Democratic ICAI: Debating Our Way to Steering Principles from Preferences

Anonymous ACL submission

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* (DICAI), 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 preference benchmarks, MuCE-Pref and LiT-Bench, across multiple creative task categories show that Democratic ICAI yields a more faithful preference structure. It improves average preference prediction across tasks relative to deliberative prompting and principle-based baselines, while producing constitutions that LLM annotators prefer.

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 and 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 annotators 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

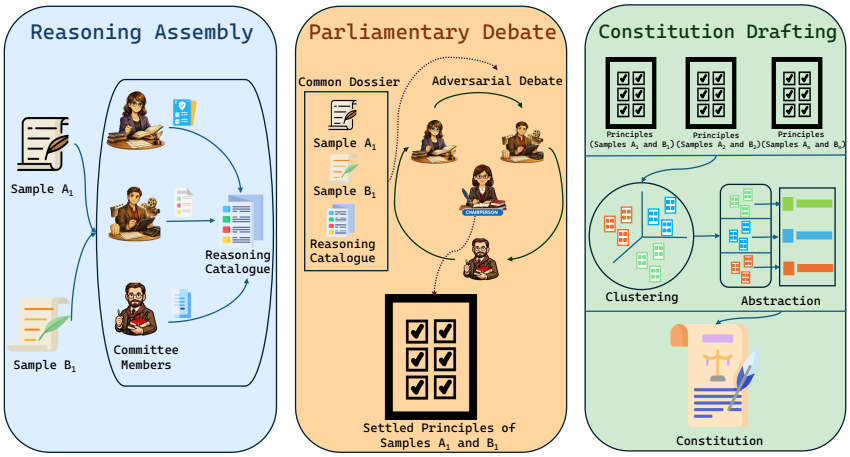


Figure 1: **Architecture of *Democratic ICAI*.** A committee of domain-expert personas first generates detailed rationales for each preference pair. These rationales are then subjected to an adversarial debate procedure, through which the evaluative principles relevant to each comparison are surfaced.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

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 and 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. Complementary to principle induction, AutoRubric (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

Given a dataset of paired samples (A_i, B_i) with human preference labels, we generate rationales explaining why the selected option is preferred. These rationales are produced by a committee of P expert personas, designed to represent expertise relevant to the task. To further diversify the produced rationales, each persona is paired with a distinct reasoning strategy. These strategies encourage complementary modes of reasoning (e.g., step-by-step decomposition, self-consistency, or critical reflection), improving coverage of the reasoning space and reducing reliance on any single heuristic. As a result, personas vary in *what* each member treats as a signal of quality, while strategies shape *how* each member arrives at a justification.

For each pair (A_i, B_i), every persona independently generates a rationale conditioned on the preference label using its assigned reasoning strategy. The resulting set of rationales is collected as

$$R_i = \{r_i^{(1)}, r_i^{(2)}, \ldots, r_i^{(n)}\}, \tag{1}$$

forming a reasoning catalogue, which is then passed to the parliamentary debate stage.

3.2 Parliamentary Debate

The assembled rationales capture diverse perspectives but may also contain redundancy, superficial reasoning or conflicting interpretations. We refine them through a structured debate that extracts and consolidates evaluative criteria, with a judge enforcing de-duplication and consistency (Liang et al., 2024).

To operationalize this, we introduce a structured adversarial debate mechanism (Foster, 2004; Huang et al., 2025), illustrated in Figure 1. For each pair (A_i, B_i), the reasoning assembly (Section 3.1) proceeds to a structured debate phase. In each round, personas challenge, defend, and refine the rationales. While a judge keeps arguments grounded in the sample content, removes weak or redundant justifications, and consolidates overlapping criteria. After T rounds, the judge complies the final principles into a set of settled principles

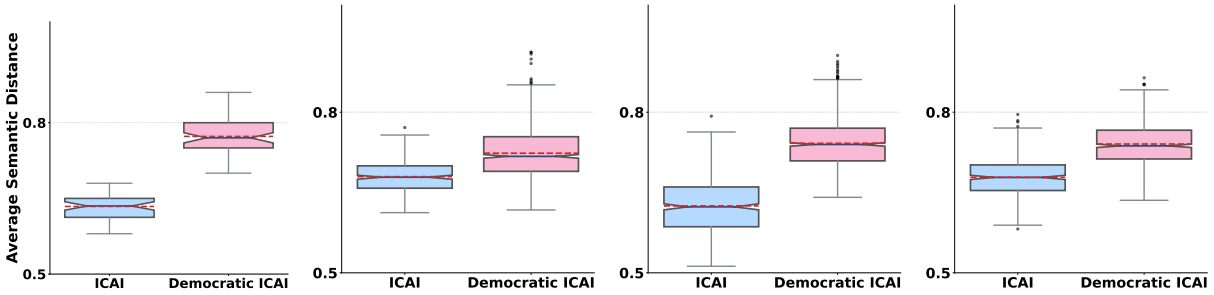
$$Q_i = \left\{Q_i^{(1)}, Q_i^{(2)}, \ldots\right\}. \tag{2}$$

3.3 Constitution Drafting

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

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

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



(a) Long Stories (b) Hypothesis Generation (c) Metaphors Generation (d) Creative Problem Solving

Figure 2: **Distribution of average semantic distance between principles within a constitution.** For each method, the distance is computed as the average cosine distance of a principle from all other principles in the constitution. Lower values indicate reduced diversity (narrower constitutional scope), while higher values reflect greater conceptual separation and normative breadth.

To transform this large collection into a compact, generalized constitution, we adopt 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 $C_j \subseteq Q^{\text{all}}$ contains principles with similar semantic intent. Next, for each cluster C_j, we generate an abstract representative principle $\tilde{Q}_j$ that captures its shared rationale 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 defined as:

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

where each $\tilde{Q}_j$ is a human-readable principle representing a core dimension of human preference.

3.4 Constitution-Guided Inference

Once the constitution $\mathcal{K}$ is induced, we operationalize it through two complementary judges:

- **LLM-based judge**: The LLM judge is prompted with the constitution and the pair, and asked to adjudicate between the two samples according to the listed principles, following the paradigm of (Findeis et al., 2025). Comparing its judgments against the human preference labels measures how well constitution captures the underlying preference structure under language-based application of the principles.
- **Decision Tree Judge**: The LLM judge applies principles holistically and offers no audit trail at the level of individual criteria. To complement it, we operationalize decision tree judge. For each pair (A_i, B_i), we construct a feature table by scoring both samples against every principle $\tilde{Q}_j \in \mathcal{K}$ on a 1–5 rubric. We train a decision tree on the feature table to predict the human preference y_i, and use the trained tree as final judge. We acknowledge the relatively higher cost of decision-tree judge and it is intended as an interpretability-focused alternative rather than the default inference mechanism. Additional details on training and inference using decision trees are provided in the Appendix (Section I).

4 Experimental Setup

4.1 Dataset and Models

We conduct our experiments primarily using the MuCE-Pref dataset (Ismayilzada et al., 2025), which captures preference judgments across a diverse range of creative tasks. We list the MuCE tasks in Table 10. For each task, we sample 500 training pairs to derive task-specific constitutions, ensuring that each pair contains unique accepted and rejected responses. Evaluation is performed on the official test split for each category, allowing us to assess whether the learned principles generalize to unseen creative examples. Additionally, we include LiTBench (Fein et al., 2025) to complement our evaluation with long-form creative writing, particularly story generation. For this setting, we use 1,000 pairs for constitution induction and 2,000 held-out pairs for evaluation.

We use two models: GPT-4o (Hurst et al., 2024) and GPT-5 (Singh et al., 2025) in our experiments.

