.0087 with length added, 0.0255 with the harm indicators and 0.0808 in the full model. Every specification explains only a small fraction of the variation in the outcome, and the largest single increment in the ladder is the one that adds the source-observable controls rather than any content or presentation variable.

#### 4.6 Robustness checks

Table 5, Panel A reports the robustness battery. The baseline conflict rate is 16.725% (2,009 of 12,012).

**Table 5**

Robustness of the primary estimate and the dataset-version comparison. Panel A perturbs the estimation sample or the standard errors and reports the movement in the conflict rate relative to the baseline of 16.725%. Panel B compares descriptors across the two releases; the two estimation samples are never pooled.

*Panel A. Robustness checks on the primary estimate*

| Check                                                                            | n     | Denominator | Conflict rate (%)         | Change (pp)                                               |
|----------------------------------------------------------------------------------|-------|-------------|---------------------------|-----------------------------------------------------------|
| Baseline: v1 all valid discordant pairs                                          | 2,009 | 12,012      | 16.725                    | +0.000                                                    |
| Duplicates: excluding duplicate pairs                                            | 2,009 | 12,012      | 16.725                    | +0.000                                                    |
| Duplicates: exact (prompt,A,B) dedup via shipped sha256                          | 2,009 | 12,012      | 16.725                    | +0.000                                                    |
| Standard errors: non-clustered SE                                                | —     | 12,012      | —                         | coefficient = -0.41637, SE=0.04905, OR=0.6594, p=<1e-15   |
| Standard errors: cluster-robust SE                                               | —     | 12,012      | —                         | coefficient = -0.41637, SE=0.05634, OR=0.6594, p=1.47e-13 |
| Prompt concentration: most repeated 1% of prompts, share of all discordant pairs | 645   | 12,012      | 5.370 (share, not a rate) | —                                                         |
| Prompt dominance: one-prompt-one-weight (unweighted mean of prompt means)        | —     | 9,316       | 16.629                    | -0.096                                                    |
| Prompt dominance: cap 1 pair(s) per prompt                                       | 1,585 | 9,316       | 17.014                    | +0.289                                                    |
| Prompt dominance: cap 5 pair(s) per prompt                                       | 1,950 | 11,697      | 16.671                    | -0.054                                                    |
| Prompt dominance: cap 20 pair(s) per prompt                                      | 1,975 | 11,822      | 16.706                    | -0.019                                                    |
| Prompt dominance: cap 50 pair(s) per prompt                                      | 1,992 | 11,925      | 16.704                    | -0.021                                                    |
| Prompt dominance: drop top 1 prompt(s)                                           | 2,005 | 11,910      | 16.835                    | +0.110                                                    |
| Prompt dominance: drop top 5 prompt(s)                                           | 1,965 | 11,722      | 16.763                    | +0.038                                                    |
| Prompt dominance: drop top 20 prompt(s)                                          | 1,932 | 11,597      | 16.659                    | -0.065                                                    |

*Panel B. Primary against replication dataset*

| Quantity                                                  | PKU-SafeRLHF v1 | PKU-SafeRLHF-30K | Difference |
|-----------------------------------------------------------|-----------------|------------------|------------|
| Conflict rate (%)                                         | 16.72           | 61.06            | +44.33 pp  |
| Safety-discordant pairs                                   | 12,012          | 5,685            | -6,327     |
| Pairs preferring the less-safe response                   | 2,009           | 3,471            | +1,462     |
| Both safe (% of analytic)                                 | 41.08           | 33.07            | -8.02      |
| Both unsafe (% of analytic)                               | 44.28           | 47.90            | +3.62      |
| Second-listed response preferred, concordant pairs (%)    | 64.90           | 45.42            | -19.48     |
| Mean unsafe-minus-safe length (characters)                | 72.50           | 21.76            | -50.74     |
| Unsafe response is longer (%)                             | 60.60           | 53.97            | -6.63      |
| Unique prompts                                            | 39,990          | 7,773            | -32,217    |
| Cluster-robust odds ratio for safe response listed second | 0.659           | 1.395            | —          |
| Cluster-robust odds ratio for length gap                  | 1.125           | 4.326            | —          |

*Note.* Panel A. The concentration row reports how much of the discordant sample the most repeated prompts contribute. It is a share of pairs rather than a conflict rate, so it carries no change against the baseline. For the one-prompt-one-weight row the denominator column reports the number of prompts rather than the number of pairs. Every other row in the panel is a conflict rate over the stated denominator.

