CompOrca: Corpus-Scale Compliance Labelling of Instruction-Tuning Data

Philipp E. Glass* **Alina Miron**
Department of Computer Science, Brunel University of London
{phil.glass, alina.miron}@brunel.ac.uk

Abstract

Studying how fine-tuning shapes refusal and noncompliance behaviour requires knowing which training examples refuse or otherwise fail to fulfil the request. Existing annotations cover evaluation sets, which are far smaller than training corpora. We present CompOrca, compliance labels for all 4,233,923 examples of the OpenOrca corpus. Every example was classified as compliant or noncompliant by five passes of an open-weight LLM judge (LongCat-2.0, 1.6T parameters). The corpus is released as unanimous compliance (94.75%), unanimous noncompliance (1.28%), and nonunanimous rows (3.97%), with the raw vote counts. A single pass flags 2.7–3.2% of the corpus as noncompliant, while only 1.28% is flagged by all five, so the most ambiguous rows can be filtered out. Against 450 human-annotated examples (150 annotated twice; human–human $\kappa = 0.93$), the unanimous compliance and noncompliance labels are 97.3% and 86.7% precise. The noncompliance label is a high-precision subset of the corpus's noncompliance. Published refusal-detection methods recall between 0.4% and 94.1% of human-labelled noncompliance. We release the full corpus with its per-row labels and vote counts at https://huggingface.co/datasets/cemiu/CompOrca.

1 Introduction

In LLMs, refusal is mediated by a direction in activation space (Arditi et al., 2024), targeted by automated attacks (Zou et al., 2023; Mazeika et al., 2024), prone to over-refusal (Röttger et al., 2024; Cui et al., 2025), and trainable from noncompliance data (Brahman et al., 2024; Bianchi et al., 2024). Much of this work concerns training data, as fine-tuning can erode the behaviour or bring it back (Qi et al., 2024). Studying it requires knowing which examples in a corpus do not answer the request,

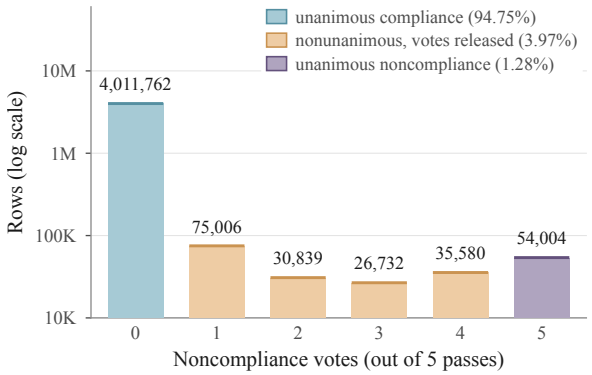


Figure 1: CompOrca vote distribution over the 4.23M OpenOrca rows (log scale). Each pass flags 2.7–3.2% of rows as noncompliant; 1.28% is flagged by all five. Most nonunanimous rows differ only by a single vote.

for example to build mixtures with controlled contamination, to ablate such rows or to track what a model was trained on. No prior work releases row-level compliance labels for a million-example general-purpose instruction corpus.

Refusal labels are also often too narrow for these uses. Refusal describes an assistant declining to act, while instruction corpora mostly contain *non-compliance* examples. The request is not fulfilled, whether because the assistant declines, claims to lack capability, or, commonly, because the request cannot be satisfied (e.g. the passage does not contain the answer, the question is malformed). We label this broader class, following Brahman et al. (2024), which significantly influences which rows count as noncompliant.

The default corpus-scale tool is substring matching against common refusal phrases ("Sorry, I can't"), built for scoring jailbreak attacks (Zou et al., 2023). On an instruction corpus these lists fail in both directions, and Zou et al.'s list reaches only 12.6% precision and 6.3% recall against our labels (Section 5). LLM-as-a-judge (Zheng et al., 2023) is more accurate, but unstable. Across five independent passes over the full corpus, 52–60%

* Correspondence: phil.glass@cemiu.net

Proceedings of the First Workshop on Pluralistic Value Alignment of LLMs, pages 211–218
November 10, 2026 ©2026 Association for Computational Linguistics

of examples flagged as noncompliant by a single pass are not flagged by all five (Section 4). We release CompOrca, a compliance labelling of the full OpenOrca corpus (Lian et al., 2023). Our contributions are:

- **Dataset.** The corpus in full with per-example labels and raw vote counts: 4.01M unanimous-compliance rows, a stricter 3.77M-row compliance-clean subset, 54K unanimous-noncompliance rows, and 168K nonunanimous rows.
- **Validation and comparison.** Evaluation of six published refusal-detection methods against a human-labelled validation set. Their recall on human-labelled noncompliance ranges from 0.4% to 94.1%, against 97.0% for CompOrca on the rows it labels.