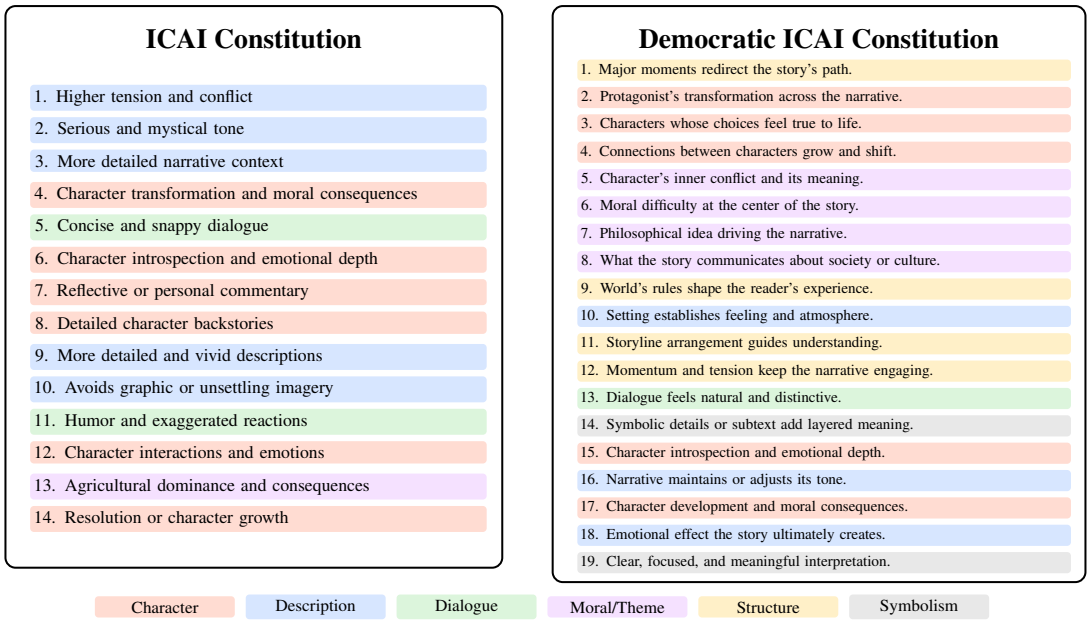


Figure 3: **Qualitative comparison of constitution for Stories (GPT-4o)**. ICAI repeatedly emphasizes overlapping character-growth, emotion, reflection, and description criteria. In contrast, Democratic ICAI distributes principles across character, moral reasoning, narrative structure, setting, dialogue, symbolism, tone, and interpretation, indicating broader coverage.

Task	CoT	CoT-SC	ToT	Self-Refine	AutoRubric	ICAI	DICAI
Alternate Uses of Objects	**78.61**	70.30	76.46	73.78	58.70	66.40	74.23
Consequences	48.29	47.24	47.40	51.10	52.44	61.20	**73.21**
Design Solutions	49.25	42.29	40.20	38.61	45.05	75.20	**75.26**
Experiment Design	65.66	63.31	62.41	64.53	55.46	73.80	**79.93**
Hypothesis Generation	56.10	54.96	55.42	52.56	54.92	64.60	**77.22**
Metaphors	40.00	60.00	60.00	60.00	60.00	71.40	**74.01**
Real-Life Creative Problem Solving	38.07	35.99	35.60	32.01	42.39	56.00	**70.03**
Research Questions	56.80	50.78	53.55	53.79	53.55	70.20	**80.21**
Long Stories (LiTBench)	70.69	66.15	69.28	**71.90**	59.22	62.89	68.70
Short Stories (MuCE)	71.79	67.09	68.75	67.50	69.51	71.40	**78.22**
Average	57.53 ± 13.15	55.81 ± 10.92	56.91 ± 12.46	56.58 ± 13.06	55.12 ± 7.31	67.31 ± 5.82	$\mathbf{75.10 \pm 3.69}$

Table 1: **Preference accuracy (%) across tasks for DICAI and baseline methods using LLM as a judge (GPT-4o).** Bold values indicate the best performance per task.

4.2 Baselines

We compare against two classes of baselines. First, **Deliberative prompting methods** include **Chain-of-Thought (CoT)** (Wei et al., 2022), which prompts the LLM judge to produce step-by-step reasoning before issuing a preference label; **Self-Consistency (CoT-SC)** (Wang et al., 2022), which samples multiple reasoning paths and aggregates their outputs via majority vote; **Tree of Thought (ToT)** (Yao et al., 2023), which formulates preference prediction as a branching search over intermediate judgments; and **Self-Refine** (Madaan et al., 2023), which iteratively improves an initial prediction via self-critique.

Second, **Principle-based methods** include **ICAI** (Findeis et al., 2025), which induces natural-language principles from preference pairs and uses them for prediction in a single pass, and **AutoRubric** (Xie et al., 2025), which extracts rubrics from preference data and applies them through an LLM judge to produce preference labels.

4.3 Training Details

We first generate candidate reasoning traces and principles using a reasoning committee composed of $P = 3$ specialized personas tailored to the target domain. Each agent is assigned one reasoning strategy: Chain-of-Thought, Self-Refine, or Self-Consistency, to guide the generation process. The corresponding prompts for these strategies are provided in the Appendix (Figures 10, 11, and 12). Following that the principles are refined through a structured multi-agent debate implemented with

Task	CoT	CoT-SC	ToT	Self-Refine	AutoRubric	ICAI	DICAI
Alternate Uses of Objects	74.94	70.77	**77.72**	75.35	59.22	69.40	73.22
Consequences	54.56	50.33	52.27	52.51	56.77	67.80	**75.23**
Design Solutions	46.27	43.50	43.56	45.05	45.88	79.80	**80.21**
Experiment Design	66.22	64.96	64.87	62.52	57.12	73.40	**81.77**
Hypothesis Generation	56.10	55.16	54.13	48.43	56.79	66.40	**72.80**
Metaphors	60.00	60.00	60.00	60.00	62.33	74.40	**75.22**
Real-Life Creative Problem Solving	38.37	34.33	35.18	32.11	43.65	64.20	**77.05**
Research Questions	58.12	57.28	56.96	54.87	58.76	71.80	**76.20**
Long Stories (LiTBench)	70.69	68.43	**71.29**	68.02	63.21	66.80	71.20
Short Stories (MuCE)	75.64	73.42	73.41	67.09	72.22	**77.60**	77.00
Average	60.09 ± 11.51	57.82 ± 11.83	58.94 ± 12.74	56.59 ± 12.08	57.59 ± 7.78	71.16 ± 4.86	$\mathbf{75.99 \pm 3.09}$

Table 2: **Preference accuracy (%) across tasks for DICAI and baseline methods using LLM as a judge (GPT-5).** Bold values indicate the best performance per task.

AutoGen (Wu et al., 2024) (prompts shown in Figures 20 and 21). The debate runs for three rounds. Following which the judge prepares the settled principles. To further ensure diversity and remove redundancy among principles, we embed all candidates and apply K-Means clustering. The embedding model used in OpenAI's Text-Embedding-3 small. We determine the number of clusters using the silhouette score, which quantifies the trade-off between intra-cluster cohesion and inter-cluster separation. We set $K = K^*$, where K^* denotes the value that maximizes the silhouette score (refer Table 3). Additional experimental details can be found in the Appendix (refer Section J).

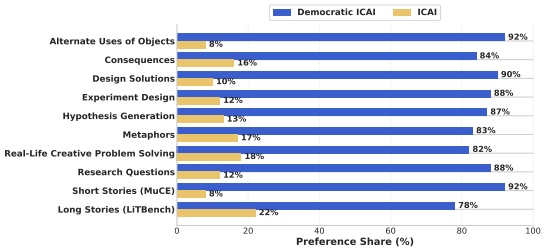
4.4 Evaluation Details

For evaluation, we consider three dimensions:

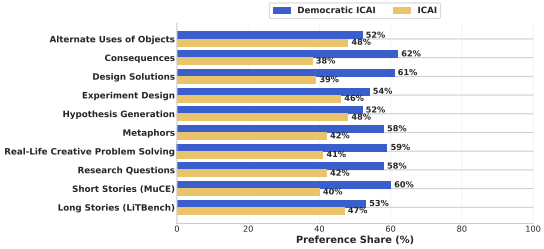
- **Constitution Quality Evaluation.** First, we directly evaluate the quality of the constitutions themselves. Drawing from foundational work in political philosophy, constitutional theory, and recent AI alignment literature, we assess constitutions along five dimensions. Following Lon Fuller's *The Morality of Law* (Fuller, 1969), and subsequent work by Nwokora (Nwokora, 2022) and Hedling (Hedling, 2023), we evaluate constitutions along five dimensions: **generality**, **clarity**, **coherence**, **feasibility**, and **faithfulness**. An LLM judge is prompted with both constitutions and asked to compare them pairwise along each of these dimensions. To mitigate circularity and model-specific bias, we perform this evaluation using three different models Qwen2.5-32B (Qwen et al., 2025), GPT-4o, and GPT-5. The prompt for the LLM annotator is presented in Figure 23. Detailed definition of these dimension are provided in Section H in the Appendix.
- **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.
- **Bias and spurious-criteria audit.** To assess whether induced constitutions encode harmful biases or reward superficial properties unrelated to genuine quality, we audit each principle along two axes using Qwen2.5-32B as an external auditor. The two axes are: (i) *demographic and stereotypical bias*, covering gender, race, ethnicity, age, religion, nationality, socioeconomic class, disability, sexual orientation, and culture; and (ii) *spurious or superficial criteria*, covering length, verbosity, formatting, and stylistic mannerisms unrelated to genuine quality. Each principle is flagged on both axes with a severity rating (high, medium, low), and we report the resulting flag rates per constitution in Table 4. Prompt for the same is provided in the Appendix (refer Figure 30)

