Refining Inverse Constitutional AI for Dataset Validation under the EU AI Act

Carl-Leander Henneking[*]
Cornell University
Ithaca, NY 14850, USA
ch2273@cornell.edu

Claas Beger[*]
Cornell University
Ithaca, NY 14850, USA
cbb89@cornell.edu

Abstract

The recent proposal of the EU AI Act sets ambitious requirements for regulating state-of-the-art AI models. In particular, Article 10(2)(f–g) mandates the examination and application of appropriate measures to ensure that datasets are assessed with respect to potential biases. Traditional alignment methods for Large Language Models (LLMs), such as Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO), rely on pairwise preferences encoding implicit principles, which are inherently mismatched with this explicit regulatory framework. In contrast, Constitutional AI (CAI) offers a transparent, rule-based approach to alignment, making it a natural fit for bias detection and governance. Building on this foundation, we refine the Inverse Constitutional AI (ICAI) algorithm by enhancing principle generation, clustering, and embedding, thereby enabling more systematic extraction of constitutions from preference datasets. Finally, we outline a potential framework for employing ICAI as a tool for validating datasets in accordance with Article 10 of the EU AI Act, offering a pathway toward alignment methods that are both technically robust and regulatorily compliant.

1 Introduction

Aligning pre-trained Large Language Models (LLMs) to human preferences commonly relies on pairwise preference data, e.g., Reinforcement Learning from Human Feedback (RLHF) with a learned reward model, or Direct Preference Optimization (DPO) which encodes preferences directly in the fine-tuning loss. Both are effective but leave the underlying human (preference) principles largely implicit.

Constitutional AI (CAI) [Bai et al., 2022] instead uses an explicit set of principles ("a constitution") to guide generation via self-critique. This improves interpretability relative to RLHF/DPO, though it still inherits biases present in data Sorensen et al. [2024], Chakraborty et al. [2024]; explicit goals also align with long-standing arguments for rule-guided AI Ji et al. [2025], Gabriel [2020], Tennant et al. [2025].

Building on this idea, Findeis et al. [2024] proposed Inverse Constitutional AI (ICAI), which infers a constitution from preference datasets via prompting, clustering, and LLM-as-a-judge feedback, surfacing latent values and enabling prompting-based steering.

Our work aims to improve the shortcomings that we identified within the ICAI algorithm to produce constitutional principles that represent dataset preferences more accurately and show how such functionality could be employed for regulatory dataset validation. For this purpose, we address different

[*]Equal contribution.

Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

weaknesses regarding generalizability and sampling. We also experiment with utilizing various embeddings to perform grouping of related preference pairs prior to the initial principle generation. We evaluate our changes in three settings, ranging from synthetic to semi-synthetic and realistic data, and report improvements over the baseline ICAI algorithm in all three.

In parallel, we ground the extraction and use of these principles in the EU AI Act's data- and governance-centric requirements for high-risk AI. Article 10 mandates that training, validation, and testing datasets be subject to data governance practices and specifically, that providers examine and mitigate bias in their data [European Parliament, 2024]. We therefore treat the inferred constitution as an auditable artifact that surfaces dataset choices, bias analyses, and traceable decision rationales of the human annotators. The produced set of principles can then also be handed off to an external auditor for third-party approval, adding governance to model alignment. The interplay of all actors is outlined in Figure 1.

Overall, our work aims to tackle the following research questions:
(1) How can we improve the existing constitutional extraction method on pair-wise preference human datasets?
(2) What are the key use cases and implications that such representative constitutions provide us with?
(3) How may Inverse Constitutional Generation be employed to enable (external) dataset validation in line with regulatory requirements?

