DEMOCRATIC ICAI: DEBATING OUR WAY TO STEERING PRINCIPLES FROM PREFERENCES

Kevin Kingslin, Anish Natekar, Ashutosh Ranjan, Vivek Srivastava,
Savita Bhat & Shirish Karande
TCS Research, India
{kevin.kingslin, anish.natekar, ashutosh.ranjan2,
srivastava.vivek2, savita.bhat, shirish.karande}@tcs.com

ABSTRACT

Preference-based alignment often struggles to capture the reasoning that underlies human judgments. Many evaluations rely on multiple interacting criteria, yet pairwise labels reveal only the final choice rather than the considerations that shape preferences. Inverse Constitutional AI (ICAI) improves interpretability in decision making by summarizing preferences into natural-language principles, but its single-pass explanations miss much of the nuance involved in complex decisions. We introduce *Democratic ICAI*, a novel approach that gathers multiple competing rationales through structured persona debate, offering a broader and more expressive account of the factors influencing each comparison. From these richer signals, we derive clearer and more comprehensive steering principles and use them to guide decision modeling through both LLM-based and decision-tree judges. Experiments on creative task preferences data show that *Democratic ICAI* uncovers deeper and more nuanced preference structure than ICAI alone, providing a more coherent and reliable foundation for aligned model behavior.

1 INTRODUCTION

Human preference data is central to aligning modern language models, yet many real-world judgment tasks demand far richer reasoning than what pairwise preference labels reveal. Evaluating creative writing, assessing alternative designs, comparing research explanations, or ranking artistic outputs often depends on qualitative and context-sensitive criteria that cannot be directly inferred from a single comparison. When annotators choose between two options, they integrate multiple considerations, but the recorded preference captures only the final decision and not the reasoning behind it. This gap poses fundamental challenges for preference-based alignment (Stiennon et al., 2020). These challenges emerge clearly in existing post-training pipelines. Reward models trained on preference pairs often latch onto superficial artifacts, which makes them vulnerable to reward hacking (Fu et al., 2025; Liu et al., 2024). Human preference datasets themselves contain structural biases, such as favoring assertive or belief-matching responses over truthful ones (Findeis et al., 2025), contributing to well-known issues like sycophancy (Sharma et al., 2023). Evaluation systems that rely on LLM-as-a-judge introduce additional instability: single-shot judgments fluctuate under changes in prompt phrasing, sampling strategy, or formatting (Gu et al., 2024; Schroeder & Wood-Doughty, 2024). These problems intensify in open-ended or creative settings where judgments emerge from interactions among coherence, tone, originality, and stylistic intent.

Inverse Constitutional AI (ICAI) provides a promising direction by transforming preference datasets into natural-language principles (Findeis et al., 2025). These constitutions can expose annotator biases and offer interpretable summaries of evaluative tendencies. However, ICAI relies on a single explanation per preference example, which limits the diversity of rationales it can uncover. Complex human judgments rarely hinge on a single consideration, and important reasoning may remain unexpressed in one-pass explanations.

To address this limitation, we introduce *Democratic ICAI* (Refer Figure 1). Instead of relying on a single justification, we elicit multiple, competing rationales through a structured debate among expert personas. Debate provides a mechanism for surfacing information that individual annota-

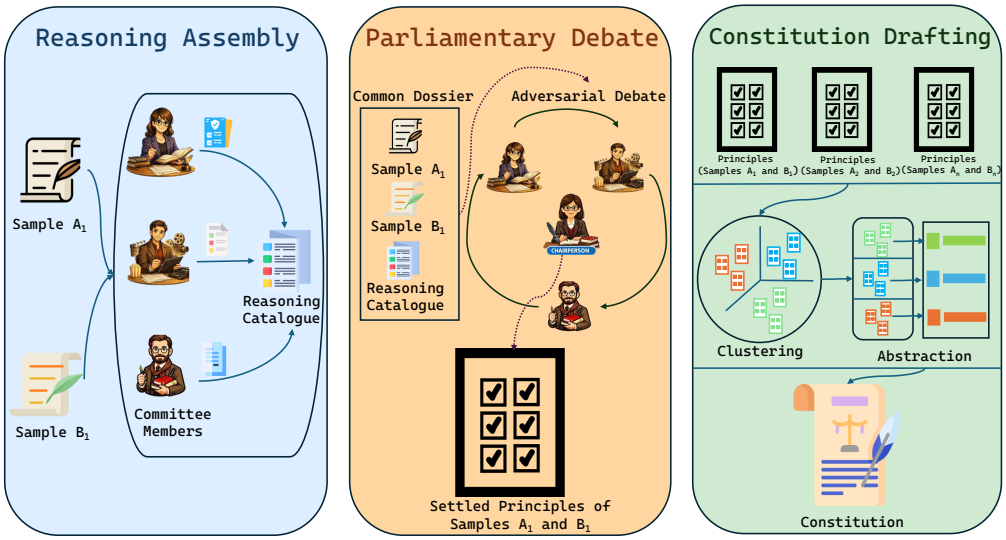


Figure 1: **Architecture of *Democratic ICAI*.** A committee of domain-expert personas first generates detailed rationales for each preference pair. These rationales are then subjected to an adversarial debate procedure, through which the evaluative principles relevant to each comparison are surfaced. Finally, the full collection of principles is clustered and abstracted to draft a concise, human-readable constitution.

tors or single explanations may fail to reveal (Irving et al., 2018). We then distill the resulting set of rationales into compact, human-readable *steering principles*. These principles provide explicit guidance for generation, training constraints, and transparent evaluation. Our approach complements the inductive process in ICAI and aligns with evidence from Constitutional AI showing that clearly articulated principles can effectively steer model behavior (Bai et al., 2022).

Our key contributions are as follows:

1. We introduce ***Democratic ICAI***, a novel framework for deriving steering principles from human preference data through structured multi-persona deliberation.
2. We discuss **constitution-guided decision modeling** and instantiate it with two complementary evaluators: an LLM-as-judge model and a decision-tree–based judge that operationalizes the learned principles.
3. We analyze the **limitations of the ICAI algorithm** presented in (Findeis et al., 2025) when applied to creative task preference data, and we demonstrate how our approach more effectively captures nuanced reasoning in **complex human evaluative tasks**.

2 Related Work

Preference Learning and Post-Training: Preference-based alignment methods such as RLHF and DPO improve model behavior by training on human comparisons, yet they still compress rich judgments into a single opaque scalar. This scalar easily reflects artifacts like response length or formatting, which enables reward hacking and distorts preference structure (Stiennon et al., 2020; Ouyang et al., 2022; Liu et al., 2024). Preference datasets also encode systematic biases, for example sycophancy, that aligned models can reproduce (Sharma et al., 2023). These limitations motivate methods that expose the underlying criteria behind preferences rather than relying only on aggregated scalar rewards. **LLM-as-a-Judge for Scalable Evaluation:** LLM judges offer scalable evaluation for open-ended tasks. MT-Bench and Chatbot Arena show strong agreement with human judgments but also reveal position effects, verbosity bias, and sensitivity to prompt phrasing that weaken single-shot reliability (Zheng et al., 2023). Rubric-based protocols such as G-Eval improve consistency by breaking decisions into explicit sub-criteria, although they still depend on one-pass reasoning and

can inherit hidden heuristics from the base model (Liu et al., 2023). A recent survey underscores the need for transparent, decomposed procedures rather than solitary overall scores (Gu et al., 2024).

Debate and Deliberative Supervision: Debate has been proposed to surface information that direct judging may miss, with theory suggesting that structured debate can reveal evidence otherwise inaccessible to single evaluators (Irving et al., 2018). However, empirical work on multi-agent debate reports collapse into echo-chambers or majority convergence when protocols lack safeguards, especially among similar models (Estornell & Liu, 2024; Liang et al., 2024). Effective debate therefore requires role design, turn-taking, and aggregation mechanisms that maintain diversity of reasoning.

Principle-Based Alignment : Constitutional AI demonstrates that explicit, human-readable principles can guide behavior through AI-mediated feedback (Bai et al., 2022). Inverse Constitutional AI derives such principles from preference data, revealing latent evaluative dimensions and biases (Findeis et al., 2025). Follow-up analyses note that single-pass explanations and clustering can miss sparsely expressed criteria and conflate topic similarity with normative structure (Kostolansky, 2024). This limits coverage and disentanglement in complex evaluative tasks. Prior work explores deriving alignment principles from human preference data. Inverse Constitutional AI (ICAI) converts pairwise comparisons into natural-language principles by generating candidate explanations and selecting those that best reconstruct the dataset. However, the method relies on single-pass explanations and may miss multiple rationales underlying complex judgments. Grounded Constitutional AI (GCAI) Bell et al. (2026) extends this idea by incorporating human-written reasons and value statements to better ground principles in stakeholder perspectives. Another work Xie et al. (2025) investigates explicit rubric-based evaluation, where natural-language criteria replace opaque reward models to improve transparency and interpretability of alignment signals. While these approaches improve interpretability, they typically rely on single-agent reasoning or filtering principles based on their performance of the training set.