Task	DICAI	ICAI
Alternate Uses of Objects	1050 (0.1411)	750 (0.1165)
Consequences	850 (0.0929)	600 (0.0900)
Design Solutions	500 (0.0959)	650 (0.0970)
Experiment Design	1000 (0.0851)	600 (0.0800)
Hypothesis Generation	700 (0.1048)	800 (0.0900)
Metaphors	850 (0.1409)	750 (0.1290)
Real-Life Creative Problem Solving	850 (0.0899)	700 (0.0850)
Research Questions	700 (0.0871)	600 (0.0850)
Stories	600 (0.0947)	450 (0.0830)

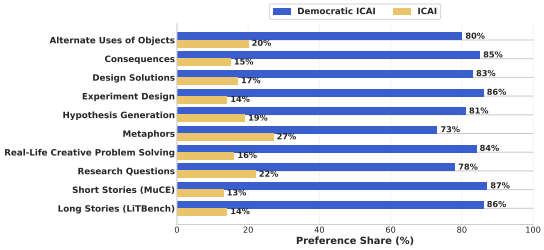
Table 3: **Optimal number of clusters (K^*).** The values in parentheses denote the corresponding peak silhouette scores obtained at K^*, measuring clustering quality in terms of intra-cluster cohesion and inter-cluster separation. Higher scores indicate better-defined and more separable clusters.



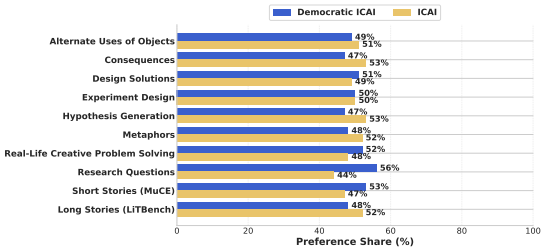
(a) Generality



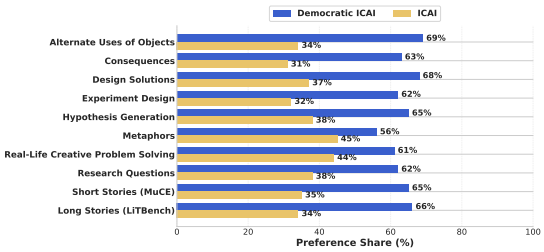
(b) Clarity



(c) Coherence



(d) Feasibility



(e) Faithfulness

Figure 4: **Comparison of Democratic ICAI and ICAI across five dimensions using Qwen-2.5-32B.** 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.

5 Results and Analyses

1. **Democratic ICAI produces relatively more informative, interpretable, and robust constitutions:** Across all tasks and under both GPT-4o and GPT-5, Democratic ICAI achieves higher preference accuracy than ICAI and AutoRubric (refer Tables 1, 2, and Figure 3), while remaining competitive with deliberative prompting baselines such as CoT, ToT, and Self-Refine. Although it does not outperform all baselines on every task, Democratic ICAI achieves the highest average preference accuracy across tasks and exhibits the lowest standard deviation under both models. The gains are most pronounced on tasks, where single-pass explanations often fail to capture or adequately balance competing rationales, such as Research Questions and Real-Life Creative Problem Solving.
   These improvements are further supported by an independent Decision Tree judge, where Democratic ICAI achieves higher preference accuracy (refer Figure 29 in the Appendix), suggesting that the gains are not tied to a specific judge. Finally, LLM judges consistently prefer Democratic ICAI-generated constitutions across qualitative dimensions; including generality, clarity, coherence, and faithfulness; highlighting the interpretability and robustness of the induced principles (refer Figure 4).
2. **Cognitively aligned principles enable the construction of diverse constitutions:** Democratic ICAI constructs principles through adversarial debate, where agents justify competing preferences. This mirrors the multi-dimensional reasoning humans employ when evaluating creative outputs, enabling more cognitively aligned principle formation (Huang et al., 2025). As a result, it produces more diverse constitutions: embeddings show broader dispersion, larger interquartile ranges, and more pronounced outliers compared to ICAI (refer Figures 2 and 28). This diversity reflects richer conceptual coverage and

Constitution	Bias flag	Spurious flag	Bias (high/med)	Spurious (high/med)
Alternate Uses of Objects	0.19%	1.62%	0/2	0/17
Consequences	0.00%	2.47%	0/0	0/21
Design Solutions	1.8%	2.2%	0/9	0/11
Experiment Design	0.00%	0.00%	0/0	0/0
Hypothesis Generation	1.71%	4.57%	0/12	0/32
Metaphors Generation	0.04%	3.11%	0/3	0/23
Real-Life Creative Problem Solving	0.00%	0.00%	0/0	0/0
Research Questions	0.00%	0.04%	0/0	0/4
Long Stories (LiTBench)	0.00%	1.25%	0/0	0/0
Short Stories (MuCE)	0.5%	3.40%	0/3	0/17

Table 4: **Bias and spurious-criteria audit of D-ICAI constitutions**. We using Qwen2.5-32B as an external auditor to avoid model bias

Task	ICAI	1 round	2 rounds	3 rounds
Research Questions	70.20	78.22	80.01	80.21
Metaphors	71.40	73.21	73.80	74.80
Long Stories (LiTBench)	62.89	67.63	68.00	68.70

Table 5: Ablation on debate rounds. We vary the number of debate rounds (1, 2, 3) within the DICAI pipeline and compare against ICAI.

the emergence of distinctive principles beyond the dominant semantic cluster, improving generalization (refer Figure 4(a) and 22(a) in the Appendix).

3. **Democratic ICAI achieves relatively lesser lossy compression:** Democratic ICAI exhibits reduced lossy compression by inducing a more diverse and semantically broad set of principles. Its optimal silhouette scores occur at higher cluster counts, indicating support for a larger number of well-separated conceptual groups. It also achieves higher peak silhouette values across tasks, suggesting that its principles are both more coherent and more discretely structured (refer Table 3). This distinction is evident in Figure 3, where Democratic ICAI's principles span a wider range of mechanisms, while ICAI's are more overlapping and narrowly focused. Consistent with this, LLM annotators rate Democratic ICAI higher in clarity, indicating that increased semantic diversity translates into more interpretable and better-aligned principles.
4. **Complex decision-making processes are harder to compress:** Democratic ICAI induces constitutions that are more diverse (refer 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). 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 feasibility compared to simpler heuristics as can be seen in Figure 4(d), it improves the ability to capture overall quality, as reflected in LLM-based evaluations. In contrast, ICAI's compact principles emphasize simplicity at the cost of expressive fidelity.
5. **Induced principles do not encode harmful bias or reward spurious properties:** An audit of the induced constitutions, using Qwen2.5-32B as an external auditor, reports no high-severity flags on either the demographic-bias or spurious-criteria axis (refer Table 4). The rare medium-severity flags correspond to protective guardrails (e.g., avoiding culturally marked framing) rather than encoded stereotypes, while low-severity spurious flags reflect creative-writing trade-offs such as tonal or stylistic preferences that genuinely correlate with quality (refer Table 11 in the Appendix). Because the induced principles are human-readable, they can be directly inspected and edited before deployment, offering a critical mitigation absent in black-box reward models.
6. **Both adversarial debate and persona diversity are necessary for the observed gains:** Ablations isolate the contribution of each component (refer Tables 5 and 6). Removing the debate stage and extracting principles directly from rationales reduces performance to the level of ICAI, even when using multiple personas. While a single persona without debate performs even worse. In contrast, the full DICAI pipeline outperforms these variants, indicating that gains arise from the combination of adversarial debate and persona diversity rather than either component alone. Within the debate stage, most improvements emerge early: even a single round substantially improves over ICAI, with diminishing returns in later rounds. This suggests that DICAI achieves strong performance without requiring extended multi-round debate, maintaining computational efficiency.