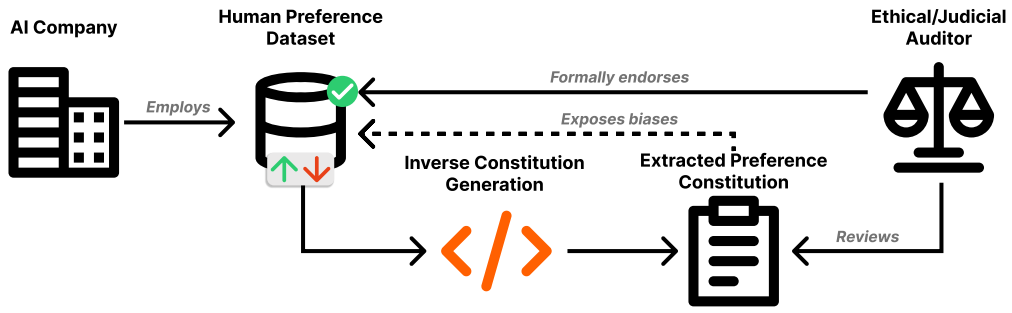

Figure 1: Overview of the regulatable inverse-constitutional pipeline: an AI company wants to use a human preferences dataset for model alignment, inverse constitution generation infers an extracted preference constitution, which exposes implicit biases in the dataset, and an ethical/judicial auditor reviews said constitution for irregularities/misalignment, and formally endorses or rejects the usage of the dataset.

2 Methodology

Findeis et al. [2024] utilize a combination of prompting, clustering, and voting to extract a representative set of rules. The ICAI algorithm takes a pairwise preference dataset as input and performs the following five steps to derive a constitution:

(1) Initial candidate generation: Using a pair of chosen and rejected replies, the algorithm prompts an LLM to generate a set of candidate principles that reflect the preference rating.
(2) Clustering: All principles are embedded and clustered using KMeans.
(3) Subsampling: A random principle is chosen from each cluster to represent a candidate for the final constitution.
(4) Testing: Using an LLM, each candidate principle is evaluated against every pair in the preference dataset. Ratings are collected on whether a candidate is in favor/against/not applicable to each preference pair.
(5) Filtering: Finally, based on a set of rules/thresholds, the list of candidate principles is reduced to the final constitution.

We first analyze the implementation of the ICAI algorithm and identify possible improvements in multiple steps. From the architecture, it is clear that the pipeline heavily relies on the ability to generate representative and generalizable principles from a single preference pair, which is rooted

in step one. This essentially constitutes a bottleneck since rules are not changed in later stages; they are only filtered. This is problematic since our test runs of the pipeline revealed that the principles were often heavily tailored to the specific sample they were derived from. For example, "Select the response that engages with the user's interest in Sahawiq." was one of the identified principles (Sahawiq being a hot sauce). Because of such candidates, we believe limiting effects in later processing steps reduce overall effectiveness.

Further, in step three, a principle is randomly selected from the resulting clusters. As described above, some principles are highly specific and should not be considered; however, principles of such sub-par quality may still be selected in this step and misrepresent the cluster contents. Overall, we believe that the ICAI architecture suffers from the propagation of weak, non-representative principles into late processing steps, which replace promising alternatives. Additionally, the model heavily relies on its initial principles, which require drafting a large number to ensure stable results. This is a bottleneck concerning result quality and incurs unnecessary performance overhead.

Our proposed improvements address these issues. First, we modify the principle generation prompt to nudge the LLM to generate a set of more generalizable principles. Secondly, during principle sub-sampling, our approach selects the principle from the cluster closest to its centroid (as measured by the cosine similarity of embeddings). These two joint improvements result in our first improved version.

We also explore further enhancements aimed at improving the quality of the initial principal candidates. This step is based on the assumption that the rule we seek to extract is reflected in the difference between the two responses. This difference can potentially be represented as a tensor in latent space, provided we have an appropriate representation of the responses. We further believe that different rules may target different aspects of alignment, including style ("Speak in a kind and concise manner"), content ("Do not give information on illegal activities"), and sentiment ("Do not agree on problematic claims and try to resolve them"). In line with these assumptions, we employ different embedding models to produce the needed representation of pairs and consequently their difference. In particular, we employ all-mpnet-base-v2 [Henderson et al., 2019] for content, Wegmann et al. [2022] for content-independent style, and Sanh et al. [2020] for sentiment classification.
Using these models, we arrive at three embedding difference maps and use KMeans to identify clusters, which we hope will feature a joint principle. To consolidate the best candidates from every map, we compute a combined score of inter- and intra-cluster distance and only extract the top k clusters over all three. We also ensure no overlaps between nodes in different chosen clusters. Since some clusters tend to incorporate a large number of nodes, we refine the extraction further by focusing on node triplets within a cluster. We reuse the distance metrics from the cluster selection to choose representative triplets. In this step, we aim to ensure a rule majority in cluster representations, which we validated on a synthetic dataset, where the top 5 triplets consistently achieved a purity of at least 66.7%. Since the underlying data is synthetic, this value can be computed by inspecting the ground truth principle used to generate the preference pair. These triplets are then used in a joint prompt to generate initial candidates and complete the second improved version of the pipeline. We illustrate this in Figure 2.