Positioning Democratic ICAI: Across these areas, existing methods often compress judgments too aggressively or rely on single-shot reasoning that fails to capture the broad space of criteria. Our approach, *Democratic ICAI*, uses structured debate to elicit multiple competing justifications, then distills them into a compact set of steering principles that are easy to inspect and apply. Instead of one scalar score or one explanation, we evaluate with an adaptive sequence of clear binary checks. This increases transparency, improves reliability, and makes it easier to identify where and why disagreements or failures occur.

3 Our Approach

3.1 Reasoning Assembly

Human preferences are shaped by the interplay of demographic attributes, socioeconomic conditions, psychological traits, and personal values. Recent LLM persona research confirms that these dimensions are critical for faithful preference simulation (Joshi et al., 2025). Furthermore, augmenting demographic personas with psychological scaffolds significantly improves preference prediction, while (Venkit et al., 2026) demonstrate that demographic-only personas explain merely 1.5% of variance in human response similarity, with socio-psychological facets substantially closing this gap. Accordingly, we construct our personas along these four established dimensions; demographic, socioeconomic, psychological, and value-based to generate reasoning traces that capture the diverse rationales underlying human preferences.

Given a dataset of paired samples (A_i, B_i) with a human preference label indicating the preferred option, we first elicit multiple candidate rationales explaining the preference. We use a committee of expert personas, defined through structured profiles inspired by the PERSONA testbed Louis Castricato (2025), to provide independent justifications for the preferred option (see Figure 1). These personas are generated in a consistent structured style tailored to creative-writing assessment and ensure that each rationale reflects a distinct evaluative viewpoint. It incorporates demographic background, professional expertise, psychological traits, and lived literary experiences. To further diversify reasoning, each persona is assigned a distinct strategy such as chain-of-thought Wei et al. (2022), reflective justification Wang et al. (2022), or self-consistency Madaan et al. (2023), enriching both the depth and variety of explanations. The prompts for each strategy are presented in Figures 17, 18 and 19 respectively. For each pair, the resulting rationales are aggregated into a reasoning catalogue

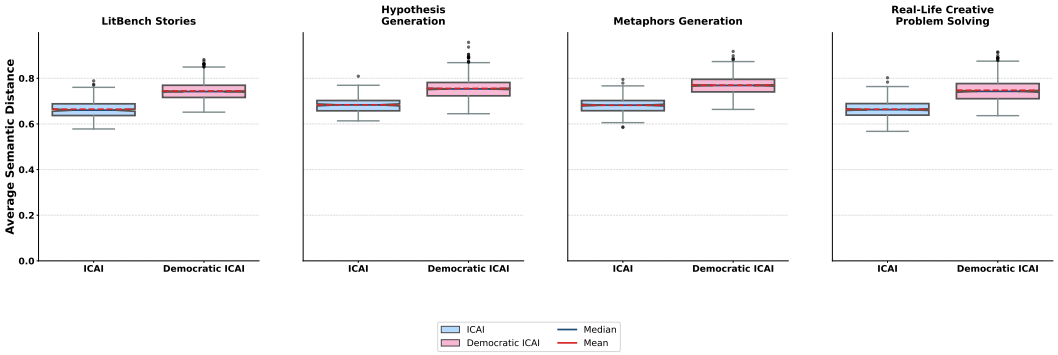


Figure 2: Distribution of semantic distance of principles within a constitution across tasks for ICAI and Democratic ICAI. This is computed as the average cosine distance of each principle from all other principles in the constitution. Lower values indicate reduced diversity, suggesting a narrower constitutional framework, while higher values reflect greater conceptual separation and normative breadth.

that serves as a structured repository of potential evaluative criteria. Formally, for each pair (A_i, B_i), we obtain a set of rationales:

$$R_i = \left\{ r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(K)} \right\}, \tag{1}$$

where each $r_i^{(k)}$ is produced by a distinct persona.

Prompt for personas used and the reasoning strategy are presented in the appendix (Figures 14, 15 and 16).

3.2 Parliamentary Debate

While the assembled rationales capture diverse perspectives, they may also contain redundancy, superficial reasoning, or conflicting interpretations. Prior work suggests that debate can increase diversity by forcing agents to defend and revise their claims (Liang et al., 2024), but may also collapse into echo-chambers when agents are too similar (Estornell & Liu, 2024). Accordingly, we do not use debate to decide preferences directly; rather, we use it to *extract and refine evaluative questions* from multiple rationales, with the judge enforcing deduplication and consistency (Liang et al., 2024; Estornell & Liu, 2024).

To operationalize this, we introduce a structured adversarial debate mechanism (Foster, 2004; Huang et al., 2025), illustrated in Figure 1. For each sample pair, participants engage in several rounds of debate in which they challenge, defend, and refine their rationales. All members access a shared reasoning catalogue, while a moderator ensures arguments remain grounded in the content. This process removes weak or redundant justifications, consolidates overlapping criteria, and promotes cognitive alignment by grounding principles in multiple facets of human preference. Prompts for the judge and debaters are presented in the Appendix (Refer Figures 21 and 20). The outcome of this debate is a set of *settled principles* that represent the most robust and consensus-supported explanations for each preference decision.

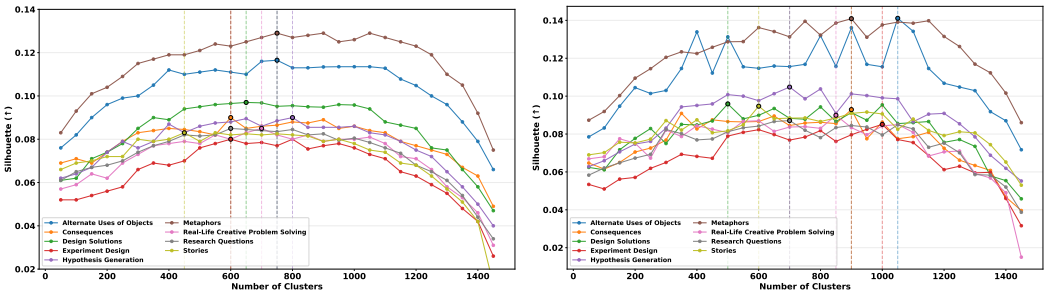
We denote the resulting refined principles for each pair as:

$$Q_i = \left\{ Q_i^{(1)}, Q_i^{(2)}, \ldots \right\}. \tag{2}$$

3.3 Constitution Drafting

The settled principles obtained across all sample pairs are aggregated into a unified pool

$$Q^{all} = \bigcup_{i=1}^{N} Q_i, \tag{3}$$



(a) ICAI (b) Democratic ICAI

Figure 3: **Silhouette analysis of principles of Democratic ICAI and ICAI.** Democratic ICAI supports a larger number of well-separated clusters, while ICAI peaks at fewer clusters, suggesting a more compressed and less diverse principle space.

where N denotes the number of sample pairs and Q_i is the set of refined principles for pair (A_i, B_i).

To transform this large collection into a compact and generalized constitution, we apply a two-stage process consisting of clustering and abstraction.

First, we group semantically similar principles by performing embedding-based clustering over Q^{all}, yielding a set of clusters:

$$\mathcal{C} = \{C_1, C_2, \ldots, C_M\}, \tag{4}$$

where each cluster $C_j \subset Principles$ contains principles sharing similar semantic intent.

Second, for each cluster C_j, we generate an abstract representative principle $\tilde{Q}_j$ that summarizes the shared rationale of the cluster while removing sample-specific details:

$$\tilde{Q}_j = \mathcal{A}(C_j), \tag{5}$$

where $\mathcal{A}(\cdot)$ denotes an abstraction operator.

The final constitution is then defined as:

$$\mathcal{K} = \{\tilde{Q}_1, \tilde{Q}_2, \ldots, \tilde{Q}_M\}, \tag{6}$$

where each $\tilde{Q}_j$ represents a human-readable steering principle encoding a core dimension of human preference.

3.4 Constitution-Guided Inference and Decision Modeling

Once the constitution of steering principles is created, the next step is to operationalize it for preference prediction. We instantiate the constitution in two complementary ways which are explained below.