6 Conclusion

Democratic ICAI provides a stronger framework for modeling and structuring the reasoning underlying human preferences through multi-persona deliberation, adversarial refinement, and abstraction. It produces constitutions that are more diverse, expressive, and better aligned with human judgment than ICAI, while achieving higher average performance and lower variance across tasks compared to ICAI and AutoRubric. Gains are most evident in tasks requiring multi-criteria judgments, where single-pass methods tend to underperform. Overall, Democratic ICAI more faithfully captures the evaluative structure of complex human decisions rather than reducing them to a single dominant heuristic.

7 Limitations

Explicit ground-truth principles underlying human judgments are often unavailable, making it difficult to determine whether learned principles are fully accurate or complete. As a result, evaluation relies on indirect proxy measures rather than direct validation against human intent. Moreover, preference data may encode systematic biases, which can propagate into induced principles and influence downstream behavior. Democratic ICAI mitigates some of these risks by producing human-readable principles that support inspection and iterative refinement. However, this transparency does not eliminate bias, and expert oversight remains necessary to identify subtle or systemic issues and ensure alignment with intended values.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

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

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

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

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

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2025. Litbench: A benchmark and dataset for reliable evaluation of creative writing. *arXiv preprint arXiv:2507.00769*.

Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert Mullins. 2025. Inverse constitutional ai: Compressing preferences into principles. *arXiv preprint arXiv:2406.06560*.

David E Foster. 2004. 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).

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

Lon L. Fuller. 1969. *The Morality of Law: Revised Edition*. Yale University Press. Accessed: 14 May 2026.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Nora Hedling. 2023. *A Practical Guide to Constitution Building*. International IDEA.

Yuesheng Huang, Meiqi Feng, Zhenming He, Yueyuan Peng, and Jiawen Li. 2025. 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*, pages 41–50.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

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

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, and 1 others. 2024. 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.

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

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

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, and 1 others. 2024. Rrm: Robust reward model training mitigates reward hacking. *arXiv preprint arXiv:2409.13156*.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

Zim Nwokora. 2022. Constitutional design for dynamic democracies: A framework for analysis. *International Journal of Constitutional Law*, 20(2):580–610.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

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

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.

Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, and 1 others. 2025. Auto-rubric: Learning from implicit weights to explicit rubrics for reward modeling. *arXiv preprint arXiv:2510.17314*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822. Curran Associates, Inc.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Appendix

A Ethical Considerations and Societal Implications **12**

B Constitutions **12**
- B.1 Constitution generated with ICAI (GPT-4o) on LitBench Stories . . 12
- B.2 Constitution generated with Democratic ICAI (GPT-4o) on LitBench Stories 12
- B.3 Constitution generated with ICAI (GPT-5) on LitBench Stories . . . 13
- B.4 Constitution generated with Democratic ICAI (GPT-5) on LitBench Stories 14
- B.5 Constitution generated with ICAI (GPT-4o) on the MuCE dataset on the "Hypothesis Generation" task . 14
- B.6 Constitution generated with Democratic ICAI (GPT-4o) on the MuCE dataset on the "Hypothesis Generation" task 15
- B.7 Constitution generated with ICAI (GPT-4o) on the MuCE dataset on the "Metaphors" task 15
- B.8 Constitution generated with Democratic ICAI (GPT-4o) on the MuCE dataset on the "Metaphors" task . 16

C Ablation Studies **16**

D Reasoning Assembly Persona **16**
- D.1 Task wise committee 17
- D.2 Screenwriter Expert Persona Prompt 17
- D.3 Professor of Literature Expert Persona Prompt 18
- D.4 Literary Critic Expert Persona Prompt 18
- D.5 Innovation Consultant Expert Persona Prompt 19
- D.6 Design Engineer Expert Persona Prompt 20

E Reasoning Assembly Prompts **21**
- E.1 Chain of Thought Prompt 21
- E.2 Self Refine Prompt 21
- E.3 Self Consistency Prompt 22

F Deliberative Prompting Method Prompts **22**
- F.1 Chain of Thought Prompt 22
- F.2 Chain of Thought Self Consistency Prompt 23
- F.3 Tree of Thought Prompts 23
- F.4 Self Refine Prompts 24

G Adversarial Debate Prompts **26**
- G.1 Judge System Prompt 26
- G.2 Debate Participant System Prompt 26

H Definition of five dimension to evaluate quality of Constitution **27**

I Constitution-Based Inference with Decision Tree Judge **33**

J Dataset and Baselines **33**

K Licenses of Artifacts **34**

A Ethical Considerations and Societal Implications

Our work aims to improve the transparency and usability of preference data, which plays a central role in training and evaluating large language models. By inducing human-readable constitutions, Democratic ICAI offers a more interpretable alternative to black-box reward models, enabling practitioners to understand, inspect, and refine the principles guiding model behavior. This increased transparency can support better accountability and more informed decision-making in downstream applications.

At the same time, the use of preference data raises important ethical considerations. Biases present in underlying datasets can be reflected in induced principles and subsequently influence model behavior. Democratic ICAI partially addresses this challenge by producing explicit, editable principles, allowing practitioners to identify and revise undesirable patterns before deployment. This visibility distinguishes it from approaches where such biases remain hidden within opaque models.

We recommend that generated constitutions be treated as editable artifacts rather than fixed outputs. Careful review is essential to ensure alignment with intended values and to mitigate the risk of reinforcing harmful or unintended biases. When used responsibly, we believe Democratic ICAI can contribute to more interpretable, controllable, and accountable AI systems.

B Constitutions

B.1 Constitution generated with ICAI (GPT-4o) on LitBench Stories

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.

B.2 Constitution generated with Democratic ICAI (GPT-4o) on LitBench Stories

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 resonates with characters whose choices feel true to life.
4. Select the response that reveals how connections between characters grow and shift.

5. Select the response that brings forward the character's inner conflict and its meaning.
6. Select the response that communicates the moral difficulty at the center of the story.
7. Select the response that draws out the deeper philosophical idea driving the narrative.
8. Select the response that reflects what the story communicates about its society or culture.
9. Select the response that explains how the world's rules shape the reader's experience.
10. Select the response that conveys how the setting establishes feeling and atmosphere.
11. Select the response that shows how the arrangement of the storyline guides understanding.
12. Select the response that captures how momentum and tension keep the narrative engaging.
13. Select the response that brings attention to dialogue that feels natural and distinctive.
14. Select the response that highlights symbolic details or subtext adding layered meaning.
15. Select the response that explores character introspection and emotional depth.
16. Select the response that expresses how the narrative maintains or adjusts its tone.
17. Select the response that focuses on the character's development and the moral consequences of their actions.
18. Select the response that conveys the emotional effect the story ultimately creates.
19. Select the response that offers the most clear, focused, and meaningful interpretation.

B.3 Constitution generated with ICAI (GPT-5) on LitBench Stories

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.

B.4 Constitution generated with Democratic ICAI (GPT-5) on LitBench Stories

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.

B.5 Constitution generated with ICAI (GPT-4o) on the MuCE dataset on the "Hypothesis Generation" task

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.

B.6 Constitution generated with Democratic ICAI (GPT-4o) on the MuCE dataset on 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.

B.7 Constitution generated with ICAI (GPT-4o) on the MuCE dataset on the "Metaphors" 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.

B.8 Constitution generated with Democratic ICAI (GPT-4o) on the MuCE dataset on 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.

C Ablation Studies

A single persona without MAD underperforms even ICAI, while adding personas without debate only recovers ICAI-level performance; the full DICAI pipeline yields 6–8 point gains over the three-persona variant, attributing the improvement to the combination of adversarial debate and persona diversity (Table 6).

A single round of debate already substantially exceeds ICAI, with returns saturating at 2 rounds across all three tasks, indicating that the benefit is front-loaded and DICAI remains compute-efficient relative to longer multi-round setups (Table 5).