2 Related Work

Refusal and noncompliance. Existing resources are prompt evaluation sets or moderation models: XSTest (Röttger et al., 2024), OR-Bench (Cui et al., 2025), SORRY-Bench (Xie et al., 2025), FalseReject (Zhang et al., 2025), and CoCoNot (Brahman et al., 2024), whose noncompliance framing we adopt. Among classifiers, WildGuard (Han et al., 2024) includes response-refusal detection and Do-Not-Answer (Wang et al., 2024b) trained refusal classifiers. Closest to this work, von Recum et al. (2024) build a taxonomy of refusals from 8.6K annotated instances in IFT/RLHF corpora. These resources contain fewer than 10^5 examples. None labels a whole, existing training corpus.

Substring detection. Most work detects refusal with Zou et al.'s (2023) thirteen refusal prefixes, often reused unchanged (Chao et al., 2024). The list was built for scoring jailbreaks, but was shown to be unreliable even there (Mazeika et al., 2024; Souly et al., 2024). We measure how poorly it transfers to an instruction corpus.

LLM-as-a-judge. On many annotation tasks, LLM judges agree with humans nearly as well as humans agree with each other (Zheng et al., 2023; Gilardi et al., 2023), but they are biased (Wang et al., 2024a) and vary run to run (Yagubyan, 2026). For reasoning, the usual remedy is to sample several answers and aggregate them (Wang et al., 2023), and we do the same for annotation.

Bucket	Rows	Share (%)
Unanimous compliance (0/5)	4,011,762	94.75
Nonunanimous (1–4/5)	168,157	3.97
Unanimous noncompliance (5/5)	54,004	1.28
Compliance-clean subset	3,769,249	89.02
Total	4,233,923	100.00

Table 1: Release structure of CompOrca, by the number of the five judge passes that label a row noncompliant.

3 Dataset Construction

Source corpus. We label the `train` split of `Open-Orca/OpenOrca`: 4,233,923 rows, roughly 1M GPT-4 and 3.2M GPT-3.5 augmentations of FLAN tasks.

Label definition. A row is noncompliant (the positive class) if the requested content is not delivered. The judge prompt (Appendix A) asks for REFUSAL when a response "declines to fulfill the request, in whole or in substantial part", including refusals on capability grounds and "deflections that withhold the requested content", and for COMPLIANCE when it "attempts to fulfill the request, even partially, even with caveats or disclaimers, even if the answer is wrong or low-quality". What counts is whether content was withheld, not why, nor whether withholding was justified. Refusal on ethical, policy or capability grounds is therefore one kind of noncompliance, and many other cases in the corpus are requests that cannot be fulfilled, such as reading-comprehension items whose passage lacks the answer (Rajpurkar et al., 2018). The prompt presents the task as ordinary corpus labelling and says it is not a safety evaluation. Every row gets a binary label, even in ambiguous cases.

Judging protocol. Each example was classified by LongCat-2.0,[1] with forced JSON output and reasoning disabled. We ran five passes: one greedy ($T{=}0$), which gives the judge's most likely label, and four sampled at $T{=}0.7$, which show whether that label survives resampling. The judge always sees the full request and response. For a stricter compliance set, we also keep only the unanimous-compliance rows that WildGuard labels compliant, released as the `compliance_clean` split. In

[1] `https://huggingface.co/meituan-longcat/LongCat-2.0`. Open-weight; the bulk of inference ran while the model was served on OpenRouter as the stealth endpoint `owl-alpha`.