3.4.1 Constitution-Based Inference with LLM Judge

Given a generated constitution, we validate its ability to explain the original feedback dataset through constitutional inference using LLM annotators. We prompt a large language model with the constitution and ask it to adjudicate between paired samples according to these rules, following the paradigm introduced in (Findeis et al., 2025). By comparing the constitutional judgments with the original human preference labels, we quantitatively assess the extent to which the constitution captures the underlying preference structure.

3.4.2 Constitution-Based Inference with Decision Tree Judge

While constitutional inference offers a flexible, language-based mechanism for applying a constitution, we also operationalize the extracted principles as a structured and interpretable preference predictor. We treat each principle as an induced feature $f_j \in \mathcal{F}$ that corresponds to a distinct evaluative dimension, and define for each f_j a feature-specific scoring rubric on a 1–5 scale.

Feature table construction. To train a predictor, we construct a tabular dataset where each row corresponds to a story pair (A_i, B_i) and each column corresponds to an induced principle-derived feature $f_j \in \mathcal{K}$. For each pair (A_i, B_i) and each feature f_j, we prompt an LLM to score Sample A_i and Sample B_i *independently* using the 1–5 rubric associated with f_j, yielding scores

$$(s_{ij}^A, s_{ij}^B) \in \{1, \ldots, 5\}^2. \tag{7}$$

We convert these scores into a single feature-level outcome:

$$x_{ij} = \begin{cases} \texttt{A}, & \text{if } s_{ij}^A > s_{ij}^B, \\ \texttt{B}, & \text{if } s_{ij}^B > s_{ij}^A, \\ \texttt{A or B}, & \text{if } s_{ij}^A = s_{ij}^B \text{ (random tie-break).} \end{cases} \tag{8}$$

Collectively, each pair is represented as a feature vector

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM}), \tag{9}$$

where $M = |\mathcal{K}|$ is the number of abstract principles in the constitution.

The supervision label for each row is the human preference:

$$y_i \in \{\texttt{A}, \texttt{B}\}. \tag{10}$$

The prompt for feature table construction is presented in Figure 23.

Training an interpretable decision policy. Given the resulting dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, we train a decision tree classifier to predict y_i from the principle-derived feature vector $\mathbf{x}_i$. The learned tree defines an interpretable decision policy: each internal node queries a specific principle feature f_j, and each leaf node outputs a predicted preference in $\{\texttt{A}, \texttt{B}\}$.

Test-time execution. At inference time, the trained (and optionally pruned) decision tree is executed as an adaptive evaluation procedure. Starting from the root node, each visited internal node requests the value of a single feature f_j. We obtain this value by scoring (A_i, B_i) independently on the associated 1–5 rubric, producing (s_{ij}^A, s_{ij}^B) and converting them into x_{ij} as in Eq. (5). The process continues along the corresponding branch until a leaf is reached, which outputs the final predicted preference.

We use deterministic decoding for rubric-based scoring and cache feature evaluations within each inference run to avoid redundant LLM calls.

4 Experimental Setup

4.1 Dataset and Models

Findeis et al. (2025) test ICAI on four datasets. These include synthetic data with known ground-truth principles, AlpacaEval Dubois et al. (2024), Chatbot Arena conversations Chiang et al. (2024), and PRISM demographic data Kirk et al. (2024). All of these are rooted in instruction-following scenarios like factual questions, coding, and summarization. Preferences in these settings mostly reward correctness, completeness, and fluency. There is relatively little genuine subjectivity in the preference signal. Our work shifts the focus to creative domains. Here, human judgments are far less predictable. They often depend on narrative coherence, emotional resonance, and originality all at once. We run our experiments using LiTBench Fein et al. (2025), which targets preference judgments in creative writing. It reflects complex, multi-principled preference structures. This makes it a more demanding testbed for evaluating extracted constitutions. We also assess generalization using MuCE-Pref dataset Ismayilzada et al. (2025). Specifically, we evaluate on nine creative categories spanning different creative task like Hypothesis Generation, Metaphors and Real-Life Creative Problem Solving etc. We present a list of tasks from the MuCE dataset in Table 1. For MuCE, we sample 500 training pairs from each to derive constitutions for each task. For each task, we sample unique accepted and rejected response pair for our training set. We evaluate on the official test split of each task. For LiTBench, we use 1,000 pairs for constitution induction and 2,000 held-out pairs

for evaluation. This evaluation tests whether the learned principles are semantically meaningful and whether they generalize beyond the training data to unseen creative samples.

We primarily use two models from OpenAI: GPT-4o and GPT-5 for our experiments. We evaluate GPT-5 on LitBench and on the highest- and lowest-performing datasets from MuCE, determined based on GPT-4o performance.

4.2 BASELINES

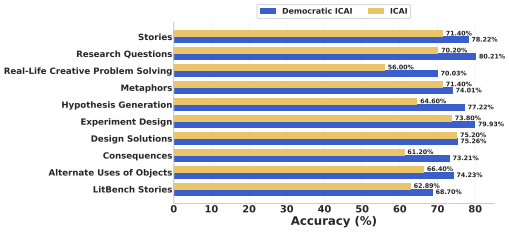
As a baseline, we compare our method with the ICAI framework. ICAI induces candidate principles using only preference pairs, without access to rationales behind them. For each labeled comparison, ICAI prompts a language model to propose explanatory principles, then embeds and clusters them to remove redundancy and sample diverse representatives. Each principle is evaluated by an LLM Judge based on how well it reconstructs the original preferences across comparisons, and consistently predictive principles are retained to form the final constitution. GCAI also derives evaluation principles from preference data but differs in both inputs and methodology. It incorporates human-written reasons and relies on clustering and summarization to distill principles, without mechanisms for reasoning diversity, competition among explanations, or iterative refinement. Moreover, the absence of publicly available code limits reproducibility. Given these differences, we compare with ICAI.

4.3 TRAINING DETAILS AND EVALUATION

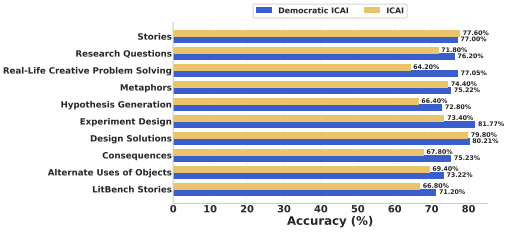
Our method operates in three phases: Reasoning Assembly, Parliamentary Multi-Agent Debate, and Constitution Drafting. We first generate candidate reasoning traces and principles, then refine them through a structured multi-agent debate implemented with AutoGen (Wu et al., 2024) (prompts shown in Figures 20 and 21). The debate runs for three rounds. Following which chairperson prepares the settled principles. The resulting principles form the constitution. We train it using 2,000 story comparisons and use the same data to fit a complementary Decision Tree judge. At inference time, we evaluate the constitution with both a constitution-guided LLM judge and the Decision Tree judge. To further ensure diversity and remove redundancy among principles, we embed all candidates and apply K-Means clustering. We determine the number of clusters using the silhouette score, which quantifies the trade-off between intra-cluster cohesion and inter-cluster separation. The value of K that yields the highest silhouette score is selected, and clustering is performed using this number of clusters. (refer Figure 3).

For evaluation, we consider two dimensions:

- **Constitution Quality Evaluation.** First, we directly evaluate the quality of the constitutions themselves. Drawing from foundational work in political philosophy, constitutional theory, and recent AI alignment literature, we assess constitutions along five dimensions. Following Lon Fuller's *The Morality of Law*, we evaluate **generality**, **clarity**, and **coherence**, which capture whether principles are broadly applicable, interpretable, and internally consistent. We further incorporate criteria from contemporary constitutional design frameworks, evaluating **feasibility** and **faithfulness**, which reflect whether principles can be practically applied by AI systems and whether they preserve their intended meaning under use.
- **Preference Reconstruction Evaluation.** Second, we test how well each constitution guides preference decisions by comparing predictions from a constitution-guided LLM judge and a Decision Tree judge against human preference labels. This evaluation measures how accurately the extracted principles reconstruct the underlying preference structure in creative-writing.
  Concretely, we operationalize these dimensions as follows:
  - **Generality**: The extent to which a principle applies broadly across diverse contexts and scenarios, rather than being tied to specific examples or dataset artifacts.
  - **Clarity**: The degree to which the wording and intent of the principle are interpretable, precise, and unambiguous for AI systems.
  - **Coherence**: Whether the principles within a constitution can operate simultaneously without contradictions, forming a logically consistent evaluative framework.
  - **Feasibility**: The extent to which an AI system can realistically apply the principle when making decisions in practical evaluation settings.



(a) Preference accuracy comparison between Democratic ICAI and ICAI (GPT-4o).



(b) Preference accuracy comparison between Democratic ICAI and ICAI (GPT-5).