We evaluate the baseline and both improved versions in three settings:
(1) A synthetic setting where we explicitly control for the preferences elicited in the pairwise preference data.
(2) A semi-synthetic setting derived from pairs that elicit significant differences in scored preference.
(3) A realistic/original setting where we sample principles from a pairwise preference dataset without intervention.

The synthetic dataset is generated by choosing five constitutional principles from the original CAI paper [Bai et al., 2022]. Next, we sample 150 pairs of chosen and rejected outputs from Anthropic's HH dataset's harmlessness subset. We keep the rejected output and prompt GPT-4o to re-write the output based on one of the constitutional principles, obtaining a synthetic chosen output. The chosen principles include an explicit rewriting prompt, which we utilize for this purpose. Each principle is represented equally among the train set (100) and test set (50). The process is visually described in Figure 3.

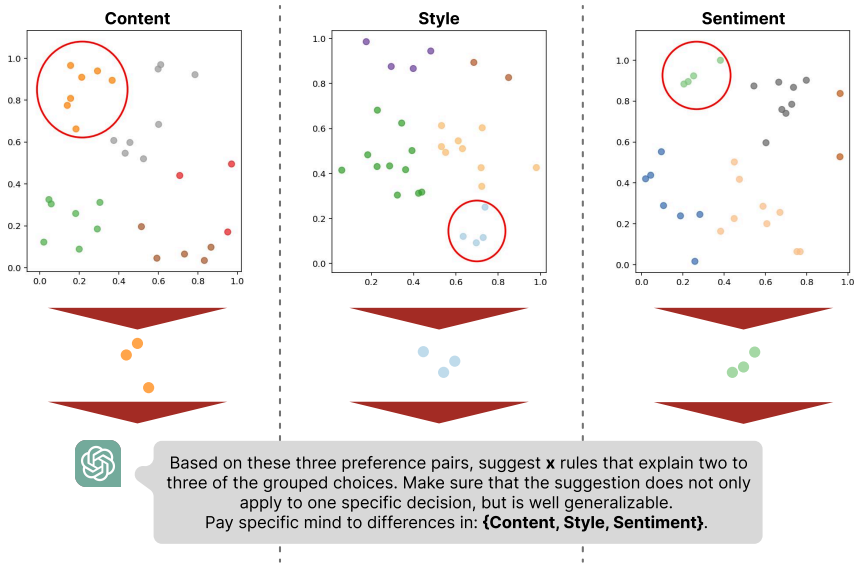

Figure 2: We apply KMeans on the difference between the embeddings of the preference pairs. After estimating cluster potential, we extract representative node triplets, which are inserted in one joint principle generation prompt. Prompts are executed separately for each dimension.

To obtain a semi-synthetic dataset, we choose to filter a subsample of data points from the UltraFeedback [Cui et al., 2024] dataset, which conveniently includes ratings (ranging between 1 and 5) for the rejected and chosen outputs. We filter to only include samples where the rating difference between chosen and rejected is at least two and then, using weighted sampling based on the distances, sample 1000 data points for the train set and 500 for the test set. We display the corresponding process in Figure 4. We apply this filtering since outputs with a similar rating according to the annotator's preference do not carry much useful information for extracting an underlying rule according to the original ICAI algorithm and our improved versions. On the contrary, very slight differences may lead to the extraction of unwanted or incorrect rules. However, we believe this may change when the scores are an active part of the extraction algorithm, which we investigate in section 3.