**Note.** Largest movement across any perturbation of the headline estimate: 0.289 percentage points (C prompt dominance: cap 1 pair(s) per prompt).

**Note.** Risk difference across releases: +44.33 percentage points (95% CI +42.90 to +45.76); risk ratio 3.651 (95% CI 3.49 to 3.82); odds ratio 7.806 (95% CI 7.27 to 8.39) unclustered and 7.806 (95% CI 7.14 to 8.54) clustered by prompt.

**Note.** The comparison is interpretable only as a difference between dataset versions. It is not a trend, not evidence of improvement or deterioration, and not causal.

Duplicate pairs do not affect the estimate. No exact duplicate comparison pair survives into the discordant sample, and de-duplicating on the shipped response digests removes 0 pairs and leaves the rate unchanged at 16.725%.

Clustering matters for precision but not for the point estimate. Standard errors clustered by prompt are 1.149 times the non-clustered standard errors for the position coefficient (0.04905 against 0.05634). Confidence intervals widen accordingly, and all substantive conclusions are unchanged.

Prompt dominance is the only structural feature of these data that could plausibly distort an unweighted rate. The most repeated 1% of prompts, 93 of the 9,316 prompts that appear in the discordant sample, account for 645 of the 12,012 discordant pairs, or 5.370% of the sample. That figure is a share of pairs and not a conflict rate, and it is the only entry in Table 5, Panel A of that kind. Weighting every prompt equally rather than every pair equally gives an unweighted mean of the prompt-level rates of 16.629%, which is -0.096 percentage points from the baseline. Capping the number of pairs per prompt at 1, 5, 20 and 50 pairs gives rates of 17.014, 16.671, 16.706, 16.704 per cent, and dropping the 1, 5 and 20 most frequent prompts gives rates of 16.835, 16.763, 16.659 per cent. The largest movement across any perturbation of the headline estimate is 0.289 percentage points, produced by the most restrictive one-pair-per-prompt cap. Prompt repetition is therefore a large feature of the sample composition and a small influence on the estimate.

The position and length stratifications reported in Sections 4.4 and 4.5 are contrasts rather than perturbations, and they move the rate by design. Conditioning on where the safe response sits moves the conflict rate from 14.242% to 20.118%, and conditioning on length direction moves it from 13.945% to 18.533%. Both contrasts are consistent with the regression results and neither is a threat to the headline estimate, which is the unconditional rate on all discordant pairs.

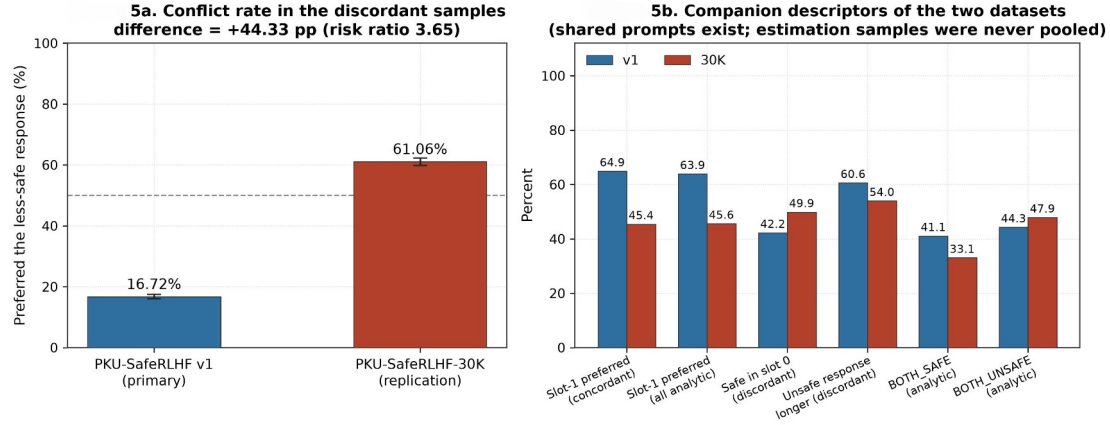
#### **4.7 Replication in PKU-SafeRLHF-30K**

The replication sample contains 29,861 analytic pairs, of which 5,685 (19.04%) are safety-discordant. The recorded preference selects the less-safe response in 61.06% of those pairs (3,471 of 5,685; 95% CI 59.78 to 62.32). Against the primary estimate of 16.72%, the risk difference is 44.33 percentage points (95% CI 42.90 to 45.76), the risk ratio is 3.651 (95% CI 3.49

to 3.82) and the odds ratio is 7.806 (95% CI 7.27 to 8.39) without clustering and 7.806 (95% CI 7.14 to 8.54) with standard errors clustered by prompt. Figure 5a shows the two rates.