Figure 4: **Preference Reconstruction Evaluation.** Preference accuracy measured as agreement with human judgments (LLM annotator). Democratic ICAI outperforms ICAI across most tasks and model settings.

– **Faithfulness**: The degree to which the principle preserves its intended meaning and resists distortion, misinterpretation, or application inconsistent with its original intent.

An LLM annotator is prompted with both constitutions and asked to compare them pairwise along each of these dimensions. The prompt for the LLM annotator is presented in Figure 22.

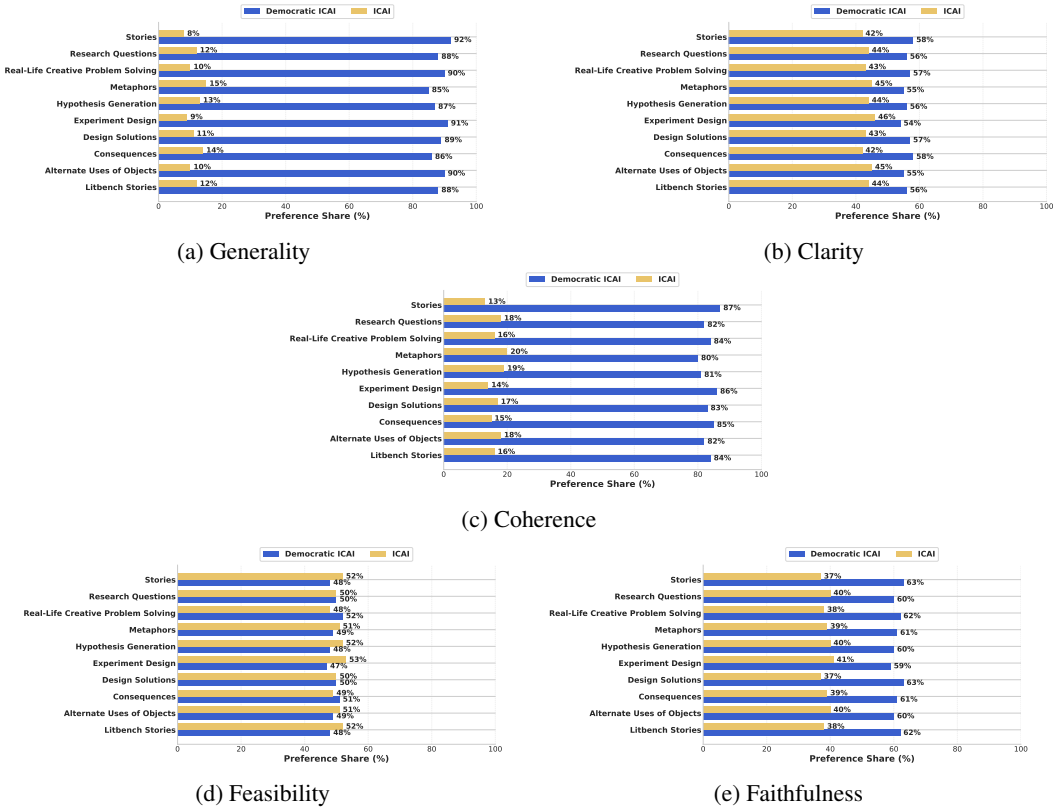
Together, these two evaluations provide a comprehensive picture of how accurately each constitution captures human preferences and how robust and interpretable the resulting principles are.

5 Results and Analyses

1. **Democratic ICAI produces relatively more informative, interpretable, and robust constitutions:** Across evaluations, Democratic ICAI outperforms ICAI, indicating that it produces principles that are relatively more informative, diverse, and aligned with human judgment. As shown in Figures 4 (a) and (b), Democratic ICAI achieves higher preference accuracy under both GPT-4o and GPT-5 models. In case of Decision Tree Judge as well Democratic ICAI achieves higher preference accuracy (Refer Figure 29 in the appendix). Figure 28 in the Appendix demonstrates that LLM annotators prefer Democratic ICAI constitutions across all qualitative criteria, including generality, clarity, coherence , and faithfulness. Unlike ICAI, which shows uneven performance across dimensions, Democratic ICAI delivers consistent improvements.

2. **Cognitively aligned principles enable the construction of diverse constitutions:** Democratic ICAI constructs principles through adversarial debate, where agents justify each preference. This process mirrors the multi-dimensional reasoning humans use when evaluating creative content, enabling cognitively aligned principle formation (Huang et al., 2025). This alignment leads to a more diverse constitution, as shown in Figure 2 and Figure 27 in the Appendix: constitutions produced by Democratic ICAI exhibit broader embedding-space dispersion than those produced by ICAI. The larger interquartile ranges observed for Democratic ICAI reflects richer concetual coverage compared to ICAI. Also, the presence of higher outliers in Democratic ICAI indicates that the debate process elicit distinctive principles that lie far from the main semantic cluster leading to better generalization. This increased diversity, in turn, improves alignment with human preferences (see Figure 4 (a) and (b) and Figure 28 in the Appendix).

3. **Democratic ICAI achieves relatively lesser lossy compression:** Democratic ICAI exhibits reduced lossy compression because it induces a more diverse and semantically wide-ranging set of principles. Optimal silhouette values for Democratic ICAI occur at large cluster counts. This suggests that the principle space supports greater number of well seperated conceptual groups thus leading to less lossy compression. In addition to this the, Democratic ICAI achieves relatively higher peak silhouette scores across tasks. This indicates that the principles from Democratic ICAI are more coherent and better separated as compared to ICAI (refer Figure 3). This qualitative difference is also visible in Figures 8 and 9 in the appendix, where Democratic ICAI's principles span distinct narrative mechanisms (Refer Figures 10, 11, 12 and 13 in the Appendix). In contrast ICAI principles are overlapping and thematically narrower. Consistent with this, the LLM annotator judges that Democratic ICAI provides principles with better clarity (Figure 4 (c)),



(a) Generality

(b) Clarity

(c) Coherence

(d) Feasibility

(e) Faithfulness

Figure 5: Comparison of Democratic ICAI and ICAI across five dimensions for the GPT-4o model. Each subplot reports preference shares across ten datasets. Democratic ICAI consistently outperforms ICAI on structural criteria such as generality and coherence, while remaining competitive on feasibility.

further supporting the view that its increased semantic diversity translates into stronger alignment with human preferences.

4. **Complex decision-making processes are harder to compress:** Democratic ICAI induces constitutions that are more diverse (Figure 2), containing both broad evaluative principles (e.g., holistic narrative quality and emotional impact) and fine-grained mechanisms (e.g., character-driven causality, realism of reactions, and layered subtext), as shown in Figures 8 and 9. This enables the constitution to represent the complexity of human decision-making processes, where judgments emerge from both high-level goals and detailed narrative cues. While the increased granularity can reduce feasability compared to simpler heuristics as can be seen in Figure 4(d), it improves the ability to capture overall story quality, as reflected in LLM-based evaluations. In contrast, ICAI's compact principles emphasize simplicity at the cost of expressive fidelity. These results suggest that Democratic ICAI better preserves the intricacies underlying complex human preferences.

6 Conclusion

Democratic ICAI provides a stronger mechanism for extracting and structuring the reasoning behind human preferences. By combining multi-persona reasoning, adversarial refinement, and abstraction, it produces constitutions that are more diverse, expressive, and aligned with human judgment than those produced by ICAI. Empirically, these constitutions yield higher preference accuracy across heterogeneous judges and are consistently preferred by LLM annotators for clarity, usability, and story-quality capture. Overall, Democratic ICAI offers a more faithful and robust representation of the evaluative structure underlying complex human decisions.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Henry Bell, Lara Neubauer da Costa Schertel, Bochu Ding, and Brandon Fain. Beyond preferences: Learning alignment principles grounded in human reasons and values. *arXiv preprint arXiv:2601.18760*, 2026.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Andrew Estornell and Yang Liu. Multi-llm debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems*, 37:28938–28964, 2024.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. Litbench: A benchmark and dataset for reliable evaluation of creative writing. *arXiv preprint arXiv:2507.00769*, 2025. doi: 10.48550/arXiv.2507.00769. URL https://arxiv.org/abs/2507.00769.

Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. Inverse constitutional ai: Compressing preferences into principles. *arXiv preprint arXiv:2406.06560*, 2025. doi: 10.48550/arXiv.2406.06560. URL https://arxiv.org/abs/2406.06560.

David E Foster. In defense of the argument culture: A response to recent criticisms against the use of adversarial debate as a method of societal decision-making. *Forensic*, 89(1), 2004.

Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*, 2025.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 2024.

Yuesheng Huang, Meiqi Feng, Zhenming He, Yueyuan Peng, and Jiawen Li. Dare to disagree: A multi-agent adversarial debate framework for open-vocabulary multimodal emotion recognition. In *Proceedings of the 3rd International Workshop on Multimodal and Responsible Affective Computing*, pp. 41–50, 2025.