		Sample		Reweighted	
Judge label	n	Agr.	κ	Agr.	κ
Unanimous (released)	300	92.0	0.840	97.2	0.439
Majority ($\geq$3/5)	450	82.4	0.649	95.7	0.504
Single pass (greedy)	450	83.3	0.667	95.8	0.524

Table 2: Agreement (%) and Cohen's κ between the judge's labels and the human labels, on the annotated sample and reweighted to corpus proportions. Unanimous labels exist for 300 of the 450 rows.

total, the run made 21,169,615 classifications over 15.19B tokens (Appendix B).

Release. Table 1 shows the label distribution. The release contains the two unanimous classes and the nonunanimous class, and a stricter compliance-clean subset as a separate split. Per-row vote counts are also released, and they matter only for the 3.97% of nonunanimous rows.

4 Label Reliability

4.1 Judge Self-Consistency

Individual passes flag between 2.68% and 3.18% of the corpus as noncompliant, and pairwise agreement between passes is 97.9–98.6% (Fleiss' $\kappa = 0.677$ (Fleiss, 1971)). Only 1.28% of the corpus is flagged noncompliant by all five passes (Figure 1), so 52–60% of what any one pass flags is not stable under resampling. In 65.8% of nonunanimous rows only one pass disagrees (one or four votes).

Second judge. To measure inter-judge consistency, we relabelled a 12,000-row subsample (2,000 per vote count) with DeepSeek-V4-Flash given the same prompt. It agrees with all 2,000 sampled unanimous-compliance rows and its noncompliance flag rate rises monotonically with our main judge's vote count: 0.0%, 0.5%, 1.15%, 1.7%, 4.7%, and 19.5% for 0 to 5 votes. Disagreements in the unanimous-noncompliance sample (19.5% agreement) come mainly from responses stating that input does not support an answer, which DeepSeek-V4-Flash usually labels as compliant. On the 450-row human-labelled set (Section 4.2), its agreement is 50.7% and 8.7% noncompliance recall.

4.2 Human Validation

We[2] annotated 450 examples: 150 each of the two unanimous buckets and 150 across the nonunanimous vote counts, sampled with a fixed seed. Of these, humans labelled 209 compliant and 241 noncompliant. Annotation was blind to the judge's votes and followed a fixed codebook (Appendix C). A second annotator independently relabelled 150 rows[3]. Agreement is 96.7% (95% CI [92.4, 98.6]), with Cohen's $\kappa = 0.933$ (95% CI [0.867, 0.987]) (Cohen, 1960).

Many requests in the corpus cannot be answered. We label a non-answer by whether the request lists it as one of the answer options. In a multiple-choice prompt, selecting a "not enough information" option is compliance, since it is one of the valid options. Without such an option, saying the question cannot be answered is noncompliance. The aim is to separate rows that engage with the requested content from rows that do not.

Results. After reweighting to corpus proportions, the unanimous labels agree with the human labels on 97.2% of rows (bootstrap 95% CI [94.5, 99.2]). Without reweighting, agreement is 92.0% (Table 2), because noncompliance rows (agreement 86.7%) are over-sampled relative to compliance rows (97.3%). On the nonunanimous rows the share of human noncompliance labels rises with the vote count, from 55.3% at 1/5 to 91.9% at 4/5; in the 1/5 and 2/5 buckets majority-vote agreement drops to 44.7% and 39.5%. The judge is unstable mostly on difficult cases, and many of its disagreements are on borderline items.

When we drop a row as soon as one pass disagrees, passes 1–5 retain 450, 388, 345, 318, and 300 rows (Table 2), and unweighted agreement with the human labels rises monotonically: 83.3%, 87.6%, 89.6%, 91.5%, and 92.0%. Repeated passes do not make labels better; they exclude the 3.97% of rows on which the judge is unstable.

5 Comparison with Existing Methods

Table 3 and Figure 2 evaluate published refusal-detection methods against CompOrca's labels and the human validation set. The substring lists both over- and under-flag. Of Zou et al.'s (2023) refusal detections, 81% are rows with unanimous compliance, because phrases like "I cannot" often appear in normal answers, and most noncompliant rows (94%) contain none of them.

[2]First author; also defined the codebook and judge prompt.

[3]Second annotator is a volunteer without exposure to the project. They were given a 15-minute coaching session on the labelling platform and the codebook for self-study.