Task	ICAI	1 Persona w/o MAD	3 Personas w/o MAD	DICAI
Research Questions	70.20	66.22	72.77	80.21
Metaphors	71.40	68.27	70.01	74.01
Long Stories (LiTBench)	62.89	59.22	62.13	68.70

Table 6: Ablation on MAD and persona count. We remove the debate stage and extract principles directly from rationales, varying the number of personas (single vs. three) without MAD, and compare against ICAI and the DICAI pipeline.

D Reasoning Assembly Persona

For reasoning assembly, we employ three expert personas per task. The complete list of personas is provided in Table 8. Example persona prompts are shown in Figures 5, 6, and 7.

Task	Committee Experts
Real-Life Creative Problem Solving	Innovation Consultant / Design Thinking Expert; Organizational Psychologist; Business Strategy Expert
Metaphors	Cognitive Linguist; Poet / Creative Writer; Professor of Literature
Alternate Uses of Objects Task	Creativity Researcher (Divergent Thinking Specialist); Product / Industrial Designer; Cognitive Psychologist
Design Solutions	UX / Product Designer; Design Engineer; Innovation Consultant / Design Thinking Expert
Stories	Screenwriter; Professor of Literature; Literary Critic
Consequences Task	Futurist / Scenario Planner; Cognitive Psychologist (Counterfactual Thinking); Science Fiction Writer
Experiment Design	Experimental Scientist; Research Methods / Statistics Professor; Principal Investigator (PI)
Hypothesis Generation	Domain Scientist; Theory / Conceptual Framework Researcher; Science Writer / Scientific Communicator
Research Questions	Principal Investigator (PI); Journal Editor or Academic Reviewer; Research Methods Professor

Table 8: Task-wise assignment of committee experts used for reasoning.

D.1 Task wise committee

D.2 Screenwriter Expert Persona Prompt

Screenwriter Expert 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 5: Screenwriter expert persona prompt for reasoning and debate agents

D.3 Professor of Literature Expert Persona Prompt

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 6: Professor of literature expert persona prompt for reasoning and debate agents

D.4 Literary Critic Expert Persona Prompt

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

D.5 Innovation Consultant Expert Persona Prompt

Innovation Consultant Expert Persona

Age: 38
Sex: Female
Race: Asian
Ancestry: Korean
Household Language: English and Korean
Education: Master's degree
Employment Status: Employed
Class Of Worker: Private wage and salary worker
Industry Category: Innovation Consulting
Occupation Category: Innovation Consultant / Design Thinking Expert
Detailed Job Description: Works at the intersection of human needs, business viability, and technical feasibility, helping organizations translate ambiguous challenges into validated, scalable solutions through research, synthesis, and experimentation.
Income: 120000.0
Marital Status: Married
Household Type: Married couple household
Family Presence And Age: Spouse, no children
Place Of Birth: California/CA
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: Very High, Conscientiousness: High, Extraversion: Moderate, Agreeableness: High, Neuroticism: Low
Defining Quirks: Maintains detailed assumption logs and visual problem maps across projects.
Mannerisms: Frames ideas as hypotheses and frequently asks evidence-based questions.
Personal Time: Explores design communities, travels, and experiments with creative side projects.
Lifestyle: Fast-paced and collaborative, focused on rapid learning cycles and measurable impact.
Ideology: Believes in human-centered, ethical, and inclusive design.
Political Views: Moderate
Religion: Not religious

Figure 8: Innovation consultant expert persona prompt for reasoning and debate agents

D.6 Design Engineer Expert Persona Prompt

Design Engineer Expert Persona

Age: 34
Sex: Male
Race: White alone
Ancestry: European
Household Language: English only
Education: Bachelor's degree (Mechanical Engineering)
Employment Status: Employed
Class Of Worker: Private wage and salary worker
Industry Category: Manufacturing / Product Development
Occupation Category: Design Engineer
Detailed Job Description: A mid-career design engineer specializing in electro-mechanical product development, working across CAD modeling, simulation, supplier coordination, and shop-floor validation to translate ambiguous requirements into manufacturable, reliable designs under cost and regulatory constraints.
Income: 95000.0
Marital Status: Married
Household Type: Married couple household
Family Presence And Age: Spouse and one child
Place Of Birth: Texas/TX
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: Moderate, Conscientiousness: Very High, Extraversion: Low, Agreeableness: Moderate, Neuroticism: Low
Defining Quirks: Keeps detailed tolerance stack-up sheets and sanity-checks all designs with quick hand calculations.
Mannerisms: Breaks problems into first-principles reasoning and frequently sketches mechanisms while explaining.
Personal Time: Enjoys DIY mechanical projects, reading engineering forums, and experimenting with CAD designs.
Lifestyle: Structured and detail-oriented, balancing office work with hands-on prototyping and shop-floor interactions.
Ideology: Believes good engineering should be practical, reliable, and grounded in real-world constraints.
Political Views: Moderate
Religion: Not religious

Figure 9: Design engineer expert persona prompt for reasoning and debate agents

E Reasoning Assembly Prompts

E.1 Chain of Thought Prompt

As the persona you are embodying, privately think through your preference step-by-step.

Evaluate the outputs according to your persona's worldview, values, communication style, and priorities.

Consider various factors to reach your reasoning.

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):

```
{"final_reasoning": "<one concise paragraph written in the persona's voice>"}
```

Figure 10: Reasoning assembly prompt using chain of thought strategy.

E.2 Self Refine Prompt

As the persona you are embodying, privately generate three different reasoning paths from distinct perspectives that your persona might consider.

Privately compare these paths and determine the most consistent justification, based on your persona's worldview.

IMPORTANT:

Do not reveal the three paths or any internal deliberations.

After completing this internal process, output ONLY valid JSON in this exact schema (no mark-

down, no extra text):

```
{"final_reasoning": "<one concise paragraph written in the persona's voice>"}
```

Figure 11: Reasoning assembly prompt using reflective justification strategy.

E.3 Self Consistency 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):

```
{"final_reasoning": "<one concise paragraph written in the persona's voice>"}
```

Figure 12: Reasoning assembly prompt using self consistency strategy.

F Deliberative Prompting Method Prompts

F.1 Chain of Thought Prompt

You are a careful preference judge. You will compare two candidate responses to the same user request.

Output JSON only.

Task Category: {task}

User Request: {inp}

Response A: {a}

Response B: {b}

Instructions:

- Think step-by-step internally, but DO NOT reveal your full reasoning.
- Output MUST be valid JSON only, with keys:

`choice`: "A" or "B"

`justification`: string (2–4 sentences max)

Figure 13: Chain of Thought Prompt

F.2 Chain of Thought Self Consistency Prompt

You are a careful preference judge. You will compare two candidate responses to the same user request.

You must NOT reveal hidden chain-of-thought. Output JSON only.

As the preference judge, privately generate {k_paths} different reasoning paths from distinct perspectives.

Privately compare these paths and determine the most consistent justification, based on your worldview.

IMPORTANT:

Do not reveal the {k_paths} paths or internal deliberations.

Task Category: {task}

User Request: {inp}

Response A: {a}

Response B: {b}

{sc_instructions}

After completing the internal process, output ONLY valid JSON in this exact schema (no markdown, no extra text):

```
{ "choice": "A" or "B", "final_reasoning": "one concise paragraph in your judge persona voice (2-5 sentences)" }
```

Figure 14: Chain of Thought Self Consistency Prompt

F.3 Tree of Thought Prompts

F.3.1 Thought Generation Prompt

Generate {k} distinct candidate preference decisions ("thoughts").

Each thought must be complete: choose A or B and provide a brief justification (2–4 sentences).

Make the thoughts meaningfully different by emphasizing different rubric angles.

IMPORTANT:

- Do NOT reveal chain-of-thought or step-by-step reasoning.

- Output ONLY valid JSON (no markdown, no extra text).

Schema:

```
{{ "thoughts": [ {{"id": 1, "choice": "A" or "B", "justification": "..." }},
..., {{"id": {k}, "choice": "A" or "B", "justification": "..." }} ] }}
```

Task Category: {task}

User Request: {inp}

Response A: {a}

Response B: {b}

{rubric}

Figure 15: Tree of Thought Thought Generation Prompt

F.3.2 Thought Evaluation Prompt

You are the state evaluator V in a Tree-of-Thoughts search.

Score each candidate thought by how well the decision aligns with the rubric and how well-supported it is.