**Figure 5**

The primary and replication datasets side by side.



The four-way relation distribution differs as well. In the replication sample 33.07% of pairs are both-safe and 47.90% are both-unsafe, against 41.08% and 44.28% in the primary sample. The composition differences across all four relation classes are reported in Table 5, Panel B.

The position and length effects do not merely weaken in the replication sample. They reverse. In the primary sample the second-listed response is preferred 64.90% of the time among concordant pairs; in the replication sample the corresponding figure is 45.42%, and the cluster-robust odds ratio for position in the replication regression is 1.395 (95% CI 1.220 to 1.596) against 0.659 in the primary sample. The length gap odds ratio rises from 1.125 to 4.326 (95% CI 3.788 to 4.941). The unsafe response is the longer one in 53.97% of replication discordant pairs against 60.60% in the primary sample, while the mean unsafe-minus-safe length gap falls from 72.5 characters to 21.76 characters. Figure 5b places the two sets of descriptors side by side.

The prompt-level 14-category taxonomy available for the replication sample is a secondary derived analysis and is not comparable with the response-level 19-category scheme. Its conflict rates are uniformly higher than the primary category rates, ranging from 48.91% for *misinformation\_regarding\_ethics\_laws\_and\_safety* to 74.83% for *privacy\_violation*, and the range across those 14 categories is wider than in the primary sample. Because the two taxonomies differ in level, scheme and derivation, these results are reported in the Supplementary Material and are not interpreted against the primary categories.

The replication result is not a failure to reproduce in the ordinary sense. The same operational preference-safety relation was evaluated using the released fields on a separately collected release drawn from the same prompt pool, and it produced an estimate more than three times larger with the opposite sign on every presentation variable. Sections 5.4 and 6 discuss what can and cannot be concluded from that.

## 5. Discussion

### *5.1 Human preference is not a direct safety signal*

In the primary dataset, 83.28% of safety-discordant comparisons are decided in favour of the safer response. Any claim that people generally prefer unsafe answers would be inconsistent with this finding, and no such claim is made here. The overall preference and the independent safety annotation agree in the large majority of cases where they can be compared.

The residual matters precisely because the agreement is not complete. Among the 12,012 discordant pairs, the recorded preference selected the response that the safety annotation marked as unsafe in 2,009 of them, a rate of 16.72% with a confidence interval that excludes zero by a wide margin. Two annotations elicited for the same pair of responses therefore disagree roughly one time in six. If the preference label were a faithful measurement of safety-relevant judgement, that disagreement would be difficult to explain.

Three interpretations of the disagreement are consistent with the data, and the data do not distinguish between them. The safety annotation may be wrong in some cases, particularly where the harm category is contested or context-dependent. The preference judgement may weight a property other than safety, which is the natural reading if the annotator was asked which response was better overall rather than which was safer. Or the two judgements may be answering genuinely different questions and disagreeing where those questions come apart. The last reading is the most conservative and requires no assumption of error on either side. On any of the three readings, the practical implication is the same: an overall preference label and a safety label are not interchangeable, and a pipeline that needs a safety signal should measure safety rather than infer it from preference.

### *5.2 Preference-safety conflict is not primarily taxonomy-driven*

The second question was whether conflict concentrates in particular kinds of harmful content. It does not, at least not to a degree that would make harm category the organising explanatory variable. Conflict rates across the 19 categories span 12.30 percentage points, from 9.36% to 21.66%, and a substantial number of the differences are distinguishable from zero. The magnitudes, however, are small. The largest Cramér's  $V$  across the category tests is 0.0571, which is negligible on any conventional reading of that statistic, and the category odds ratios stay close to one. When the categories are entered alongside position and length, their estimates barely move, which indicates that they are not absorbing the presentation effects.