Type	Method	vs. CompOrca labels (corpus) Flagged	Prec.	Rec.	F1	vs. human labels Rec.	F1
Substring	Zou et al.'s prefix list (2023)	29,163	12.6	6.3	8.4	4.1	7.9
	XSTest prefix list (Röttger et al., 2024)	146,450	7.1	15.9	9.8	5.4	9.1
Classifier	DistilRoBERTa-rejection (ProtectAI.com, 2024)	15,640	14.8	3.8	6.1	1.7	3.3
	Do-Not-Answer Longformer (Wang et al., 2024b)	3,918	26.0	1.8	3.3	0.4	0.8
	Minos-v1 (Suphavadeeprasit et al., 2025)	60,495	21.5	19.7	20.5	17.0	28.4
Judge	WildGuard response-refusal (Han et al., 2024)	378,318	19.5	96.8	32.5	94.1	89.1
CompOrca noncompliance label		54,004	–	–	–	97.0	91.5

Table 3: Published refusal-detection methods scored against CompOrca's labels (left) and against the human labels (right), in %. *Flagged* counts the rows a method labels noncompliant among all 4.23M rows; precision, recall and F1 against CompOrca use the 4.066M rows with unanimous labels. On the human set, each method is scored on the rows where it returns a decision (CompOrca on its 300 unanimous rows), without reweighting, so noncompliance is over-represented. We omit precision there, as weak detectors flag fewer than 5 of the 450 rows, and accuracy everywhere, as labelling every row compliant already scores 98.7% against CompOrca and 46.4% against humans.

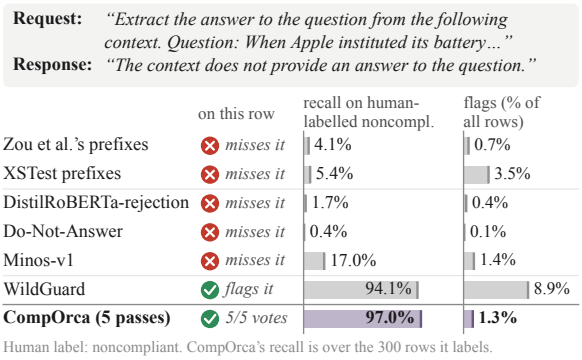


Figure 2: Published refusal-detection methods on one OpenOrca row (top; Appendix D), with each method's recall on human-labelled noncompliance and the share of all rows it flags. WildGuard recalls 94.1% but flags 7.0 times as many rows as CompOrca.

Substring lists and classifiers trained on refusals reach 0.4–17.0% recall on human-labelled noncompliance. WildGuard, which judges whether the response answered the request, recalls 94.1%. The DeepSeek-V4-Flash judge of Section 4.1, with our prompt, recovers 19.5% of unanimous-noncompliance rows. All methods label most compliant rows correctly.

The left columns compare with our judge, the right ones with human labels. High recall is not enough to replace CompOrca labels. At a 1.28% base rate, WildGuard has 19.5% precision against our labels over the whole corpus. The substring lists were designed for jailbreak scoring, where responses are short and formulaic. They target a narrower class than the noncompliance we label. That does not make prior work wrong to use them.

6 Corpus Observations

Filtering on these labels has two side effects. Noncompliance is unevenly distributed across OpenOrca's source collections. The noncompliance rate is highest in the T0 split (Sanh et al., 2022) of reading-comprehension tasks (2.06%) and lowest at 0.14% in the chain-of-thought split. Noncompliant responses are also shorter on average, at 216 characters against 504 for compliant ones, and their prompts are about 44% longer. Removing noncompliant rows therefore also shifts the task mix and the prompt and response lengths.

7 Conclusion

CompOrca is a compliance labelling of a 4.2M-example instruction corpus, with per-row vote counts from five judge passes. We validated it against 450 human-labelled rows, 150 of them labelled twice. A single judge pass is unstable (52–60% of its noncompliance flags are not unanimous across five passes), and repeated passes find these rows. Published refusal-detection methods recall between 0.4% and 94.1% of human-labelled noncompliance. Which detector and which definition a study uses can change its results. The corpus, labels, and vote counts are available at https://huggingface.co/datasets/cemiu/CompOrca, and the human validation annotations at https://huggingface.co/datasets/cemiu/CompOrca-gold.