Lastly, our original/realistic setting consists of a subset of the HH-Harmlessness dataset of 1000 pairs for the train set and 500 for the test set. We note that the HH dataset includes very subtle differences between the replies, which makes it a challenging target, so much so that even humans struggle to infer preferences when the labels are removed.

For our main evaluation, we use an LLM-as-a-judge approach and iterate over all samples in the test set, prompt the LLM with the chosen and rejected sample, include all constitutional values, and ask the model to choose the output that aligns best with the provided principles. We believe that regenerating the preferences is a natural measure of the quality of our approximation of the underlying latent preference set. We adjust for ordering bias by gathering scores in both orders for each pair and then averaging the two scores.

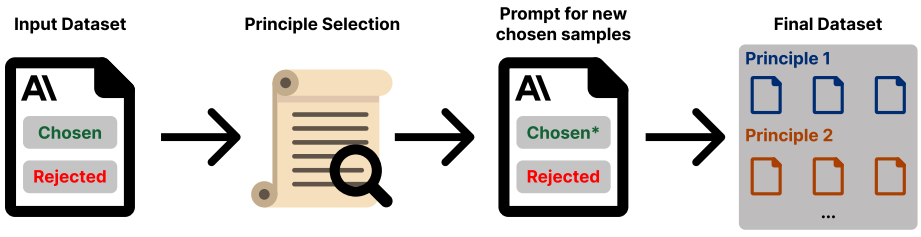

Figure 3: Dataset generation process for the synthetic dataset. Colors in the final dataset represent the samples that were generated using the same principle.

Figure 4: Dataset generation process for the semi-synthetic dataset. Different shades of purple in the final dataset indicate different deltas between chosen and rejected ratings.

3 Experimental Results

3.1 Regenerating Preferences

Our experimental results on the regeneration of preferences in the three different settings are outlined in Table 1. We include an orthogonal approach (unrelated constitutional values, such as prefer cats over dogs) to show reference win rates and control for inherent model bias. For this purpose, we adapt the orthogonal constitution used in Findeis et al. [2024]. The orthogonal constitution is held constant for all three datasets. Specifically, for the latter two datasets, we see accuracies of around 50%, which we would expect, as this represents a random choice. The accuracy for the synthetic dataset is slightly higher; however, accuracies for all other approaches outperform it significantly. We hypothesize that the harmful character of the rejected prompts makes it difficult for the model to choose them despite being instructed to abandon judgment beyond the constitutional rules.

Dataset	Baseline	Orthogonal	Improved 1	Improved 2
Synthetic	92.00%	62.50%	**94.00%**	93.00%
Semi-Synthetic	71.20%	46.95%	73.80%	**76.20%**
Original	60.65%	56.60%	60.55%	**60.75%**

Table 1: We report test-set accuracy as measured by how many sample preferences can be regenerated using the extracted constitution. Our results show slight improvements in the synthetic and realistic setting and significant improvements in the semi-synthetic setting.

3.2 Scored Preference Datasets

Some alignment datasets, such as UltraFeedback, also include numerical ratings of both outputs of each preference pair. We hypothesize that these ratings could be employed in our pipeline to further improve the extraction of principles, as they contain valuable information about the extent to which a rule expresses a preference for one reply over the other. We slightly modify our first improved algorithm to test our hypothesis and include the UltraFeedback ratings in the initial principle generation prompt. We consciously do not adjust the rest of the algorithm to gain a sense of the performance improvement the model gains by employing the scores in the prompting step. Interestingly, this simple modification led to an accuracy of 76.80%, or an improvement of 3%, even slightly higher than our Improved version 2. We thus conclude that preference ratings bear significant potential for proper constitution extraction. Even without functional adaption of the algorithm, the model is able to extract valuable information. Further optimization in this direction will likely enable us to extract an even more representative constitution in future work.