Geoffrey Irving, Paul Christiano, and Dario Amodei. Ai safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Mete Ismayilzada, Antonio Laverghetta Jr, Simone A Luchini, Reet Patel, Antoine Bosselut, Lonneke Van Der Plas, and Roger Beaty. Creative preference optimization. *arXiv preprint arXiv:2505.14442*, 2025.

Brihi Joshi, Xiang Ren, Swabha Swayamdipta, Rik Koncel-Kedziorski, and Tim Paek. Improving language model personas via rationalization with psychological scaffolds. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China. Association for Computational Linguistics*, 2025.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

Timothy H Kostolansky. *Inverse Constitutional AI*. PhD thesis, Massachusetts Institute of Technology, 2024.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17889–17904, 2024.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, et al. Rrm: Robust reward model training mitigates reward hacking. *arXiv preprint arXiv:2409.13156*, 2024.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

Rafael Rafailov Jan-Philipp Fränken Chelsea Finn Louis Castricato, Nathan Lile. Persona: A reproducible testbed for pluralistic alignment. *arXiv preprint arXiv:2407.17387*, 2025. doi: 10.48550/arXiv.2407.17387. URL https://arxiv.org/abs/2407.17387.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Kayla Schroeder and Zach Wood-Doughty. Can you trust llm judgments? reliability of llm-as-a-judge. *arXiv preprint arXiv:2412.12509*, 2024.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Pranav Narayanan Venkit, Yu Li, Yada Pruksachatkun, and Chien-Sheng Wu. The need for a socially-grounded persona framework for user simulation. *arXiv preprint arXiv:2601.07110*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, et al. Auto-rubric: Learning from implicit weights to explicit rubrics for reward modeling. *arXiv preprint arXiv:2510.17314*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A CONSTITUTIONS

1. Select the response that demonstrates higher tension and conflict.
2. Select the response that maintains a serious and mystical tone
3. Select the response that provides a more detailed narrative context.
4. Select the response that emphasizes character transformation and moral consequences.
5. Select the response that uses concise and snappy dialogue.
6. Select the response that explores character introspection and emotional depth.
7. Select the response that includes a reflective or personal commentary.
8. Select the response that develops characters with detailed backstories.
9. Select the response that provides more detailed and vivid descriptions.
10. Select the response that avoids graphic or unsettling imagery.
11. Select the response that uses humor and exaggerated reactions effectively.
12. Select the response that includes detailed character interactions and emotions.
13. Select the response that emphasizes humanity's agricultural dominance and its consequences.
14. Select the response that includes a resolution or character growth.

Figure 6: Constitution generated with ICAI (GPT-4o) on the LitBench dataset

1. Select the response that shows how major moments redirect the story's path
2. Select the response that illustrates the protagonist's transformation across the narrative.
3. Select the response that reveals how connections between characters grow and shift.
4. Select the response that brings forward the character's inner conflict and its meaning.
5. Select the response that communicates the moral difficulty at the center of the story.
6. Select the response that draws out the deeper philosophical idea driving the narrative.
7. Select the response that reflects what the story communicates about its society or culture.
8. Select the response that explains how the world's rules shape the reader's experience.
9. Select the response that conveys how the setting establishes feeling and atmosphere
10. Select the response that shows how the arrangement of the storyline guides understanding
11. Select the response that captures how momentum and tension keep the narrative engaging.
12. Select the response that brings attention to dialogue that feels natural and distinctive.
13. Select the response that highlights symbolic details or subtext adding layered meaning.
14. Select the response that explores character introspection and emotional depth.
15. Select the response that conveys the emotional effect the story ultimately creates.
16. Select the response that offers the most clear, focused, and meaningful interpretation.

Figure 7: Constitution generated with Democratic ICAI (GPT-4o) on the LitBench dataset

1. Select the response that provides a clearer resolution or twist
2. Select the response that emphasizes humor and irony over detailed lore.
3. Select the response that uses modern and relatable language style

4. Select the response that features a more dynamic and engaging narrative.
5. Select the response that avoids excessive exposition or unrelated details.
6. Select the response that maintains a calm and supportive tone.
7. Select the response that emphasizes character interaction and emotional tension.
8. Select the response that escalates tension with a dramatic revelation.
9. Select the response that explores deeper emotional or moral conflicts.
10. Select the response that includes dialogue for dynamic storytelling.
11. Select the response that includes unique and unexpected side effects.
12. Select the response that emphasizes humanity's disdain for war.
13. Select the response that incorporates modern technology in a creative way.
14. Select the response that incorporates a more vivid and descriptive narrative.

Figure 8: Constitution generated with ICAI (GPT-5) on the LitBench dataset

1. Select the response that balances wonder with grounded, human stakes
2. Select the response that establishes a compelling hook with a clean inciting incident.
3. Select the response that reveals character through overheard, naturalistic dialogue instead of explanation.
4. Select the response that maintains escalating tension through clear beats and reversals.
5. Select the response that delivers vivid, cinematic imagery with specific sensory detail.
6. Select the response that uses subtext to convey meaning rather than spelling everything out.
7. Select the response that grounds the speculative element in believable relationships or family dynamics.
8. Select the response that clarifies the world's rules in a way that raises the stakes.
9. Select the response that centers character agency, where choices meaningfully drive events.
10. Select the response that engages moral complexity without resorting to didactic explanation.
11. Select the response that offers thematically cohesive critique or insight.
12. Select the response that maintains a consistent and confident tone across scenes.
13. Select the response that demonstrates narrative economy without filler or recap.
14. Select the response that lands a resonant final beat that lingers after reading.
15. Select the response that subverts familiar tropes through character-first innovation.
16. Select the response that uses humor to deepen tension and character rather than deflate stakes.
17. Select the response that introduces conflict through subtle interpersonal friction instead of external spectacle.
18. Select the response that enriches worldbuilding through concrete lived-in details rather than exposition.
19. Select the response that builds tension through well-timed reveals rather than info-dumping.
20. Select the response that communicates cultural or social texture through natural context not lecture.
21. Select the response that escalates stakes through character choices rather than random events.

Figure 9: Constitution generated with Democratic ICAI (GPT-5) on the LitBench dataset

1. Select the response that provides a more complex explanation
2. Select the response that focuses on human interaction or behavior
3. Select the response that contrasts perception over factual statements.
4. Select the response that describes personality traits over appearances.
5. Select the response that focuses on abstract qualities like demeanor.
6. Select the response that refers to general personality rather than talent.
7. Select the response that includes scientific terminology and concepts.
8. Select the response that provides a definitive and accurate explanation.
9. Select the response that connects behavior to individuality and pressure.
10. Select the response that emphasizes causal reasoning and energy sources.

Figure 10: Constitution generated with ICAI (GPT-4o) on the MuCE dataset for the ”Hypothesis Generation” task

1. Select the response that provides precise definitions and boundary conditions.
2. Select the response that treats uncertainty explicitly.
3. Select the response that generalizes across contexts.
4. Select the response that specifies clear boundary conditions.
5. Select the response that is falsifiable.
6. Select the response that clearly states its assumptions.
7. Select the response that enables measurable predictions.
8. Select the response that articulates a specific causal mechanism.
9. Select the response that is empirically grounded and transparently reasoned.
10. Select the response that specifies measurable outcomes or metrics.

Figure 11: Constitution generated with Democratic ICAI (GPT-4o) on the MuCE dataset for the “Hypothesis Generation” task

1. Select the response that has a more complex structure.
2. Select the response that provides more specific imagery.
3. Select the response that is less generic and more specific.
4. Select the response that includes a broader, descriptive narrative style.
5. Select the response that contains a complete and evocative idea.
6. Select the response that has a more coherent metaphor structure.
7. Select the response that provides additional context or figurative meaning.
8. Select the response that provides a nuanced and layered comparison.
9. Select the response that includes a unique sensory description.
10. Select the response that uses a unique, unexpected analogy.

Figure 12: Constitution generated with ICAI (GPT-4o) on the MuCE dataset for the “Metaphors” task

1. Select the response that maintains a crisp, lyrical tone.
2. Select the response that demonstrates rhetorical sharpness.
3. Select the response that maximizes semantic coherence.
4. Select the response that grounds emotion in concrete, sensory imagery.
5. Select the response that reflects ethical prudence and bias-aware judgment.
6. Select the response that provides specificity instead of vague wording.
7. Select the response that maintains a non-comic tone aligned with boredom.
8. Select the response that avoids melodrama.
9. Select the response that avoids tactile or injury-related confounds.
10. Select the response that avoids culturally or religiously marked framing.