Return ONLY valid JSON:

```
{{ "scores": [ {{"id": <id>, "score": 1-10, "note": "one short sentence"}} ]
}}
```

Candidate thoughts: {thoughts_json}

Task Category: {task}

User Request: {inp}

Response A: {a}

Response B: {b}

{rubric}

Figure 16: Tree of Thought Thought Evaluation Prompt

F.4 Self Refine Prompts

F.4.1 Initial Generation Prompt

Choose which response a typical user would prefer overall.

Return ONLY valid JSON (no markdown, no extra text) in this schema:

```
{{ "choice": "A" or "B", "justification": "2-4 sentences explaining your choice"
}}
```

Task Category: {task}

User Request: {inp}

Response A: {a}

Response B: {b}

Figure 17: Self Refine Initial Generation Prompt

F.4.2 Feedback Prompt

You are the FEEDBACK module in an iterative SELF-REFINE loop.

Provide actionable feedback to improve the decision quality and calibration.

Evaluate the decision along these aspects:

- Helpfulness (0–5)

- Correctness (0–5)

- Completeness (0–5)

- Clarity (0–5)

- Safety (0–5)

- Overall decision quality (0–5)

Rules:

- Be specific: mention what evidence in A/B supports or contradicts the current choice.

- Provide actionable suggestions (what to reconsider, what criteria were underweighted/overweighted).

- Decide whether the decision is already strong enough to stop.

Return ONLY valid JSON in this schema:

```
{{ "scores": {{ "helpfulness": 0-5, "correctness": 0-5, "completeness": 0-5,
"clarity": 0-5, "safety": 0-5, "overall": 0-5 }}, "total": 0-30, "issues":
["..."], "suggestions": ["..."], "stop": true or false }}
```

Task Category: {task}

User Request: {inp}

Response A: {a}

Response B: {b}

Current decision JSON: {decision_json}

Figure 18: Self Refine Feedback Prompt

F.4.3 Refinement Prompt

You are the REFINE module in an iterative SELF-REFINE loop.

Use the feedback to produce an improved final decision.

Return ONLY valid JSON (no markdown, no extra text) in this schema:

```
{{ "choice": "A" or "B", "justification": "2-4 sentences explaining your choice"
}}
```

Task Category: {task}

User Request: {inp}

Response A: {a}

Response B: {b}

Current decision JSON: {decision_json}

Feedback JSON: {feedback_json}

Figure 19: Self Refine Refinement Prompt

G Adversarial Debate Prompts

G.1 Judge System Prompt

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.

G.2 Debate Participant System Prompt

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.

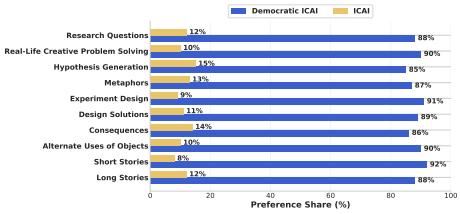
OUTPUT FORMAT (STRICT; no extra text):

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 >

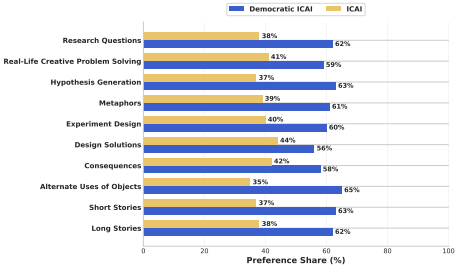
Figure 21: Debater system prompt for parliamentary debate.

H Definition of five dimension to evaluate quality of Constitution

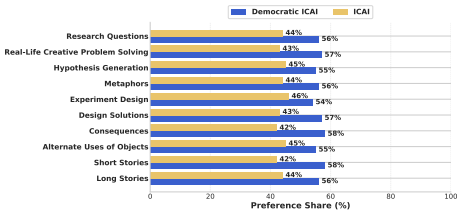
- **Generality**: The extent to which a principle applies broadly across diverse contexts and scenarios, rather than being tied to specific examples or dataset artifacts.
- **Clarity**: The degree to which the wording and intent of the principle are interpretable, precise, and unambiguous for AI systems.
- **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.
- **Faithfulness**: The degree to which the principle preserves its intended meaning and resists distortion, misinterpretation, or application inconsistent with its original intent.



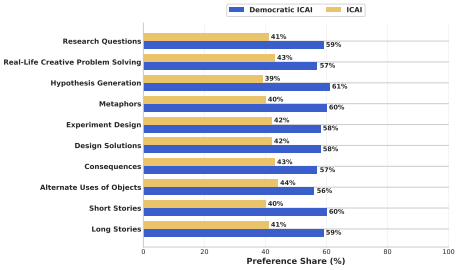
(a) GPT-4o Generality



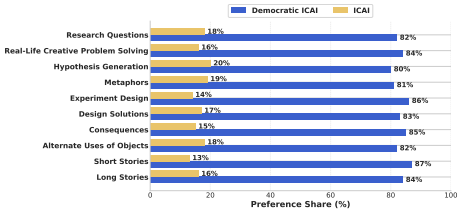
(b) GPT-5 Generality



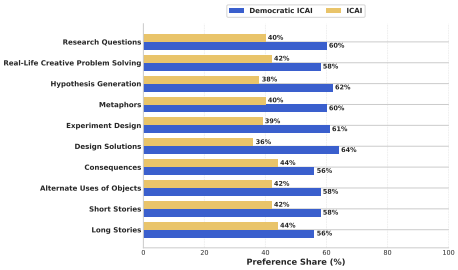
(c) GPT-4o Clarity



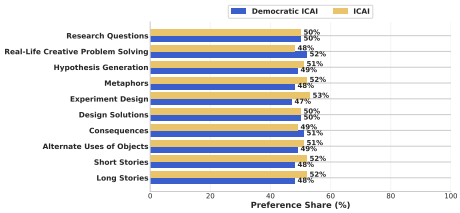
(d) GPT-5 Clarity



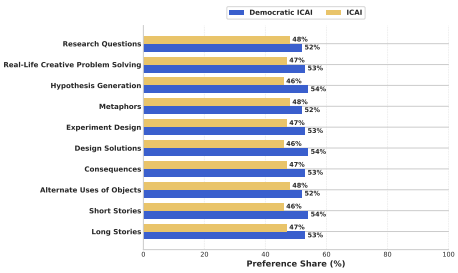
(e) GPT-4o Coherence



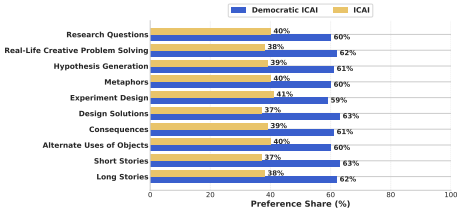
(f) GPT-5 Coherence



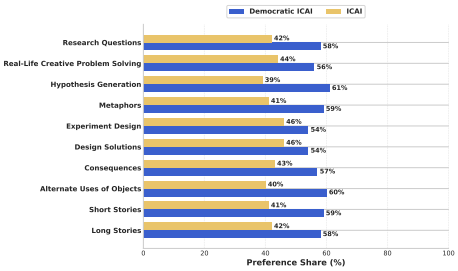
(g) GPT-4o Feasibility



(h) GPT-5 Feasibility



(i) GPT-4o Faithfulness



(j) GPT-5 Faithfulness

Figure 22: Comparison of Democratic ICAI and ICAI across five dimensions under GPT-4o (left column) and GPT-5 (right column). Each row corresponds to a qualitative dimension evaluated across ten tasks. Democratic ICAI consistently outperforms ICAI on structural criteria such as generality and coherence, while remaining competitive on feasibility.