The severity gradient is the clearest category-side result and deserves separate comment. Conflict declines monotonically from 21.83% for minor harm to 7.12% for severe harm. That is a larger spread than the range across categories, and it points in an interpretable direction: when annotators do prefer the unsafe response, the harm tends to be less severe. Severity is not an independent dimension of safety in this schema, since the safe case is defined as zero severity, so the gradient cannot be read as an independent safety signal. It can be read as internal consistency evidence about the annotation: a dataset in which the propensity to prefer an unsafe response fell as the harm became more serious is behaving the way a safety-aware annotation process would be expected to behave. That this gradient exists while category membership matters so little

suggests the categories carry less information about annotator judgement than their prominence in the taxonomy would imply.

### ***5.3 Presentation and dataset effects matter more than content categories***

The variables most strongly associated with conflict are properties of the comparison task rather than of the harmful content. Position is the clearest case. The odds of preferring the less-safe response are 0.659 times as high when the safe response is listed second as when it is listed first, an association that survives every adjustment in the ladder. On the effect-size measures reported in Section 4.3 it exceeds all 19 of the harm category associations, though not the severity gradient, which is slightly larger. The conditional rates make the mechanism visible: the probability of taking the second-listed response is 20.12% when the safe response is listed first and 85.76% when it is listed second. A position preference of this size will produce apparent safety conflict whenever the safe option happens to be listed first.

Length is weaker but points the same way. Longer responses are preferred more often, unsafe responses are longer more often than safe ones in discordant pairs, and the conflict rate is 18.53% when the unsafe response is longer against 13.94% when it is not. A modest preference for length therefore transfers into a modest preference-safety conflict without any safety content being involved. The generator model shows the largest adjusted association of any predictor, with odds ratios of 2.058 and 4.161 relative to the reference subset.

These findings converge on a single point about measurement. The observed alignment between preference and safety is not a fixed property of human judgement that the dataset happens to record. It is a function of the annotation instrument, and the instrument includes the order in which options are presented, the length of the responses, the models that produced them and the composition of the comparison set. Position bias in evaluation has been documented extensively for language-model judges (P. Wang et al., 2024; Zheng et al., 2023) and for self-preference (Panickssery et al., 2024), and length bias has been documented for judges and for reinforcement learning from human feedback (Saito et al., 2023; Singhal et al., 2023). The results here suggest the same biases operate in human preference annotation over safety-relevant pairs, and that their magnitude matches or exceeds the content-category effects that safety taxonomies are designed to capture.

### ***5.4 What the replication reversal means***

The replication sample produced a conflict rate of 61.06%, against 16.72% in the primary sample, a risk difference of 44.33 percentage points and a risk ratio of 3.651. The direction of the relationship between preference and safety is opposite in the two datasets: in the primary sample the safer response is preferred most of the time, and in the replication sample the less-safe response is preferred most of the time. The position effect reverses as well, from an odds ratio of 0.659 to 1.395, and the length effect strengthens from 1.125 to 4.326.

Several features of the two collections are consistent with this divergence, and none of them can be isolated with the metadata that accompany the releases. The two files come from different release rounds and expose different annotation fields. The primary release carries response-level harm categories and severity levels alongside the preference field, the relative-safety field and

the two binary safety labels. The replication release carries the preference field, the relative-safety field and the two binary safety labels and nothing further. The primary release documents three generator models and the replication release does not record its generators at all. Response length differs substantially, with a mean of 545 characters in the primary sample against 356 in the replication sample. The composition of the comparison set differs, with 14.64% discordant pairs in the primary sample against 19.04% in the replication sample. The position distribution differs, and position is the strongest predictor in both samples.

Because these features vary together, the present analysis cannot attribute the divergence to any one of them. It would be possible to describe the divergence as a change in annotation standards over time, and equally possible to describe it as an artefact of a different generator mix or a different response length distribution. Neither description is supported. What the comparison does establish is that the preference-safety relationship is not a stable property of the annotation target. It is sensitive to how the data were collected, and the sensitivity is large enough to reverse the sign of the headline result.

This is the central measurement claim of the paper. Human preference has been treated in parts of the alignment literature as a quantity that can be pooled, aggregated and compared across collections. The evidence here indicates that preference is instead a construction-dependent measurement, and that a preference dataset should be understood as a record of a particular annotation procedure rather than as a neutral sample of a stable human disposition. Work on annotator demographics already shows that who annotates changes the recorded preferences (H. R. Kirk et al., 2024; Sorensen et al., 2024), and work on human label noise shows that the labels are noisier than commonly assumed (Hosking et al., 2023). The present result extends the point from who annotates to how the annotation was elicited.