4 Policy and Regulatory Implications

The refinement of ICAI carries broader implications for how preference-based alignment datasets can be validated under the EU AI Act. Article 10(2)(f–g) explicitly requires providers of high-risk AI systems to examine training, validation, and test data for potential biases and to apply appropriate measures for detection, prevention, and mitigation [European Parliament, 2024]. While these obligations are conceptually clear, their practical fulfillment is far less straightforward in the

case of large-scale pairwise preference datasets. Such data encode subtle, implicit norms that only emerge statistically across many samples, making bias both difficult to identify and opaque to external scrutiny [Yan et al., 2025, Ferrara et al., 2024].

ICAI offers a proof-of-concept on how to make these latent structures auditable. By surfacing constitutional principles from raw preference data, it produces a set of explicit rules that can act as regulatory artifacts. These principles help translate implicit alignment signals into an interpretable form, enabling providers to show regulators and auditors not only that bias assessments have been attempted, but also what value trade-offs are implicitly embedded in the data. This transparency aligns with the Act's requirements for traceability and documentation [Gebru et al., 2021, Nexla, 2023], while reducing the need to disclose raw, potentially sensitive datasets to third parties.

From a governance perspective, ICAI introduces an intermediate representation that facilitates third-party auditing. Instead of granting external auditors direct access to proprietary or privacy-sensitive training data, providers can hand over extracted constitutions. Auditors can then evaluate whether these rules are consistent with fundamental rights obligations, detect systematic biases, and endorse or recommend corrective measures. This workflow (see Figure 1) provides a practicable compromise between the confidentiality concerns of providers and the accountability demands of regulators.

Policy-wise, ICAI exemplifies a shift from outcome-only regulation (evaluating model behavior post hoc) to dataset-centric oversight (examining the value-laden choices encoded in preference data prior to model training). If further developed, such methods could inform emerging standards for dataset governance [ISO/IEC, 2021], offering regulators a repeatable mechanism to benchmark fairness and representativity in subjective alignment datasets. More broadly, they highlight the necessity of novel interpretability tools if Article 10's bias auditing requirements are to be meaningfully enforced in practice.

That said, limitations remain. Even our refined version of ICAI struggles with realistic preference datasets where annotator rationales are highly context-dependent and principles can fragment into overly specific rules. Over-reliance on automatically inferred constitutions could thus give a false sense of compliance. We argue that ICAI should be viewed as a complement, not a replacement, for established bias detection and governance practices (e.g., statistical subgroup analysis, human-in-the-loop audits). Future policy frameworks may need to explicitly recognize such hybrid approaches, where algorithmic tools surface candidate value structures but human auditors retain final responsibility for bias evaluation.

5 Related Work

Although the key relevant bodies of work are outlined in more detail in section 1, we provide further details on adjacent works below.

Alignment Approaches: Inclusion of explicit rules has been adapted in several alignment techniques. Klingefjord et al. [2024] include the generation of relevant alignment targets directly by a process they refer to as Moral Graph Elicitation, which includes an explicit interviewing and reconciliation stage with annotators. Glaese et al. [2022] set out to steer the underlying applied preference exhibited by annotators through providing pre-defined rules, which was used in combination with a traditional RLHF approach. Partially inspired by this, Anthropic proposed CAI [Bai et al., 2022], an alignment approach that instills constitutional principles into model outputs. Findeis et al. [2024] presented the Inverse CAI (ICAI) algorithm to extract a set of principles from a pairwise preference alignment dataset. Kostolansky [2024] proposes a similar approach to ICAI that also leverages clustering of embeddings and prompting-based steps.

Beyond the approaches centered around constitutional values for alignment / extracting these values, alternative alignment approaches exist that are tangent to constitutional ones. One is Dromedary [Sun et al., 2023], a self-alignment approach that relies on 16 guiding principles throughout their pipeline. Their solution aims to minimize the number of required human annotations. Gao et al. [2024] propose a framework, PRELUDE, which uses user edits to infer latent preferences, enabling alignment without costly fine-tuning.