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) Constitution B (Rules)
Task
Response A {Response_A}
Response B {Response_b}
Task

1. Evaluate **Response A** and **Response B** using the following five criteria as they are defined or implied by each constitution:
   - **Generality**: How broadly the constitution's principles apply across diverse contexts and scenarios (not tied to narrow examples or dataset artifacts).
   - **Clarity**: How precise, interpretable, and unambiguous the wording and intent of the principles are for an AI system.
   - **Coherence**: Whether the principles can operate together without contradictions, forming a logically consistent evaluative framework.
   - **Feasibility**: How realistically an AI system can apply the principles when making decisions in practical evaluation settings.
   - **Faithfulness**: The extent to which application of the principles preserves their intended meaning and resists distortion or misinterpretation.
2. Apply **Constitution A**:
   - Decide which is better under Constitution A: **Response A** or **Response B**.
   - Provide **3–6** concise evidence bullets showing how your decision follows from Constitution A, explicitly referencing the five criteria where relevant.
   - Do not invent information beyond the provided responses and constitution rules.
3. Apply **Constitution B**:
   - Decide which is better under Constitution B: **Response A** or **Response B**.
   - Provide **3–6** concise evidence bullets showing how your decision follows from Constitution B, explicitly referencing the five criteria where relevant.
   - Do not invent information beyond the provided responses and constitution rules.
4. Answer the following single-choice questions:
   - Q1 Which response is better under Constitution A? (Response A / Response B)
   - Q2 Which response is better under Constitution B? (Response A / Response B)
   - Q3 Which constitution has clearer rules? (Constitution A / Constitution B) *[Clarity]*
   - Q4 Which constitution is easier to apply in practice? (Constitution A / Constitution B) *[Feasibility]*
   - Q5 Which constitution applies more broadly across contexts? (Constitution A / Constitution B) *[Generality]*
   - Q6 Which constitution is more internally consistent? (Constitution A / Constitution B) *[Coherence]*
   - Q7 Which constitution better preserves its intended meaning when applied? (Constitution A / Constitution B) *[Faithfulness]*

Figure 23: 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 24: 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 25: ICAI prompt for annotating according to constitution (Alpaca Eval variant).

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.

Figure 26: 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 27: Example of a story from LitBench.

I 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:

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

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:

$$y_i \in \{\mathsf{A}, \mathsf{B}\} \tag{10}$$

The prompt for feature table construction is presented in Figure 24.

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 $\{\mathsf{A}, \mathsf{B}\}$.

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.

J Dataset and Baselines

ICAI is tested 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 where, judgments are far less predictable. We test our work mainly on two datasets:

LiTBench. LiTBench (Fein et al., 2025) is a preference benchmark for creative writing evaluation, designed to test whether language models can reliably judge the quality of long-form narrative text. Each instance consists of a short story prompt paired with two candidate completions and a human preference label indicating which completion is preferred. Preferences in LiTBench are derived from number of upvotes the story obtained on the subreddit. We use 1,000 pairs for constitution induction and 2,000 held-out pairs for evaluation.

MuCE. MuCE (Ismayilzada et al., 2025) is a large-scale human-preference dataset spanning a diverse range of creative tasks, curated from psychological creativity assessments. The full dataset contains over 200,000 responses with ratings from more than 30 creativity tasks, of which we evaluate on nine categories that span markedly different notions of creative success: Alternate Uses of Objects, Consequences, Design Solutions, Experiment Design, Hypothesis Generation, Metaphors, Real-Life Creative Problem Solving, Research Questions, and Short Stories. This category-level diversity makes MuCE complementary to LiTBench: while LiTBench focuses on long-form narrative judgment, MuCE tests generalization across short-form creative tasks where different categories privilege different evaluative criteria. For each task we sample 500 unique accepted-rejected response pairs to derive a task-specific constitution and evaluate on the official test split.

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 be the result if everyone suddenly lost the sense of balance and were unable to stay in the upright position for more than a moment?	We would eventually die	motorcycles would no longer be a usable form of transportation
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	Come up with an original and creative experiment to test the following hypothesis or research question: You think students at your school are friendlier than students at most schools. How could you test that hypothesis?	Have a random group with a variety of demographics take a survey of how their experience is with other students at each school. Two different groups of students enrolled at each school	distribute a mandatory, randomized survey
Hypothesis Generation	You just started going to a new school. You notice that more students sing in the hallways at this school than at your old school. What hypotheses do you have about why that is?	this school's tradition	the kids are in the chorus
Metaphors	Finish the sentence with an original and creative metaphor: The sweet candy is....	life is like a box of sweet candy	horrible
Real-Life Creative Problem Solving	Come up with an original and creative solution for the following real-world problem: Clara, a junior pre-med student, is working part-time and taking a 15 hour credit load at school. Clara enjoys her pre-med courses very much, but they are very difficult and time consuming. Her current job as a 'gopher' at an architectural design firm requires her to work 25 hours a week which really cuts into her available study time. In fact, she is barely getting 'C's' in two of the classes she needs for her major. The pay in her present job is good, and she enjoys the work, although she is not getting a lot of practical experience. Clara does not want to drop any of her classes as she needs them to graduate so that she can be admitted to medical school in the coming year. She also knows that she needs good grades to be admitted to medical school. Up until now, Clara has been able to work at her job and still get good grades, but the difficult courses she is taking now require much more of her time. Clara is not sure how to solve her problem..	she should weigh out her priorities. it seems as if money is the only real reason for her staying at her curret job so she should just apply for fafsa or loans and go through school with that money.	Clara needs to cut down on her hours at work because her grades are more important, in order for her to get into med school. If she needs to work those hours for financial reasons then she needs to not have a full load of classes and only go to school part time.
Research Questions	You created the first vehicle that can reach an unexplored part of the ocean floor. What scientific questions could you ask about this part of the ocean floor?	is there life down there?	does this part of the ocean floor receive any light from above? if so, how much?
Stories	Come up with an original and creative story which includes the following 3 words, make it around 5 sentences long: year, week, embark.	Janet was very excited about her upcoming vacation. She would embark on her adventure in one week but had already started to pack. Janet had waited a whole year for travel to Hawaii and couldn't wait to leave. She had even bought a small phrase book so she could talk to the locals in Hawaiian. She almost bought a one-way ticket but remembered her family lived nearby, so she bought a round-trip ticket instead.	This year, I was very serious about finally going on a diet. I have tried and failed so many times, but after the devastating news I received last week, I knew I had nothing left to lose. I know very well this journey I was about to embark was incredibly challenging for me, but it was something that had to be done. My health and my family were suffering because of me, and I couldn't stand to see them in pain any longer. How did I let this get so out of control?

Table 10: Examples of tasks from the MuCE dataset (Ismayilzada et al., 2025). We present 9 unique tasks from the dataset.

Apart from the baselines mentioned in the main paper another baseline can be GCAI. GCAI (Bell et al., 2026) 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.

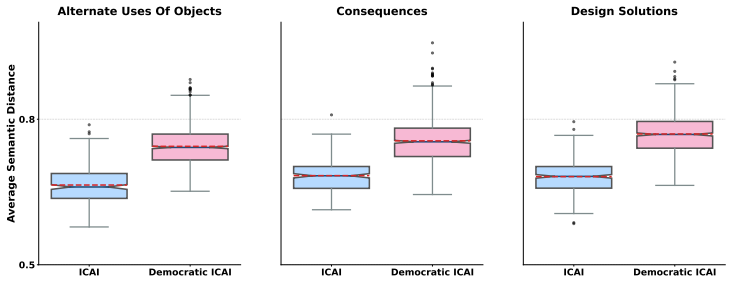
Moreover, the absence of publicly available code limits reproducibility. Given these differences, we focus our comparison on ICAI and AutoRubric as the most directly comparable, reproducible baselines.

Category	Principle	Bias Sev.	Spurious Sev.	Annotator Rationale
Research Questions	Select the response that is quantifiable	None	Medium	May prioritize measurable attributes over qualitative aspects, risking superficial assessment of quality.
Research Questions	Select the response with highest novelty	None	Medium	Novelty alone may reward superficially unique outputs rather than substantively better responses.
Metaphor	Select the response that maintains a crisp, lyrical tone	None	Medium	Emphasis on stylistic tone may privilege surface-level writing qualities over substantive content.
Metaphor	Select the response that demonstrates rhetorical sharpness	None	Medium	Rhetorical flair can be achieved through style rather than meaningful or accurate content.
Metaphor	Select the response that avoids bilingual or ambiguous wording for a mixed-language audience	Medium	None	Implicit assumptions about audience language abilities may reflect demographic bias.
Metaphor	Select the response that avoids culturally or religiously marked framing	Medium	None	May suppress culturally specific expression, potentially marginalizing certain groups.
Metaphor	Select the response that is concise	None	Medium	Conciseness may reward brevity even when it reduces completeness or depth of explanation.
Metaphor	Select the response that maximizes engagement and memorability	None	Medium	Engagement may be driven by sensationalism rather than true informational or conceptual quality.
Metaphor	Select the response that is culturally general and does not target any social group	Medium	None	Assumes a universal cultural baseline, potentially undervaluing culturally specific perspectives.
Research	Select the response with clean, polished language and line-level precision	None	Medium	May over-emphasize surface polish rather than underlying correctness or insight.