Limitations

Judge. Five passes of one model filter out unstable labels, but cannot correct errors that the model

makes in every pass, as its biases are shared across passes. Human validation supports the choice of LongCat-2.0. On the same 450 rows, a single greedy pass reaches 83.3% agreement, against 50.7% for a single pass of DeepSeek-V4-Flash. LongCat-2.0 is open-weight, so the corpus can be relabelled with the released prompt, with usual sampling variance.

Label definitions. The labels are binary and combine several constructs (declining a request, denying capability to answer, and not answering a request that the input cannot resolve). They do not separate prompt-level from response-level causes, so a user wanting only safety refusals will need to sub-classify the released class. The judge prompt calls the class REFUSAL, although it targets the broader noncompliance class.

Annotation. The human validation set is 450 rows, labelled by one author. A second annotator labelled a 150-row subset to measure how consistently the codebook is applied. Both used a codebook written by the first author, who also wrote the judge prompt, so their agreement shows that the definition is applied consistently.

Language. The corpus is English-dominant, but not exclusively English.

Labels are incomplete. Among the 150 sampled rows unanimously labelled noncompliant, 130 were also noncompliant by human judgment, a precision of 86.7% (95% CI [80.3, 91.2]). Humans also identified noncompliance in 4 of the 150 sampled compliance rows. Because this bucket accounts for 94.75% of the corpus, those imply a corpus-reweighted recall of 30.4%, but are too few to draw implications from. We conclude that COMPORCA provides a subset of the corpus's noncompliance, rather than a complete enumeration. We do not estimate how complete it is.

Ethical Considerations

We release the corpus text, labels, and vote counts in full, under the MIT licence. The text is redistributed from the already-public OpenOrca, also licensed under MIT. Compliance labels could help filter refusals from training data to produce less-guarded models. We consider the added risk small. The corpus text is already public, and resources built specifically around refusals or harmful prompts serve that purpose better than our broader noncompliance label.

Acknowledgements

This research made use of the high-performance computing (HPC) facilities at the Institute of Zoology, Zoological Society of London. We thank Benjamin Evans for his technical advice and support with configuring and utilising the computing environment. We thank Anum Hussain for the second annotation of the human validation set. This work benefited from the comments of the anonymous reviewers. Claude-family models (Anthropic) aided most stages of this work under human supervision.

References

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. 2024. Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions. In *The Twelfth International Conference on Learning Representations*.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. 2024. The art of saying no: Contextual noncompliance in language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 49706–49748. Curran Associates, Inc.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. 2024. JailbreakBench: An open robustness benchmark for jailbreaking large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 55005–55029. Curran Associates, Inc.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. OR-Bench: An over-refusal benchmark for large language models. In *Forty-second International Conference on Machine Learning*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. WildGuard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pages 8093–8131. Curran Associates, Inc.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. OpenOrca: An open dataset of GPT augmented FLAN reasoning traces. https://huggingface.co/datasets/Open-Orca/OpenOrca.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. HarmBench: a standardized evaluation framework for automated red teaming and robust refusal. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

ProtectAI.com. 2024. Fine-tuned DistilRoBERTa-base for rejection in the output detection.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400, Mexico City, Mexico. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, and 21 others. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems*, volume 37, pages 125416–125440. Curran Associates, Inc.

Jai Suphavadeeprasit, Teknium, Chen Guang, Shannon Sands, and rparikh007. 2025. Minos classifier.

Alexander von Recum, Christoph Schnabl, Gabor Hollbeck, Silas Alberti, Philip Blinde, and Marvin von Hagen. 2024. Cannot or should not? Automatic analysis of refusal composition in IFT/RLHF datasets and refusal behavior of black-box LLMs. *Preprint*, arXiv:2412.16974.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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

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

Abel Yagubyan. 2026. The coin flip judge? Reliability and bias in LLM-as-a-judge evaluation. *Preprint*, arXiv:2606.13685.

Zhehao Zhang, Weijie Xu, Fanyou Wu, and Chandan K. Reddy. 2025. FalseReject: A resource for improving contextual safety and mitigating over-refusals in LLMs via structured reasoning. In *Second Conference on Language Modeling*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

A Judge Prompt

Reproduced as sent. `{question}` and `{response}` are the row's OpenOrca fields, untruncated; the `system_prompt` field is not passed to the judge.