Value-based Methods: Value-based approaches are not limited to language model alignment. Hosking et al. [2024] argue that human preferences are inherently multi-dimensional rather than one-dimensional. Their experiments reveal that multiple factors contribute to these preferences,

with the refusal to answer unreasonable requests emerging as the most crucial. Similarly, Ji et al. [2023] introduce the BeaverTails dataset, a 330k-entry QA dataset annotated for helpfulness and harmlessness. The harmlessness labels are further divided into 14 distinct categories of harmful values, providing a more nuanced framework for evaluating responses.

Regulation and Governance: Beyond technical alignment methods, a growing body of work emphasizes dataset governance and regulatory compliance. The EU AI Act highlights the need for transparency, traceability, and bias mitigation in training data [European Parliament, 2024]. Frameworks such as Datasheets for Datasets [Gebru et al., 2021] and data lineage tooling [Nexla, 2023] provide templates for documenting dataset composition and provenance, while ISO/IEC TR 24027:2021 offers guidance on identifying and managing bias in AI systems [ISO/IEC, 2021]. These governance-oriented approaches intersect with constitutional alignment methods: both aim to surface implicit normative choices in a form that is auditable and accountable. Recent work in industry settings also illustrates how fairness audits and bias detection are being integrated into recommender pipelines [Yan et al., 2025], while bias-aware ranking algorithms show promise in addressing annotation-driven skew in pairwise preference data [Ferrara et al., 2024]. Our proposed refinement of ICAI builds on these efforts by positioning extracted constitutions as regulatory artifacts that could bridge the gap between technical alignment and legal oversight.

6 Conclusion

This work refined the Inverse Constitutional AI (ICAI) algorithm and evaluated its capacity to extract interpretable constitutional principles from pairwise preference datasets. Our empirical results suggest that our revised constitution extraction can improve preference regeneration across synthetic, semi-synthetic, and realistic datasets, while also producing human-readable artifacts that expose the implicit value trade-offs encoded in alignment data. These artifacts are not only useful for interpretability but also carry potential regulatory significance.

6.1 Discussion

The EU AI Act's Article 10(2)(f–g) requires providers of high-risk AI systems to examine training, validation, and test data for bias, and to apply appropriate mitigation measures. Meeting these obligations is particularly challenging for large-scale pairwise preference datasets, where normative patterns are deeply embedded and difficult to audit. ICAI provides a mechanism to translate these latent biases into explicit, auditable principles. In this sense, the approach contributes to the broader dataset governance requirements of the Act, including traceability, documentation, and representativity. By supplying external auditors with extracted constitutions instead of raw data, ICAI also helps balance confidentiality concerns with regulatory transparency.

Nonetheless, challenges remain. On the technical side, scalability bottlenecks and ambiguity in rule extraction persist. On the governance side, the value of extracted constitutions depends on standardized evaluation frameworks for auditors. Without agreed metrics and procedures, there is a risk of inconsistent assessments that may weaken compliance efforts. ICAI therefore, should not be seen as a standalone compliance tool, but as one layer in a hybrid governance stack that also includes statistical bias audits, subgroup analysis, and human-in-the-loop review.

6.2 Future Work

Future research can extend ICAI in several directions with regulatory utility in mind. First, preference scores and quality annotations could be more deeply integrated into the extraction process, potentially yielding more representative constitutions, as our findings in subsection 3.2 suggest. Second, adapting ICAI to datasets beyond binary pairwise preferences would expand its scope to demonstration-only or scalar-scored datasets, making it relevant across a broader range of high-risk AI systems. Third, applying ICAI to model-generated preferences could enable reverse-engineering of implicit constitutions, providing new tools for bias detection and accountability. Finally, stronger integration with existing governance standards, such as datasheets for datasets, ISO guidance on bias in AI, and lineage-tracking system, would support auditors in operationalizing Article 10. In summary, ICAI illustrates a pathway toward alignment methods that not only improve interpretability but also help operationalize the EU AI Act's data governance and bias mitigation requirements.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, and Mengdi Wang. Maxmin-rlhf: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with scaled ai feedback, 2024. URL https://arxiv.org/abs/2310.01377.

Council of the European Union European Parliament. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (AI Act). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689, 2024. Article 10: Data and Data Governance.