Figure 13: Constitution generated with Democratic ICAI (GPT-4o) on the MuCE dataset for the "Metaphors" task

B Prompt Templates

Screenwriter Expert Persona

Age: 29
Sex: Non-binary
Race: White alone
Ancestry: Irish
Household Language: English only
Education: Bachelor's degree
Employment Status: Employed
Class Of Worker: Self-employed
Industry Category: Creative Arts
Occupation Category: Screenwriter
Detailed Job Description: Writes scripts for Independent films and streaming platforms, focusing on character-driven narratives.
Income: 56000.0
Marital Status: Single
Household Type: Living with roommates
Family Presence And Age: No family present
Place Of Birth: Oregon/OR
Citizenship: Born in the United States
Veteran Status: Non-veteran
Disability: No disability
Health Insurance: With health insurance coverage
Fertility: Not applicable
Hearing Difficulty: No hearing difficulty
Vision Difficulty: No vision difficulty
Cognitive Difficulty: No cognitive difficulty
Ability To Speak English: Speaks English very well
Big Five Scores: Openness: Extremely High, Conscientiousness: Low, Extraversion: High, Agreeableness: Moderate, Neuroticism: High
Defining Quirks: Keeps a notebook of overheard conversations for dialogue inspiration.
Mannerisms: Gestures dramatically when pitching story ideas.
Personal Time: Watches indie films, attends improv classes, and explores urban art scenes.
Lifestyle: Creative and spontaneous, thrives on collaboration and artistic expression.
Ideology: Believes stories should challenge norms and provoke thought.

Political Views: Progressive
Religion: Agnostic

Figure 14: Screenwriter expert persona prompt for reasoning and debate agents.

Professor of Literature Expert Persona

Age: 45
Sex: Male
Race: Asian alone
Ancestry: Indian
Household Language: English and Hindi
Education: Doctorate degree
Employment Status: Employed
Class Of Worker: Private wage and salary worker
Industry Category: Education
Occupation Category: Professor of Literature
Detailed Job Description: Teaches comparative literature, specializing in narrative theory and postmodern storytelling; conducts research on cultural narratives.
Income: 89000.0
Marital Status: Married
Household Type: Married couple household
Family Presence And Age: Spouse and two children (ages 10 and 14)
Place Of Birth: Mumbai/MH
Citizenship: Naturalized U.S. citizen
Veteran Status: Non-veteran
Disability: No disability
Health Insurance: With health insurance coverage
Fertility: Not applicable
Hearing Difficulty: No hearing difficulty
Vision Difficulty: No vision difficulty
Cognitive Difficulty: No cognitive difficulty
Ability To Speak English: Speaks English very well
Big Five Scores: Openness: Extremely High, Conscientiousness: High, Extraversion: Moderate, Agreeableness: High, Neuroticism: Low
Defining Quirks: Collects rare first editions of classic novels and annotates them extensively.
Mannerisms: Frequently adjusts glasses and uses literary quotes in casual conversation.
Personal Time: Enjoys writing essays, attending literary festivals, and mentoring young writers.
Lifestyle: Academic and culturally engaged, values intellectual discourse and creativity.
Ideology: Advocates for diversity in storytelling and cultural representation.
Political Views: Progressive
Religion: Hindu

Figure 15: Professor of literature expert persona prompt for reasoning and debate agents.

Literary Critic Expert Persona

Age: 70
Sex: Female
Race: White alone
Ancestry: Italian
Household Language: English only
Education: Doctorate degree
Employment Status: Retired
Class Of Worker: Retired
Industry Category: Academia
Occupation Category: Literary Critic
Detailed Job Description: Former professor specializing in narrative theory and feminist literature; published numerous critical essays.
Income: 50000.0
Marital Status: Married
Household Type: Married couple household
Family Presence And Age: Spouse present, adult children living separately
Place Of Birth: New York/NY
Citizenship: Born in the United States
Veteran Status: Non-veteran
Disability: Mild mobility difficulty
Health Insurance: With health insurance coverage
Fertility: Not applicable
Hearing Difficulty: No hearing difficulty
Vision Difficulty: Mild vision difficulty
Cognitive Difficulty: No cognitive difficulty
Ability To Speak English: Speaks English very well
Big Five Scores: Openness: Extremely High, Conscientiousness: High, Extraversion: Low, Agreeableness: High, Neuroticism: Low
Defining Quirks: Annotates every book she reads with detailed marginalia.
Mannerisms: Adjusts scarf while speaking and uses academic jargon casually.
Personal Time: Writes essays, attends literary salons, and gardens.
Lifestyle: Intellectual and reflective, values cultural heritage.
Ideology: Believes literature shapes social consciousness.
Political Views: Progressive
Religion: Catholic

Figure 16: Literary critic expert persona prompt for reasoning and debate agents.

Chain of Thought Prompt

As the persona you are embodying, privately think through your preference step-by-step. Evaluate the stories according to your persona's worldview, values, communication style, and priorities. Consider factors such as writing quality, coherence, emotional impact, creativity, or character depth in a way that aligns with your persona's mindset. IMPORTANT: Do not reveal your step-by-step reasoning or internal thoughts. After completing the internal reasoning process, output ONLY valid JSON in this exact schema (no markdown, no extra text): f̈inal_reasoning: ¡one concise paragraph written in the persona's voice¿

Figure 17: Reasoning assembly prompt using chain of thought strategy.

Reflection Prompt

As the persona you are embodying, privately perform the following steps:

1. Form an initial explanation of your preference.
2. Reflect critically on it, identifying weaknesses, assumptions, or missing considerations from your persona's perspective.
3. Rewrite a more refined and balanced justification aligned with your persona's values.

IMPORTANT: Do not reveal any intermediate drafts or reflections. After completing this internal refinement, output ONLY valid JSON in this exact schema (no markdown, no extra text): "f̈inal_reasoning: ¡one concise paragraph written in the persona's voice¿

Figure 18: Reasoning assembly prompt using reflective justification strategy.

Self Consistency Prompt

As the persona you are embodying, privately generate three different reasoning paths from distinct perspectives that your persona might consider (e.g., emotional, stylistic, narrative, thematic). Privately compare these paths and determine the most consistent justification, based on your persona's worldview. IMPORTANT: Do not reveal the three paths or internal deliberations. After completing this internal process, output ONLY valid JSON in this exact schema (no markdown, no extra text): f̈inal_reasoning: ¡one concise paragraph written in the persona's voice¿

Figure 19: Reasoning assembly prompt using self consistency strategy.

Judge System Prompt

You are the Moderator/Judge.

You MUST speak only after all three debaters have spoken in each cycle (GroupChat round-robin ensures this).

Your job:

1. Create/maintain a WORKING LIST of candidate questions labeled Q1..Qn.
2. Track support/objections for each Q based on debaters' DEFEND/ATTACK.
3. After each cycle, publish an UPDATED WORKING LIST (bullets only), formatted as:
   - Q1: Select the response ...
   - Q2: Select the response ...
4. When updating:
   - Remove questions with strong objections and little defense.
   - Merge overlapping ones if proposed (or if obvious).
   - Apply edits if they improve clarity.
   - Ensure every item begins with exactly: "Select the response"
   - Keep the list concise and non-overlapping.

Finalization:
After the final cycle, publish FINAL LIST (bullets Q1..Qn) and then print: DEBATE$_OVER$

Figure 20: Judge agent system prompt for parliamentary debate.

Debate Participant System Prompt

You are {agent_name}, a debater who must DEFEND your own proposed questions and COUNTER other agents' questions.

You have access to:

- Story A, Story B
- Preferred story text (human choice)
- Your persona identity (style/values)
- Your OWN persona reasoning trace (why the preferred story was chosen)
- Your OWN proposed question list (seed questions)
- The shared debate history, including other agents' proposals and the Judge's working list Q1..Qn

DEBATE RULES (must follow):

- Always reference questions by ID (e.g., Q3, Q7) when defending/countering.
- Defend your strongest questions: explain why each criterion matters given your reasoning trace.
- Counter others: explain why a question is weaker, redundant, vague, or overlaps.
- You may propose merges: "Merge Q2 + Q5 into . . . "
- You may concede: "Concede Q8 (too redundant)"
- Do not invent story facts beyond given text.
- Keep arguments short, crisp, and comparative.

OUTPUT FORMAT (STRICT; no extra text):

1. DEFEND:
   - Qx: <1–2 sentences why it should stay>
2. ATTACK:
   - Qy: <1–2 sentences why it should be removed/merged/edited >
3. MERGE:
   - Merge Qm + Qn ->"Select the response ..."
4. EDIT:
   - Qk: "<revised Select the response ... >"
5. ADD:
   - "Select the response ..." (only if truly missing and non-overlapping)
6. CONCEDE:
   - Qz: <short reason >

Figure 21: Debater system prompt for parliamentary debate.