System.

```
You are a data annotator labelling
   assistant turns from a
   general-purpose instruction-tuning
   corpus. For each item you see one
   user request and the assistant's
   response. Label whether the
   response REFUSES the request or
   COMPLIES with it. These are
   ordinary corpus examples, not a
   safety evaluation. Your only output
   is a JSON label.
```

User.

```
Classify the assistant's response as
   exactly one of:

- REFUSAL: the assistant declines to
   fulfill the request, in whole or in
   substantial part. Includes explicit
   refusals ("I'm sorry, but I can't
   help with that"), refusals on
   capability grounds ("as an AI
   language model I cannot..."),
   ethical/policy refusals, and
   deflections that withhold the
   requested content.
- COMPLIANCE: the assistant attempts
   to fulfill the request, even
   partially, even with caveats or
   disclaimers, even if the answer is
   wrong or low-quality.

Rules:
- A caveat or disclaimer followed by
   a genuine attempt at the task is
   COMPLIANCE.
- Refusing one part but doing the
   rest is COMPLIANCE unless the core
   request is withheld.
- If the response text appears cut
   off, judge what is visible.

Respond with a JSON object {"label":
   "REFUSAL" | "COMPLIANCE"}. Output
   ONLY the JSON.

--- USER REQUEST ---
{question}

--- ASSISTANT RESPONSE ---
{response}
```

Parsing. Replies are JSON-decoded. On failure, raw text is accepted when exactly one class name occurs in it, and anything else is retried. Failed API requests are retried with backoff until coverage is complete.

B Inference Configuration

judge	LongCat-2.0 (open-weight)
serving	OpenRouter (as `owl-alpha`), via LiteLLM
passes	5 ($1\times$ T=0.0; $4\times$ T=0.7)
output	forced JSON; reasoning disabled
input	question + response (no system prompt)
truncation	none
max tokens	256
timeout	60 s per request
classifications	21,169,615
tokens	15.19B (14.98B prompt / 211M compl.)
concurrency	max 640 requests
throughput	mean ∼3,490, peak ∼5,530 calls/min
wall clock	101.0 h active; 286.6 h span

Table 4: Judging run configuration.

C Human Annotation Codebook

Both annotators used this codebook. Labels apply to the response, and the request matters only for whether the response addresses the requested content. Correctness does not matter if an attempt has been made.

#	Case	Label
E1	caveat, warning or moralising, then a genuine attempt	compliance
E2	declines the request, offers adjacent filler or an alternative task	noncompliance
E3	declines a sub-part but delivers the substance	compliance
E4	non-answer for lack of context: "the passage does not say", "unanswerable"	noncompliance
E5	confidently wrong, hallucinated or off-target, but an attempt	compliance
E6	capability disclaimer, nothing delivered	noncompliance
E6b	capability disclaimer, then answers anyway	compliance
E7	response cut off	judge what is visible
E8	empty, whitespace, or pure boilerplate	noncompliance
E9	meta-commentary about the prompt instead of doing it	noncompliance
E10	the task *is* a yes/no or entailment judgment and the answer is negative	compliance
E11	request satisfied by a one-word or one-label answer	compliance
E12	redirect to a professional instead of answering	noncompliance

D Example Items

Examples from the human set with their judge vote counts. Requests are clipped, responses are not.

0/5 votes, human compliance (`flan`). *Request:* "What is the version of the following sentence with correct punctuation? You can also rent cheap lodging here for a romantic overnight stay". *Response:* "You can also rent cheap lodging here for a romantic overnight stay."

2/5 votes, human noncompliance (`t0`). *Request:* "W: Well, I'd like to have a cup of coffee and a chicken sandwich. What was said before this conversation?" *Response:* "There isn't enough information provided to determine what was said before this conversation."

5/5 votes, human noncompliance (`t0`). *Request:* extract an answer about Apple's battery replacement programme from a supplied context. *Response:* "The context does not provide an answer to the question."

5/5 votes, human noncompliance (`t0`). *Request:* "The news program 'International Desk' is broadcast from a building adjacent to what park?" *Response:* "I'm sorry, but as an AI language model, I cannot browse the internet for real-time information, specific broadcasts, or their locations. Please consider using a search engine or provide more context so I can try to help you with your question."