### ***5.5 Implications for alignment and governance***

Debate about the governance of foundation models has repeatedly returned to the question of what evidence should be required before a system is deployed (Bommasani et al., 2021). If preference and safety are related but distinct measurements, several practical consequences follow for how preference data are used in alignment and how they are documented.

A preference dataset that will inform safety-relevant decisions should report the independent safety annotations that accompany it, if they exist, and should make clear when they do not. It should report the position distribution of the options it presents, since position carries first-order predictive weight and is trivially controllable through randomisation. It should report length distributions, because length behaves as a preference proxy that can be mistaken for a safety signal. It should document the models that generated the compared responses, because generator identity was the largest adjusted association observed here. It should record version changes explicitly, because the replication result indicates that the relationship can reverse between releases. It should report prompt clustering, because a small number of prompts can contribute a large proportion of pairs and unclustered inference will understate uncertainty. It should report the tie and disagreement structure, since the schema examined here cannot express a tie at all. And it should report category-level safety information at the level at which it

was annotated, rather than presenting a collapsed taxonomy as if it were a set of mutually exclusive classes.

Where safety is a core objective, the results argue for measuring it directly rather than reading it off an overall preference label. A preference label that is optimised during training will be optimised as it is, with whatever mix of reasons the annotators applied. Available evidence on proxy objectives suggests that this mix can be exploited (Amodei et al., 2016; Skalse et al., 2022), and the present results indicate that the mix is not dominated by safety when a safety distinction is available. Multi-objective formulations that keep safety as a separately specified constraint (Dai et al., 2023) or that specify normative content directly rather than inferring it from comparison (Bai et al., 2022a; Huang et al., 2024) are consistent with that conclusion. The recommendation is not that preference data are unusable. It is that a preference dataset should not be reported as evidence about safety alignment unless safety was measured separately, and that the documentation practices already developed for datasets and models (Geburu et al., 2021; Mitchell et al., 2019) should be applied to the properties shown here to matter.

### **5.6 Relation to model-side safety evaluation**

This study addresses one side of a two-sided question. Benchmark-style safety evaluation asks whether a model recognises or avoids harmful behaviour (Mazeika et al., 2024; Zhang et al., 2024; Röttger et al., 2024). The present analysis asks whether the human judgements used to train and select such models align with independently annotated safety. Both are calibration questions about the same systems, and they can be combined. A model could be internally well calibrated on safety while the preference data used to align it encode weak or unstable safety information, and a dataset could record strong safety preferences while the model trained on it degrades on safety benchmarks under distribution shift (Bianchi et al., 2023). The evidence here concerns the feedback side, where the preference-safety relationship is substantial but incomplete and demonstrably sensitive to collection procedure. Model-side and feedback-side calibration are complementary measurements, and reporting only one of them leaves the other unverified.

## **6. Limitations**

The limitations of this study follow from its design and from the contents of the datasets it uses. Twelve are listed.

(1) **The study is a secondary analysis of existing public datasets.** Its scope is bounded by what those releases contain, and several of the limitations below are structural rather than incidental. The prompt pool, the response generation, the annotation protocol, the instructions given to annotators and the decision of which fields to publish were all settled before this analysis began, and none of them was under the authors' control. Every question this study can answer is therefore a question that the released fields happen to support.

(2) **The preference label is an overall preference, not a helpfulness measurement.** Neither release provides a helpfulness score, and neither records the instructions given to annotators, so the study cannot determine which properties drove any individual choice. No

result here should be read as evidence about a helpfulness-harmlessness trade-off, and none is presented that way.

(3) **The datasets provide no annotator metadata.** Annotator identity, demographics, agreement statistics and the number of annotators per pair are all absent, so annotation heterogeneity cannot be modelled and no annotator-level claim is made. Evidence from other collections suggests that annotator composition affects the recorded preferences (H. R. Kirk et al., 2024), which means the aggregate rates reported here are averages over unknown and possibly heterogeneous pools.

(4) **The annotation schema cannot express a tie.** The preference field takes only two values, so the absence of ties is a property of the schema rather than a measurement. Any instruction that permitted annotators to record no preference or equal quality is invisible in these data, and pairs that a different instrument would have marked as tied are recorded as preferences.

(5) **Observations are heavily clustered within prompts.** The primary sample draws 82,067 pairs from 39,990 prompts, with a maximum of 980 pairs for a single prompt. All inference is clustered by prompt and prompt-level sensitivity checks are reported, but residual dependence through shared subject matter across different prompts cannot be excluded.