Alessandro Ferrara et al. Bias-aware ranking from pairwise comparisons. *Data Mining and Knowledge Discovery*, 38(2):512–534, 2024.

Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. Inverse constitutional ai: Compressing preferences into principles, 2024. URL https://arxiv.org/abs/2406.06560.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, September 2020. ISSN 1572-8641. doi: 10.1007/s11023-020-09539-2. URL http://dx.doi.org/10.1007/s11023-020-09539-2.

Ge Gao, Alexey Taymanov, Eduardo Salinas, Paul Mineiro, and Dipendra Misra. Aligning llm agents by learning latent preference from user edits, 2024. URL https://arxiv.org/abs/2404.15269.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, 2022. URL https://arxiv.org/abs/2209.14375.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. A repository of conversational datasets, 2019. URL https://arxiv.org/abs/1904.06472.

Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard, 2024. URL https://arxiv.org/abs/2309.16349.

ISO/IEC. Information technology — artificial intelligence (ai) — bias in ai systems and ai-aided decision making. Technical Report ISO/IEC TR 24027:2021, ISO/IEC, 2021.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 24678–24704. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4dbb61cb68671edc4ca3712d70083b9f-Paper-Datasets_and_Benchmarks.pdf.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Donghai Hong, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Kwan Yee Ng, Aidan O'Gara, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. Ai alignment: A comprehensive survey, 2025. URL https://arxiv.org/abs/2310.19852.

Oliver Klingefjord, Ryan Lowe, and Joe Edelman. What are human values, and how do we align ai to them?, 2024. URL https://arxiv.org/abs/2404.10636.

Timothy H. Kostolansky. Inverse Constitutional AI — hdl.handle.net. https://hdl.handle.net/1721.1/156804, 2024. [Accessed 22-09-2024].

Nexla. Data lineage tools—must-have features for genai. https://nexla.com/data-lineage-tools-for-genai/, 2023.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL https://arxiv.org/abs/1910.01108.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision, 2023. URL https://arxiv.org/abs/2305.03047.

Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Moral alignment for llm agents, 2025. URL https://arxiv.org/abs/2410.01639.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. Same author or just same topic? towards content-independent style representations, 2022. URL https://arxiv.org/abs/2204.04907.

Ming Yan et al. Fairness practices in industry: A case study in recommender systems. *arXiv preprint arXiv:2505.12345*, 2025.

A Embedding Model Ablation

Model Combination	**Synthetic**	**Semi-Synthetic (UF)**	**Realistic (HH)**
Content	0.910	0.690	0.600
Content + Style	0.910	**0.731**	0.595
Content + Sentiment	0.910	0.723	0.598
Style + Sentiment	0.910	0.712	0.577
All (Full System)	**0.920**	0.723	**0.608**
Improvement (All vs Best)	*+0.010*	*–0.008*	*+0.008*

Table 2: Average scores across three evaluation datasets: Synthetic, Semi-Synthetic (UltraFeedback), and Realistic (Helpfulness-Harmlessness). Content refers to `all-mpnet-base-v2`, Style to `AnnaWegmann/Style-Embedding`, and Sentiment to `distilbert-base-uncased-finetuned-sst-2-english`. "All" includes all three models. The improvement row compares the full system to the best-performing preceding combination in each setting.

To better understand the contribution of each embedding model in *Improvement Two*, we conduct an ablation study with all combinations of the three embeddings. The results are shown in Table 2. For this analysis, we employ GPT-4.1-nano for all generation steps to produce constitutions conditioned on each embedding combination. Subsequently, we use GPT-4o to annotate a randomly subsampled set of preference pairs from the test set, evaluating alignment with the generated constitutions.

Due to the random subsampling and the use of a smaller model (GPT-4.1-nano), absolute scores are generally lower than in the full evaluation reported earlier in Table 1. Nonetheless, relative differences remain informative. We find that using all three embeddings consistently improves over the base Content-only system across all settings. In the Synthetic and Realistic settings, the full system achieves the highest overall scores. However, in the Semi-Synthetic setting, the combination of Content and Style outperforms the full system, suggesting that the addition of Sentiment embeddings may introduce noise or redundancy in that context.