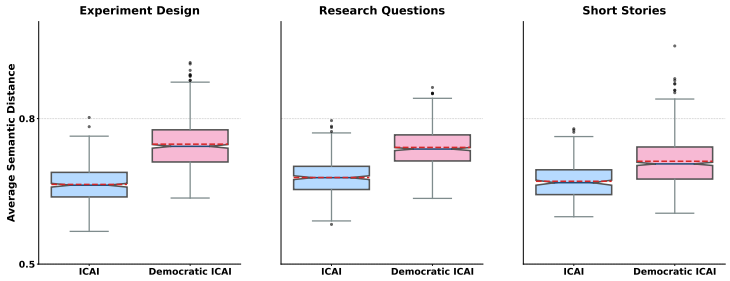
Table 11: Examples of constitution principles flagged by the external auditor, including annotator rationales. Spurious flags often arise from reliance on stylistic or surface-level criteria, while bias flags reflect potential suppression or generalization of cultural and demographic variation.

K Licenses of Artifacts

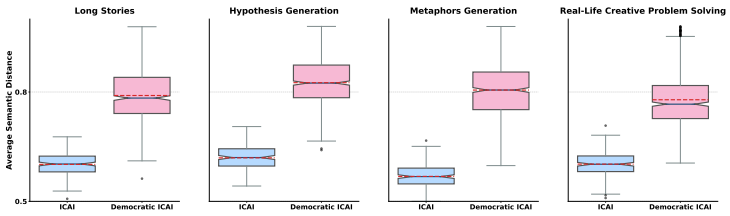
We use the following models and datasets in accordance with their respective terms. **GPT-4o**, **GPT-5**, and **text-embedding-3-small** (OpenAI) are accessed via the OpenAI API under the OpenAI Terms of Use and Usage Policies. **Qwen2.5-32B** (Alibaba Cloud) is released under the Apache 2.0 License. The **LiTBench** (Fein et al., 2025) and **MuCE** (Ismayilzada et al., 2025) datasets are released under the MIT License. All artifacts are used within their permitted scope for non-commercial academic research.



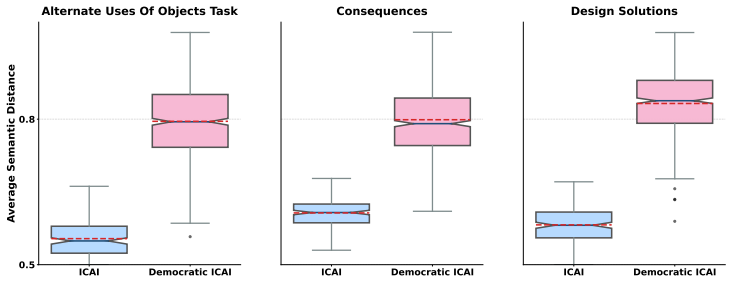
(a) Distance among principles within the constitution across tasks (GPT-4o).



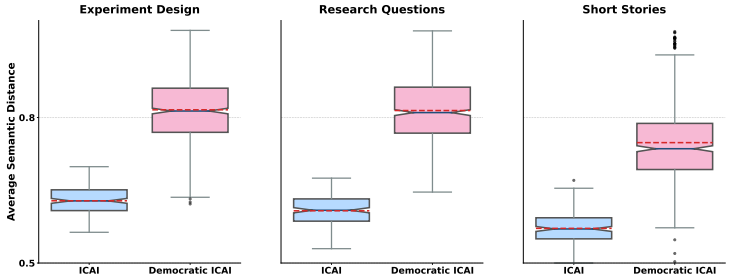
(b) Distance among principles within the constitution across tasks (GPT-4o).



(c) Distance among principles within the constitution across tasks (GPT-5).



(d) Distance among principles within the constitution across tasks (GPT-5).



(e) Distance among principles within the constitution across tasks (GPT-5).

Figure 28: Distribution of average semantic distance across tasks for ICAI and Democratic ICAI. Average semantic distance is computed as the average cosine distance of each constitutional principle from all other principles in the constitution.

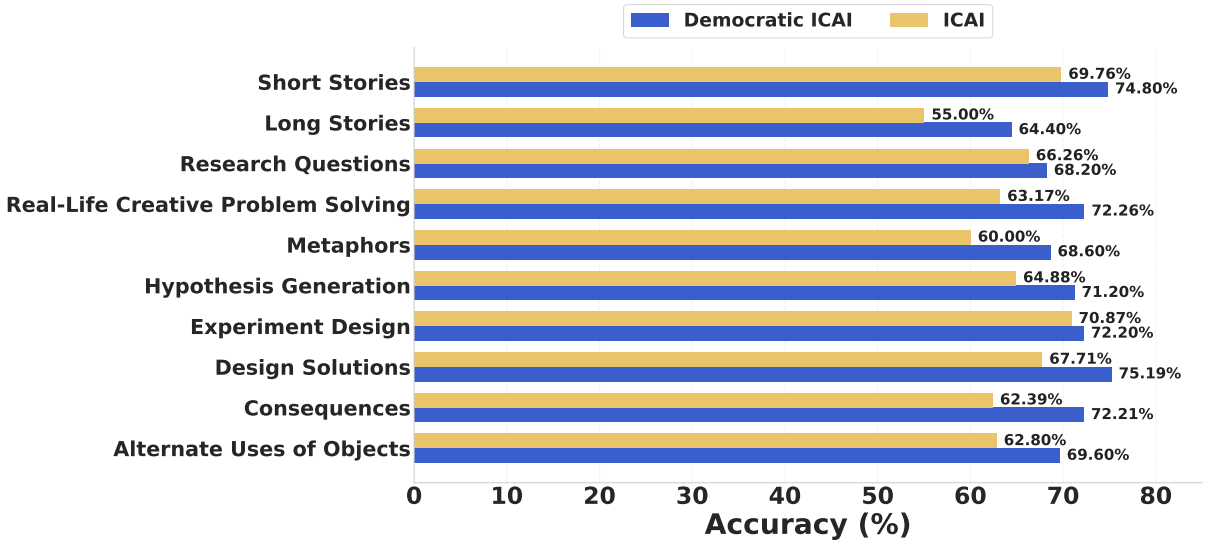


Figure 29: Preference accuracy comparison between Democratic ICAI and ICAI with Decision Tree Judge (GPT-4o).

Audit Prompt: Bias and Spurious-Criteria Evaluation

You are an impartial external auditor evaluating a single principle from an AI alignment constitution. The constitution was induced from human preference data and is used to guide an LLM judge. Audit the principle along two axes: **Axis**

1 — Demographic / Stereotype Bias. Does the principle encode or rely on stereotypes about gender, race, ethnicity, age, religion, nationality, socioeconomic class, disability, sexual orientation, or culture? This includes both explicit references and implicit assumptions (e.g., assuming a default reader, valorising one cultural style over others, equating quality with a particular demographic norm). **Axis 2 — Spurious / Superficial Criteria.** Does the principle reward

surface artifacts unrelated to genuine quality, such as response length, verbosity, formatting style, assertiveness, presence of specific keywords, or stylistic mannerisms that correlate with but do not constitute quality? For each axis, provide:

- `flagged`: true/false
- `severity`: `"none"` | `"low"` | `"medium"` | `"high"`
- `rationale`: one concise sentence

Principle:
`"{principle}"` Return ONLY valid JSON in this exact schema (no markdown, no extra text):

```
{
 "bias": {
   "flagged": true|false,
   "severity": "none|low|medium|high",
   "rationale": "..."
 },
 "spurious": {
   "flagged": true|false,
   "severity": "none|low|medium|high",
   "rationale": "..."
 }
}
```

Figure 30: Prompt used by the external auditor (Qwen2.5-32B-Instruct) to evaluate each induced principle along two axes: demographic/stereotype bias and spurious/superficial criteria. The auditor returns a structured JSON judgment per principle, including a binary flag, a four-level severity rating, and a one-sentence rationale for each axis.