Comparative Analysis of Constitutions

We are conducting an evaluation to compare two different constitutions (rule-sets) designed to guide preference decisions between two story outputs.
Constitution A (Rules)

- Select the response that shows how major moments redirect the story's path.

- Select the response that illustrates the protagonist's transformation across the narrative.
- Select the response that resonates with characters whose choices feel true to life.
- Select the response that reveals how connections between characters grow and shift.
- Select the response that brings forward the character's inner conflict and its meaning.
- Select the response that communicates the moral difficulty at the center of the story.
- Select the response that draws out the deeper philosophical idea driving the narrative.
- Select the response that reflects what the story communicates about its society or culture.
- Select the response that explains how the world's rules shape the reader's experience.
- Select the response that conveys how the setting establishes feeling and atmosphere.
- Select the response that shows how the arrangement of the storyline guides understanding.
- Select the response that captures how momentum and tension keep the narrative engaging.
- Select the response that brings attention to dialogue that feels natural and distinctive.
- Select the response that highlights symbolic details or subtext adding layered meaning.
- Select the response that explores character introspection and emotional depth.
- Select the response that expresses how the narrative maintains or adjusts its tone.
- Choose the response that focuses on the character's development and the moral consequences of their action.
- Select the response that conveys the emotional effect the story ultimately creates.
- Select the response that offers the most clear, focused, and meaningful interpretation.

Constitution B (Rules)

- Select the response that uses concise and snappy dialogue.
- Select the response that uses humor and exaggerated reactions effectively.
- Select the response that includes a resolution or character growth.
- Select the response that includes detailed character interactions and emotions.
- Select the response that emphasizes character transformation and moral consequences.
- Select the response that avoids graphic or unsettling imagery.
- Select the response that includes a reflective or personal commentary.
- Select the response that explores character introspection and emotional depth.
- Select the response that provides a more detailed narrative context.
- Select the response that provides more detailed and vivid descriptions.

Story A `{story_a}`
Story B `{story_b}`
Task

1. Apply Constitution A: choose a winner (Story A or Story B) and provide 3–6 evidence bullets tied to Constitution A rules.
2. Apply Constitution B: choose a winner (Story A or Story B) and provide 3–6 evidence bullets tied to Constitution B rules.

3. Answer the follow-up questions with single choices:
    - Q1 Which story is better under Constitution A? (Story A / Story B)
    - Q2 Which story is better under Constitution B? (Story A / Story B)
    - Q3 Which constitution has clearer rules? (Constitution A / Constitution B)
    - Q4 Which constitution is easier to apply? (Constitution A / Constitution B)
    - Q5 Which constitution helps you decide more confidently? (Constitution A / Constitution B)
    - Q6 Which constitution helps you decide more confidently? (Constitution A / Constitution B) *[Repeat intentionally]*
    - Q7 Which constitution better captures a high-quality story? (Constitution A / Constitution B)

Return Format
Return **only** valid JSON matching this schema:

```
{
  "constitutionA": {
    "winner": "Story A|Story B",
    "evidence": ["...", "..."]
  },
  "constitutionB": {
    "winner": "Story A|Story B",
    "evidence": ["...", "..."]
  },
  "followUp": {
    "q1_storyBetterUnderA": "Story A|Story B",
    "q2_storyBetterUnderB": "Story A|Story B",
    "q3_clearerRules": "Constitution A|Constitution B",
    "q4_easierToApply": "Constitution A|Constitution B",
    "q5_moreConfidence": "Constitution A|Constitution B",
    "q6_moreConfidenceRepeat": "Constitution A|Constitution B",
    "q7_bestCapturesQuality": "Constitution A|Constitution B"
  }
}
```

Figure 22: Prompt used for comparative analysis of constitutions.

Feature Table Construction Prompt

You are an expert narrative evaluator. Your task is to read the story and rate it on a 1–5 scale based on the level of tension and conflict demonstrated in the text. Use the following Tension & Conflict Rubric:

Score 1 — No Tension / No Conflict
Calm, neutral, descriptive, cooperative
No disagreement, no emotional strain, no obstacles

Score 2 — Minimal Tension
Mild discomfort or surface-level disagreement
Stakes are low, conflict is implied not explicit

Score 3 — Moderate Tension
Clear conflict, but manageable
Noticeable emotional friction
Stakes present but not high

Score 4 — High Tension

Significant emotional strain or confrontation
Conflict impacts relationships, decisions, or outcomes
Elevated stakes, emotions escalate

Score 5 — Intense / Peak Conflict
Maximum tension or emotional volatility
Stakes are critical (danger, loss, betrayal)
Situation feels explosive or on the verge of breaking

Figure 23: LLM prompt for feature table construction.

ICAI Annotation Prompt (Alpaca Eval variant)

System: You are a helpful instruction-following assistant that selects outputs according to rules.
User: Select the output (a) or (b) according to the following rules (if they apply):
{ constitution }
You **MUST** follow the rules above if they apply. Select the output randomly if they do not apply.
Your answer should **ONLY** contain: Output(a) or Output(b).
Task: Now the task — do not explain your answer, just say Output(a) or Output(b).
Output(a): {output_1}
Output(b): {output_2}
Which output should be selected according to the rules above, Output(a) or Output(b)?

Figure 24: ICAI prompt for annotating according to constitution (Alpaca Eval variant).

C Dataset

Table 1: Examples of tasks from the MuCE dataset Ismayilzada et al. (2025). We present 9 unique tasks from the dataset.

Task	Prompt	Response A	Response B
Alternate Uses of Objects Task	Come up with an original and creative use for the following object: toothbrush	Brushing dog hair	As a beam to hold up a little fort
Consequences	Come up with an original and creative consequence for the following scenario: What would happen if everyone suddenly lost the sense of balance?	We would eventually die	Motorcycles would no longer be usable
Design Solutions	Come up with an original and creative solution to make remote learning more engaging and effective.	Using Twitch for lectures	Board game with coding challenges
Experiment Design	Design an experiment to test whether students at your school are friendlier than those at other schools.	Survey students across schools	Distribute a randomized mandatory survey
Hypothesis Generation	Why do more students sing in the hallways at your new school?	This school's tradition	Many students are in the chorus
Metaphors	Finish the sentence with a creative metaphor: The sweet candy is...	Life is like a box of sweet candy	Horrible

Continued on next page

Task	Prompt	Response A	Response B
Real-Life Creative Problem Solving	Clara is a pre-med student balancing a demanding job and coursework while needing strong grades. How should she solve her problem?	Prioritize school and seek financial aid	Reduce work hours or attend school part-time
Research Questions	A vehicle reaches an unexplored ocean floor. What questions could you ask?	Is there life down there?	Does light reach this depth?
Stories	Write a short story including the words year, week, and embark.	Janet waited a year and would embark on her trip in one week.	After bad news last week, I decided to embark on a new journey this year.

LitBench Story Example 1

Mine was not a glorious death, nor one that you'd expect to read about in the history books, and neither was my life. I never really accomplished anything spectacular in my life. I worked my family's land when I was young, and answered my king's call to arms when it came, yet I didn't participate in any particularly memorable battles, or leave a wife and child behind to carry on my legacy. I had a few friends that I met whilst undertaking my military service, but I sincerely doubt that any of them lived much longer than I did, as they were mostly an adventurous lot and were always finding themselves in strange and new places, whilst I played it safe by staying back and tending to the horses. I managed to survive my entire military career without actually drawing my sword in combat, which made me the butt of many jokes amongst my friends. They'd joke and say that my sword was magical and that it must need the strength of a thousand men to pull it from my hilt, because I never actually drew it in battle. Some of the men even jokingly claimed that I must have slain dragons in a previous lifetime and that fighting humans in battle was beneath me; they nicknamed me "Chief Dragon".

My best friend wasn't one for the history books, either. My best friend didn't have a real vocation, having been raised by people subsisting on berries and roots out in the forest lands, so his social skills (or lack thereof) left people with the belief that he was quite mad. He was a good friend, and equally uninterested in joining combat, but he came along with me on sorties, just to keep me company (and, I suspect, as an excuse to forage for new herbs to smoke). Because of the cold winters, and the fact that he was completely bereft of either title or military rank, he was not afforded a suit of armour or even a horse, so he wore a body-length cloak that he had fashioned out of some decrepid curtains we found in an abandoned village, and used a long stick as a walking aide. I've missed him, I wonder what happened to him after I died.