(6) **The harm taxonomy is multi-label at response level.** Categories overlap heavily, so the per-category rates are not independent contrasts and the reported spread across categories is partly a description of overlap structure. The pairwise comparisons between categories are reported with false discovery rate control, but they inherit the same dependence.

(7) **All results are observational associations.** The design supports statements about which pair-level characteristics co-vary with the conflict outcome and not about the effect of any characteristic on the outcome. Adjusting for position and length changes the estimates of the other coefficients, which demonstrates that the covariates are related, not that any of them causes conflict.

(8) **Generator identity is confounded with everything else that varies between the three generator subsets.** Response model was the largest adjusted association observed, but the release does not document the sampling conditions, prompt subsets or annotation batches associated with each subset, so the contrast cannot be attributed to the models themselves. It is reported as an association with the metadata field.

(9) **The divergence between the two datasets cannot be decomposed.** The releases differ in annotation round, available fields, generator documentation, response length distribution, prompt set and pair composition, and these differences co-vary. The comparison is interpretable as a difference between dataset versions and not as a trend, an improvement, a deterioration or a causal effect of any identifiable factor.

(10) **External generalisation is limited.** Both datasets are drawn from a shared prompt pool, use a single annotation target and originate from one research group's releases. The conflict rates reported here are properties of these collections. Whether they transfer to other preference collections, other annotator populations, other languages or other task formats is not established by this study.

(11) **The safety labels are annotation constructs rather than ground truth about real-world harm.** The labels reflect a taxonomic judgement made by annotators under an undisclosed protocol, and disagreement between a preference and a safety label may reflect error or contestation in either annotation rather than in the preference alone. The analysis treats the safety label as the reference for the purpose of defining conflict, which is a definitional choice rather than a claim about which annotation is correct.

(12) **The preference results should not be read as statements about human moral preference.** The annotators recorded comparative judgements about short model-generated responses under an unobserved protocol. Those judgements are inputs to a technical pipeline, and treating them as a measurement of moral attitudes would extend them far beyond what the collection supports.

## 7. Conclusion

Human preference and response safety overlap substantially but are not interchangeable. Among more than twelve thousand paired comparisons in which one response was annotated as safe and the other as unsafe, the recorded human preference selected the safer response 83.28% of the time. The remaining 16.72% is a precise, stable and non-zero rate of disagreement.

The disagreement is not organised primarily by the kind of harm involved. Conflict rates across the 19 harm categories span 12.30 percentage points, but every category association is negligible in magnitude, with odds ratios close to one and confidence intervals spanning one for most of the categories. Response position, response length and the model that generated the responses are more strongly associated with conflict than any harm category.

The size and even the direction of the relationship depend on dataset construction and response presentation. A separate replication release drawn from a different release round of the same data family produces a conflict rate more than three times larger, with opposite signs on the position and length effects. The preference-safety relationship is therefore not a stable property that a preference dataset samples; it is a property of the collection procedure.

Preference-based alignment should treat safety as a separately measured objective rather than assume that an overall preference label encodes it. Where safety matters, it should be annotated, reported and verified in its own right.

## Declarations

**Data availability.** The analysis uses three publicly released datasets from the PKU-Alignment organisation, each pinned to a specific revision and redistributable under CC BY-NC 4.0. The primary dataset is PKU-Alignment/PKU-SafeRLHF at revision 9421ffa4ec3fa40a1f1a7d567b4d525079477ecb. The replication dataset is PKU-Alignment/PKU-SafeRLHF-30K at revision 334ad7d4d008421be5a9b098dd256951eba1a5c5. The supplementary taxonomy source is PKU-Alignment/BeaverTails at revision 8401fe609d288129cc684a9b3be6a93e41cfe678. A download manifest recording the official and mirror location, byte count and SHA-256 digest of each retrieved file is included with the analysis artefacts. The artefacts generated by this study comprise the data audit and audit report, the cleaned analysis files, the analysis scripts for every stage, the machine-readable results used to generate the tables and figures, the reference verification record and a numeric consistency audit. These are available from the corresponding author on request and will be deposited alongside the accepted version.

**Ethics statement.** This is a secondary analysis of existing, publicly released research datasets. No human participants were recruited, contacted or observed for this study, and no new personal data were collected. The datasets used do not contain annotator identifiers, demographic information or any other personal data. The authors did not have access to any information that could identify an individual annotator. Because the study involves no participant contact and no personal data, it falls outside the usual scope of ethics review in many jurisdictions, but institutional requirements differ: authors and readers should confirm the applicable institution-specific requirements before submission rather than assume that review is unnecessary.