These findings align with prior observations: the Content embedding often provides the most substantial lift, which is frequently enhanced by the addition of Style. Sentiment, while generally contributing modest gains, can sometimes slightly reduce performance depending on the dataset. Overall, although differences between combinations are relatively small, leveraging all three embeddings proves consistently beneficial compared to using Content alone.

B Constitution Similarity

To evaluate how our generated constitutions compare to the ground truth constitution, we conduct an experiment employing LLM-as-a-judge to estimate similarity. For this purpose we first estimate similarity values between candidates and ground truth and create an optimal matching. After that, we aggregate the similarity scores. The results are shown in Table 3.

Approach	Similarity Score
Ground Truth	5.8
Baseline	5.0
Orthogonal	2.2
Improved 1	5.0
Improved 2	**5.4**

Table 3: We report mean similarity scores (1-10) between the ground truth constitution and the constitution generated using the synthetic dataset (the only setting in which a known ground truth constitution exists). Similarity scores are averaged across all constitutional values (n=5).

We also measure the ground truth constitution against itself. As the scoring scale ranges from 1 to 10, an average score of 5.8 is significantly lower than expected, raising concerns about the LLM's ability to correctly match and rate between constitutional values. This may also partially be due to a different output format ("choose the response that", and "Explain the prompt so that"), which the model is instructed to disregard for scoring. To correctly interpret the other results, we compare values to the ground truth score of 5.8 rather than the max score of 10. Our results show a significant improvement in our Improved 2 approach over the baseline. Again, we include the orthogonal constitution to show that other scores differ significantly.

C Resource usage for IMPROVEMENT 2.

We profiled the full embedding–clustering pipeline node on our HPC system, using a dual-socket Intel Xeon machine with 32 physical cores and 384 GB RAM; one NVIDIA A100-40GB GPU was allocated for the run. The training split of 19,874 dialogue pairs (39,748 utterances) was processed end-to-end in 0.49 GPU-hours ($\approx$ 29 min wall-clock). Peak GPU memory reached 3.7 GB (queried via `torch.cuda.max_memory_allocated` and confirmed with `nvidia-smi`), while peak system RAM during k-means clustering was 3.2 GB (monitored with memory_profiler). A fine-grained timing breakdown is as follows: MiniLM content embeddings (1.9 min), BERT style embeddings (5.6 min), two sentiment passes (9.2 min total), an additional MPNet pass for the multi-model variant (12.3 min), and CPU-side clustering/group selection (7.4 min). Thus, Improvement 2 may po-

tentially even fit on a single commodity GPU (¡ 4 GB VRAM) and completes in well under one hour—orders of magnitude for small to medium sized datasets, making it non-problematic to adopt.

D Sample Constitutions

In the following we showcase three representative constitution families produced by the second improved version.

Realistic (10 clauses)

1. Select the response that addresses legal implications.
2. Select the response that avoids unnecessary assumptions.
3. Select the response that discourages harmful actions.
4. Select the response that maintains a neutral tone.
5. Select the response that seeks clarification.
6. Select the response that maintains ethical standards.
7. Select the response that is more concise and direct.
8. Select the response that addresses potential harm concerns.
9. Select the response that encourages further dialogue and understanding.
10. Select the response that avoids encouraging illegal activities.

Semi-Synthetic (10 clauses)

1. Select the response that provides detailed step-by-step instructions.
2. Select the response that uses consistent and correct language.
3. Select the response that provides specific information.
4. Select the response with higher confidence level.
5. Select the response that includes detailed code examples.
6. Select the response that provides a practical solution.
7. Select the response that maintains a neutral and factual tone.
8. Select the response that uses engaging and vivid language.
9. Select the response that provides clear logical reasoning.
10. Select the response that directly addresses the question.

Synthetic (5 clauses)

1. Select the response that provides accurate and corrected information.
2. Select the response that avoids redundancy in advice.
3. Select the response that provides a more concise explanation.
4. Select the response that promotes respect and kindness.
5. Select the response that discourages illegal activities.