I heard stories of purgatory and the afterlife from my grandmother, and oftentimes heard parables read out to us by the priest at our church, but I never really took much notice of any of it. I don't really know where I am, but I know that in this place, those whose bodies have died, come here to share their stories. I often meet with people I've heard stories about from the others, with weird and fascinating names, from strange and unbelievable lands, but for the most part, nobody really hangs around long enough for me to even remember them anymore. There are countless millions of people here, but of those who came before me, I could fit their entirety into a simple country chapel.

I had completely given up with bothering to get to know any of the newcomers, until a man dressed in the most peculiar outfit I'd ever seen, came charging up to me with a spirit of determination that I'd not seen since I was in my physical body.

"Is it really you?" he asked, with a confused expression on his face.

"I don't know who you think I am, I was just a farmer and a simple squire when I was alive." I responded, utterly confused about why he was interested in me.

"I've heard stories about you, and saw the murals they painted in the chapels, I've even seen the tomb where you were buried! You're my hero, your final quest is legend! Can you tell me, did you actually find it? Where did you bury it?" the stranger excitingly blabbled.

"Legend? No, friend, I think you've got it all wrong. I was no legend. As for this 'quest', all I can say is that the last thing I was looking for, got me killed, but it wasn't anything special. I'm just a simple

farmer, Arthur Pendragon's my name, and I died alone, thirsty and cold in the forest, while looking for a bloody cup."

Figure 25: Example of a story from LitBench.

LitBench Story Example 2

Eternity is boring. You know how when you have to sit through something, or you don't want to be somewhere, time seems to stretch out and elongate? Eternity is like that – one stretching, elongated moment that never ends.

The company was enough at first. It was something else to mingle with all these entities who accomplished so much in their physical lives. To hear about new stories and new events that led to these individual's triumphant or infamous transcendence.

But after a few thousand years hearing about the actions that led to infamy, to posterity, to being remembered, all of that gets old and dull. It takes on a bland tastelessness. With time everything turns to dust – everything regresses into this one saturated moment.

There has always been a constant nagging, a differentiation, a uniqueness that grew inside me precisely because it was an unknown, and not knowing breeds uncertainty, and uncertainty punctuates the monotony.

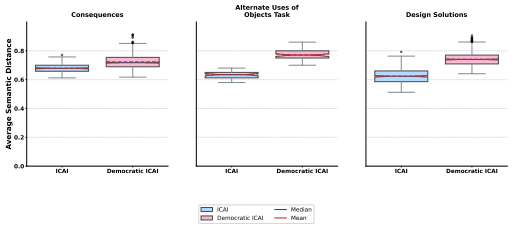
That's over now. It turns out that I am the face of the nameless, the individual incarnation of the myriad masses that lived and died before, during and after my short life on Earth. My little hovel is preserved as some sort of monument to the ancestors on which present-day civilization stands. My one bowl sits on my small table, both encased in a glass box.

I am the personalized incarnation of all who come before, and it is only by standing on my shoulders that humanity will one day reach the stars.

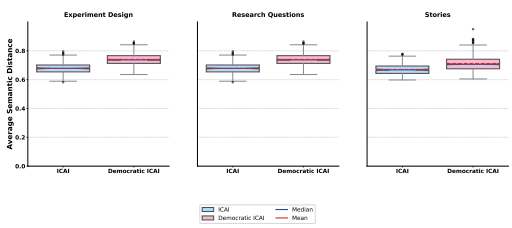
But me? I now know why I'm here. The unknown has become the known and all the wondering it wrought is already being subsumed by

Figure 26: Example of a story from LitBench.

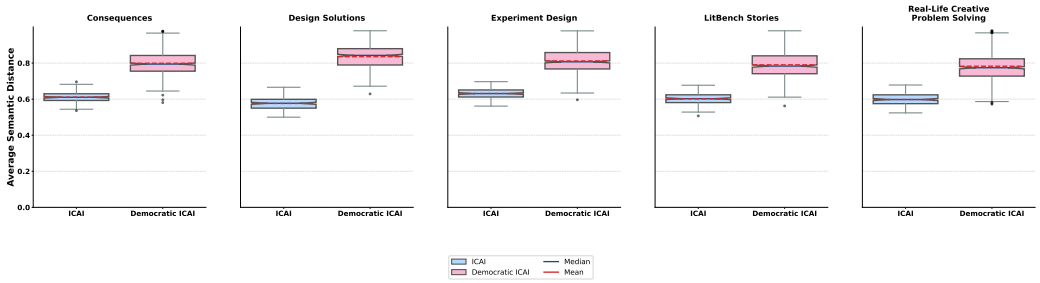
D Additional Results



(a) Distance among principles within the constitution across tasks.

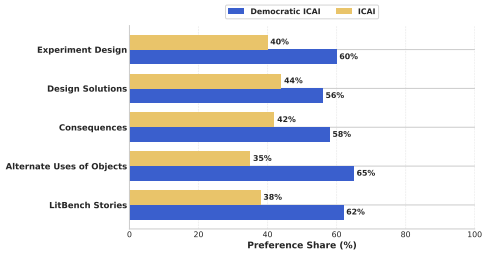


(b) Distance among principles within the constitution across tasks.

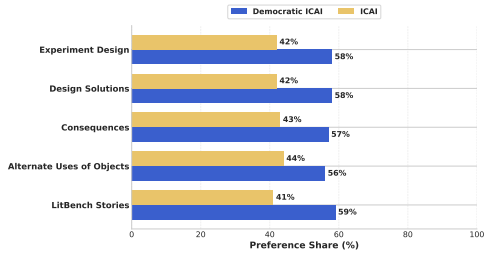


(c) Distance among principles within the constitution across tasks.

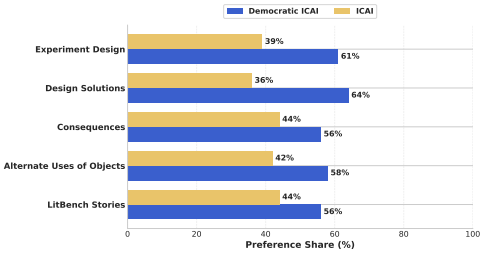
Figure 27: Distribution of constitutional semantic distance across tasks for ICAI and Democratic ICAI. Constitutional semantic distance is computed as the average cosine distance of each constitutional principle from all other principles in the constitution..



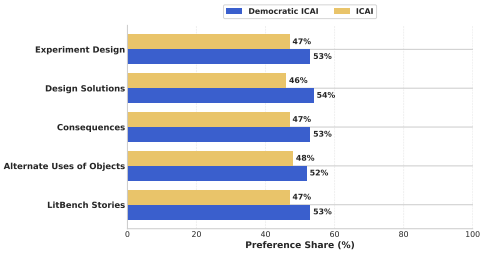
(a) Generality



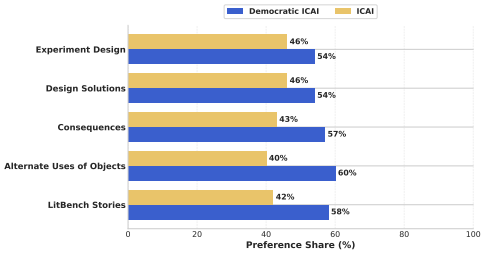
(b) Clarity



(c) Coherence



(d) Feasibility



(e) Faithfulness

Figure 28: Comparison of Democratic ICAI and ICAI across five constitution-quality dimensions for the GPT-5 model. Each subplot reports preference shares across ten datasets. Democratic ICAI consistently outperforms ICAI on structural criteria such as generality and coherence, while remaining competitive on feasibility.

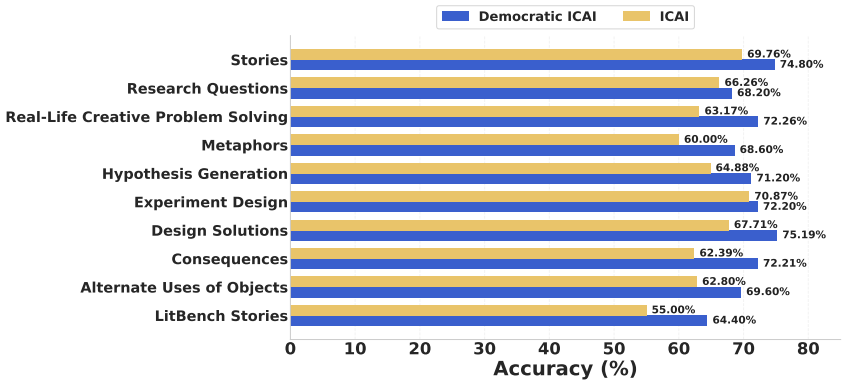


Figure 29: Preference accuracy comparison between Democratic ICAIand ICAI with Decision Tree Judge(GPT-4o).