**AI use disclosure.** AI tools were used in the preparation of this work. Specifically, AI assistance was used for code generation and debugging of the analysis scripts, for structuring the statistical workflow, for drafting and language editing of the manuscript, and for programmatic verification of the reference list against bibliographic registries. Every numeric value reported in the manuscript is computed by reproducible analysis scripts from the sourced datasets rather than generated or inferred by a language model, and each value is checked against its source file by an automated consistency audit included with the submission materials. Reference records were resolved against OpenAlex, Crossref and DataCite, and the resolution record is included. The authors retain full responsibility for the analysis, the interpretation, the accuracy of the manuscript and the decision to submit.

**Funding.** This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

**Competing interests.** The authors declare no competing interests.

**Author contributions.** All authors contributed to the conception of the study, the interpretation of the results and the revision of the manuscript.

## References

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv* <https://doi.org/10.48550/arxiv.1606.06565>
- Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., ... Kaplan, J. (2021). A general language assistant as a laboratory for alignment. *arXiv* <https://doi.org/10.48550/arxiv.2112.00861>
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv* <https://doi.org/10.48550/arxiv.2212.08073>
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., ... Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv* <https://doi.org/10.48550/arxiv.2204.05862>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bianchi, F., Suzgun, M., Attanasio, G., Röttger, P., Jurafsky, D., Hashimoto, T., & Zou, J. (2023). Safety-Tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. *arXiv* <https://doi.org/10.48550/arxiv.2309.07875>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv* <https://doi.org/10.48550/arxiv.2108.07258>
- Cameron, A. C., & Miller, D. L. (2015). A practitioner’s guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317–372. <https://doi.org/10.3368/jhr.50.2.317>
- Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C. R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., ... Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv* <https://doi.org/10.48550/arxiv.2307.15217>
- Chiang, W. L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot Arena: An open platform for evaluating LLMs by human preference. *arXiv* <https://doi.org/10.48550/arxiv.2403.04132>
- Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *arXiv* <https://doi.org/10.48550/arxiv.1706.03741>
- Cui, J., Chiang, W. L., Stoica, I., & Hsieh, C. J. (2024). OR-Bench: An over-refusal benchmark for large language models. *arXiv* <https://doi.org/10.48550/arxiv.2405.20947>
- Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., Wang, Y., & Yang, Y. (2023). Safe RLHF: Safe reinforcement learning from human feedback. *arXiv* <https://doi.org/10.48550/arxiv.2310.12773>
- Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36, 30039–30069. <https://doi.org/10.48550/arxiv.2305.14387>
- Dubois, Y., Galambosi, B., Liang, P., & Hashimoto, T. B. (2024). Length-Controlled AlpacaEval: A simple way to debias automatic evaluators. *arXiv* <https://doi.org/10.48550/arxiv.2404.04475>

- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1) <https://doi.org/10.1214/aos/1176344552>
- Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., Jones, A., Bowman, S., Chen, A., Conerly, T., DasSarma, N., Drain, D., Elhage, N., El-Showk, S., Fort, S., ... Clark, J. (2022). Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv* <https://doi.org/10.48550/arxiv.2209.07858>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
- Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., Rauh, M., Weidinger, L., Chadwick, M., Thacker, P., Campbell-Gillingham, L., Uesato, J., Huang, P. S., Comanescu, R., Yang, F., See, A., Dathathri, S., Greig, R., Chen, C., ... Irving, G. (2022). Improving alignment of dialogue agents via targeted human judgements. *arXiv* <https://doi.org/10.48550/arxiv.2209.14375>
- Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2021). Unsolved problems in ML safety. *arXiv* <https://doi.org/10.48550/arxiv.2109.13916>
- Hosking, T., Blunsom, P., & Bartolo, M. (2023). Human feedback is not gold standard. *arXiv* <https://doi.org/10.48550/arxiv.2309.16349>
- Huang, S., Siddharth, D., Lovitt, L., Liao, T. I., Durmus, E., Tamkin, A., & Ganguli, D. (2024). Collective Constitutional AI: Aligning a language model with public input. *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1395–1417. <https://doi.org/10.1145/3630106.3658979>
- Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., & Khabsa, M. (2023). Llama Guard: LLM-based input-output safeguard for Human-AI conversations. *arXiv* <https://doi.org/10.48550/arxiv.2312.06674>
- Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Zhang, C., Sun, R., Wang, Y., & Yang, Y. (2023). BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36, 24678–24704. <https://doi.org/10.48550/arxiv.2307.04657>
- Ji, J., Hong, D., Zhang, B., Chen, B., Dai, J., Zheng, B., Qiu, T., Zhou, J., Wang, K., Li, B., Han, S., Guo, Y., & Yang, Y. (2025). PKU-SafeRLHF: Towards multi-level safety alignment for LLMs with human preference. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 31983–32016. <https://doi.org/10.48550/arxiv.2406.15513>
- Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., & Hale, S. A. (2024). The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. <https://doi.org/10.48550/arxiv.2404.16019>
- Kirk, R., Mediratta, I., Nalmpantis, C., Luketina, J., Hambro, E., Grefenstette, E., & Raileanu, R. (2023). Understanding the effects of RLHF on LLM generalisation and diversity. *arXiv* <https://doi.org/10.48550/arxiv.2310.06452>
- Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z. R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., Es, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., & Mattick, A. (2023). OpenAssistant conversations—democratizing large language model alignment. *Advances in Neural Information Processing Systems*, 36, 47669–47681. <https://doi.org/10.48550/arxiv.2304.07327>
- Le Jeune, P., Liu, J., Rossi, L., & Dora, M. (2025). RealHarm: A collection of real-world language model application failures. *ACL Proceedings of the First Workshop on LLM Security (LLMSEC)*, 87–100. <https://doi.org/10.48550/arxiv.2504.10277>
- Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22. <https://doi.org/10.1093/biomet/73.1.13>
- Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., Forsyth, D., & Hendrycks, D. (2024). HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv* <https://doi.org/10.48550/arxiv.2402.04249>

- Mitchell, M., Wu, S., Zaldívar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744. <https://doi.org/10.48550/arxiv.2203.02155>
- Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37, 68772–68802. <https://doi.org/10.48550/arxiv.2404.13076>
- Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 53728–53741. <https://doi.org/10.48550/arxiv.2305.18290>
- Röttger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2024). XSTest: A test suite for identifying exaggerated safety behaviours in large language models. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5377–5400. <https://doi.org/10.48550/arxiv.2308.01263>
- Saito, K., Wachi, A., Wataoka, K., & Akimoto, Y. (2023). Verbosity bias in preference labeling by large language models. *arXiv* <https://doi.org/10.48550/arxiv.2310.10076>
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. *arXiv* <https://doi.org/10.48550/arxiv.2310.13548>
- Singhal, P., Goyal, T., Xu, J., & Durrett, G. (2023). A long way to go: Investigating length correlations in RLHF. *arXiv* <https://doi.org/10.48550/arxiv.2310.03716>
- Skalse, J., Howe, N. H. R., Krashennikov, D., & Krueger, D. (2022). Defining and characterizing reward hacking. *Advances in Neural Information Processing Systems*, 35, 9460–9471. <https://doi.org/10.48550/arxiv.2209.13085>
- Sorensen, T., Moore, J., Fisher, J., Gordon, M., Miresghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., & Choi, Y. (2024). A roadmap to pluralistic alignment. *arXiv* <https://doi.org/10.48550/arxiv.2402.05070>
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2020). Learning to summarize from human feedback. *arXiv* <https://doi.org/10.48550/arxiv.2009.01325>
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv* <https://doi.org/10.48550/arxiv.2307.09288>
- Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2024). Large language models are not fair evaluators. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9440–9450. <https://doi.org/10.48550/arxiv.2305.17926>
- Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., & Liu, Q. (2023). Aligning large language models with human: A survey. *arXiv* <https://doi.org/10.48550/arxiv.2307.12966>
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., ... Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv* <https://doi.org/10.48550/arxiv.2112.04359>
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212. <https://doi.org/10.1080/01621459.1927.10502953>

- Zhang, Z., Lei, L., Wu, L., Sun, R., Huang, Y., Long, C., Liu, X., Lei, X., Tang, J., & Huang, M. (2024). SafetyBench: Evaluating the safety of large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15537–15553. <https://doi.org/10.48550/arxiv.2309.07045>
- Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36, 46595–46623. <https://doi.org/10.48550/arxiv.2306.05685>
- Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-Tuning language models from human preferences. *arXiv* <https://doi.org/10.48550/arxiv.1909.08593>